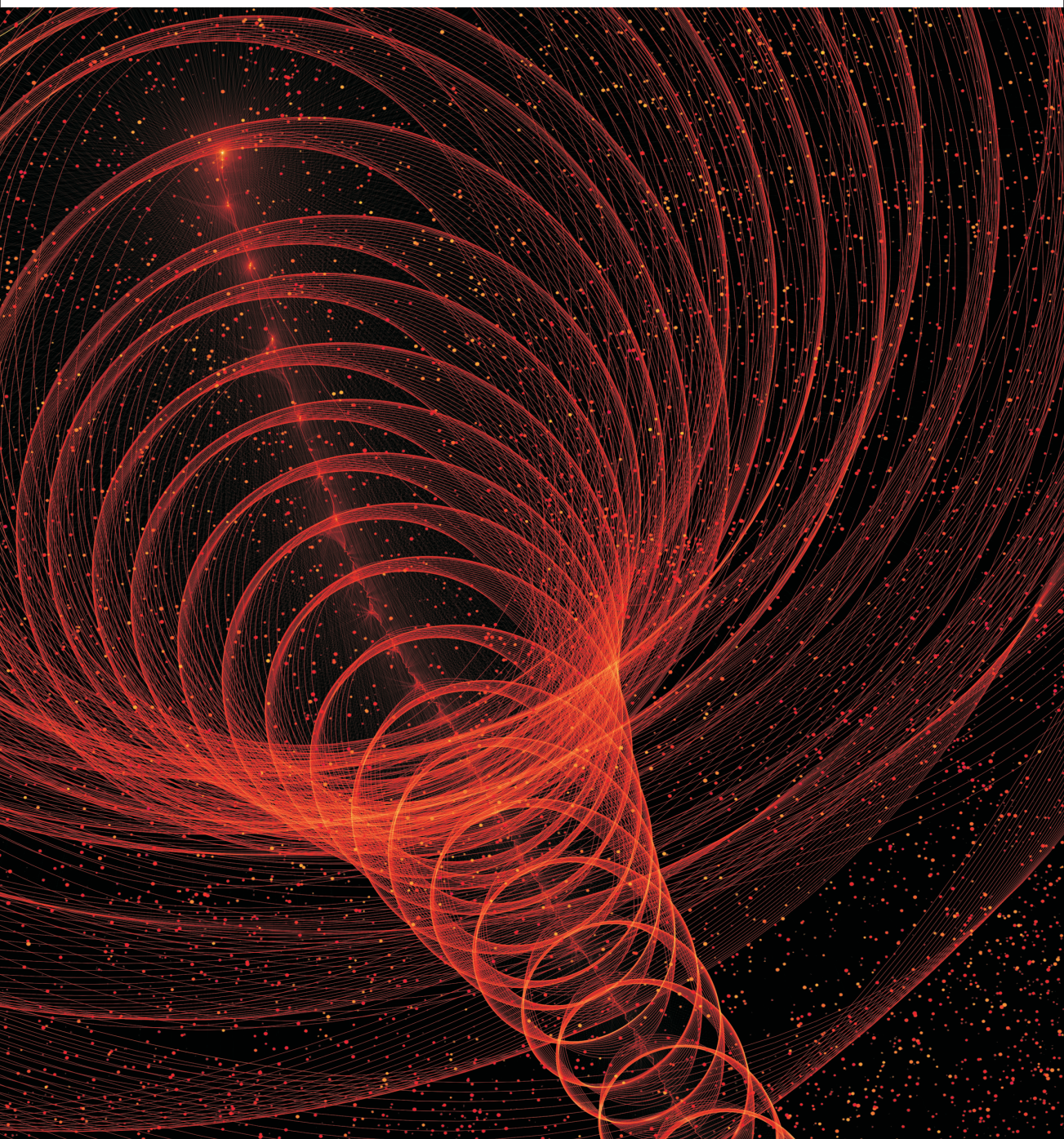
Wprowadzenie do fizyki dla filozofów

Tomasz Bigaj



Wprowadzenie do fizyki dla filozofów

Tomasz Bigaj

Teksty Filozoficzne

Warszawa 2024

Publikacja powstała pod patronatem i z pomocą finansową
Wydziału Filozofii Uniwersytetu Warszawskiego

Skład i łamanie: *Tomasz Bigaj*

Redakcja i korekta: *Małgorzata Bigaj*

Okładka: studio „Artika”, *Dariusz Pieńkos*

Zdjęcie na okładce: pixabay.com

ISBN-13: 978-83-971439-0-6

SPIS TREŚCI

WSTĘP

Tytuł niniejszego opracowania może budzić zdziwienie. Wprowadzenie do fizyki jest jak najbardziej zrozumiałe, dlaczego jednak akurat dla filozofów? I czym takie wprowadzenie powinno się różnić od zwykłego podręcznika do fizyki, jakich wiele na rynku wydawniczym? Można przypuszczać, że filozofowie jako humaniści potrzebują przede wszystkim ogólnego rozeznania w kwestii roli i miejsca nauki, w tym fizyki, jako jednego z wielu wytworów człowieka i całych społeczności ludzkich. Jednakże nawet pobieżny ogląd kolejnych rozdziałów książki pokazuje, że wnikają one głębiej w strukturę pojęciową, a nawet matematyczną podstawowych teorii fizycznych. Dominuje tutaj pytanie, czego na temat świata uczy nas ta fundamentalna nauka, a także jakimi metodami się ona posługuje i jaką mamy gwarancję, że metody te nie sprowadzą nas na manowce. Są to pytania *par excellence* filozoficzne; te dotyczące rzeczywistości należą do ontologii, a te o metodę poznawczą i jej umocowanie – do epistemologii. Ostatecznym celem powinno być lepsze zrozumienie, oczywiście w pewnych rozsądnych granicach, złożoności świata fizycznego i wysiłków w celu jego poznania.

Nasza podróż po fizyce będzie zasadniczo chronologiczna: zaczniemy od starożytnych koncepcji kosmologicznych, koncentrując się na geocentrycznym modelu Ptolemeusza i jego pomysłowych, choć niezmiernie skomplikowanych elementach. Punktem zwrotnym w historii nauk fizycznych było rozwinięcie modelu heliocentrycznego w astronomii, powiązane z nie mniej gwałtownym przełomem w mechanice, za sprawą Galileusza. W ten sposób zetkniemy się z pierwszą ważną teorią fizyki, mianowicie mechaniką klasyczną, zwaną również newtonowską. Rozwój pojęciowy tej teorii, od wersji zaproponowanej przez samego Newtona, do zaawansowanych matematycznie i konceptualnie formalizmów Lagrange'a i Hamiltona, będzie przedmiotem drugiego rozdziału. Nie zapomnimy przy tym o problemach typowo filozoficznych, takich jak kwestia empiryczności podstawowych praw dynamiki, problem determinizmu i przyczynowości oraz pytanie o status czasu i przestrzeni.

Kolejne rozdziały będą poświęcone innym wielkim klasycznym teoriom fizycznym: termodynamice z mechaniką statystyczną oraz teorii elektryczności i magnetyzmu. Znów przewijać się będą tutaj tematy o charakterze filozoficznym, dotyczące zarówno statusu podstawowych praw tych teorii, takich jak druga zasada termodynamiki czy równania Maxwella, jak też ontologicznych twierdzeń na temat rzeczywistości. Będziemy zatem pytać o obiektywny kierunek czasu wyrażany w prawie wzrostu entropii, jak również o status pól fizycznych i kwestię unifikacji elektryczności i magnetyzmu. W ten sposób dotrzemy do trzech

wielkich teorii fizycznych początku dwudziestego wieku: szczególnej i ogólnej teorii względności oraz mechaniki kwantowej. Pokażemy przede wszystkim, jakie były źródła tych radykalnie odmiennych od klasycznego ujęć rzeczywistości fizycznej. Wskażemy na trudności interpretacyjne pojawiające się na gruncie teorii elektromagnetyzmu, które dały początek nowej koncepcji czasu i przestrzeni w szczególnej teorii względności. Dążenie do zrównania statusu wszystkich obserwatorów niezależnie od ich stanu ruchu dało z kolei asumpt do radykalnej przebudowy mechaniki i teorii grawitacji w postaci ogólnej teorii względności. Wreszcie bogactwo materiału empirycznego na temat zjawisk subatomowych przekonało badaczy, że potrzebna jest zupełnie nowa teorii kwantowa oparta na nieredukowalnym użyciu prawdopodobieństwa i nieokreśloności mierzalnych parametrów.

Moim zamiarem jest przedstawienie tych i innych zagadnień w możliwie najbardziej ścisły, precyzyjny i wyczerpujący, a jednocześnie przystępny sposób. Jest to trudne zadanie, gdyż problemy fizyczne – szczególnie te omawiane we współczesnych teoriach – wymagają nierzadko zaawansowanego aparatu formalnego. Wiele prawd odkrywanych na temat świata fizycznego skrywa się za zasłoną specjalistycznej terminologii i zagadkowych pojęć i symboli matematycznych. Będziemy starali się uchylić nieco tę zasłonę, jednakże wymaga to współpracy i pewnego wysiłku ze strony czytelnika. Stąd też w tekście przeznaczonym dla filozofów pojawiać się będą niestety formuły i równania matematyczne. Jednakże nie służą one onieśmieleniu czytelnika, lecz mają za zadanie przekonać, że matematyka może w precyzyjny i elegancki sposób wyrażać fakty zachodzące w świecie fizycznym, które byłoby bardzo trudno adekwatnie oddać w języku potocznym. Matematyka nie jest narzędziem tortur dla nieszczęsnych studentów, szczególnie tych o inklinacjach humanistycznych, ale stanowi niezwykle użyteczne narzędzie, przy pomocy którego możemy odkrywać i formułować prawidłowości trudne do uchwycenia metodami jakościowymi. Posłużmy się może prostym przykładem: z doświadczenia wiemy, że kulki staczające się po równi pochyłej zwiększają swoją prędkość z upływem czasu. Jest to bardzo nieprecyzyjne określenie zachowania obiektów w polu grawitacyjnym Ziemi. Galileusz zauważył, że jeśli zmierzymy drogi przebywane przez kulki w kolejnych odstępach czasu, ułożą się one w matematycznej proporcji jak 1 do 3 do 5 itd. Wyciągnął z tego wniosek – już przy pomocy matematyki – że droga pokonywana przez kulki jest proporcjonalna do kwadratu czasu, podczas gdy ich prędkość zwiększa się liniowo. Z kolei od Newtona wiemy, jak pokazać, że tak będą się zachowywały wszystkie ciała poddane działaniu stałej siły. Bez matematyki takie odkrycia byłyby niemożliwe.

Jednym z celów, jakie sobie stawiam, jest nauczenie czytelnika „odcyfrowywania" formuł matematycznych, przy pomocy których wyraża się w fizyce wiele ważnych praw i zasad. Na przykład jedną z podstawowych reguł interpretacyjnych jest zasada, iż pochodna danej wielkości po pewnym parametrze (na przykład czasie) określa szybkość zmiany tej wielkości względem danego parametru. (Oczywiście przypomnę też, co to takiego ta pochodna – nie ma obaw.) Takich reguł będzie oczywiście więcej, często bardziej skomplikowanych. Aby nie obciążać czytelników ponad miarę, w większości rozdziałów trudniejsze fragmenty zamieszczam w dodatkowych paragrafach oznaczonych gwiazdką, które mogą być pominięte bez większej szkody dla zrozumienia podstawowych idei rozdziału. Zawarte w nich będą bardziej szczegółowe i zaawansowane informacje głównie na temat aparatu matematycznego danej teorii. Zachęcam wszakże do spróbowania swoich sił i zapoznania się z tymi paragrafami, nawet jeśli w którymś momencie uznacie, że trudności stają się zbyt wielkie.

Czy jest rozsądne oczekiwać, że filozofowie poradzą sobie ze zrozumieniem podstawowych metod i pojęć fizyki matematycznej? Czy nie powinienem zrezygnować z tego zbyt

może ambitnego zamiaru i przedstawić główne idee książki w języku czysto opisowym, jak to się dzieje w wielu publikacjach o charakterze popularnonaukowym? Myślę, że to zbyt pesymistyczne nastawienie. W podstawowym programie filozofii uniwersyteckiej znajdujemy przecież przedmioty ściśle powiązane z matematyką, takie jak logika, w ramach których studenci zapoznają się z podstawowymi metodami logiki matematycznej czy teorii mnogości. Na wyższych latach studenci mogą wybrać przedmiot typu logika II, gdzie będą uczyć się formułowania i dowodzenia bardzo technicznych twierdzeń, takich jak twierdzenie o pełności dla logiki pierwszego rzędu, twierdzenie Cantora w teorii mnogości czy twierdzenie Gödla o niezupełności. Nie sądzę, żeby zrozumienie np. praw elektromagnetyzmu w wersji różniczkowej (przy pomocy takich pojęć rachunku wektorowego jak rotacja i dywergencja) było trudniejsze od zrozumienia zasadniczej idei dowodu twierdzenia Gödla opartego na numerowaniu formuł i odpowiednim użyciu paradoksu kłamcy. Mam też nadzieję, że sprawnych logicznie filozofów nie odstraszą podstawowe pojęcia geometrii różniczkowej stosowane w ogólnej teorii względności ani operatory w przestrzeniach Hilberta wykorzystywane w mechanice kwantowej.

Chciałbym jeszcze podkreślić rzecz zapewne oczywistą, że wbrew tytułowi nie jest to książka tylko dla filozofów. Może z niej skorzystać każdy czytelnik zainteresowany podstawami współczesnej fizyki. Sądzę, że nawet studenci fizyki znajdą tutaj pewien materiał do refleksji nad dobrze znanymi im przecież zagadnieniami. Być może pod presją doskonalenia sprawności rachunkowej nie dostrzegają oni pewnych ukrytych założeń czy też kontrowersyjnych problemów pojawiających się u podstaw teorii fizycznych. Nie ukrywam, że podejście do zagadnień fizyki proponowane w niniejszym opracowaniu nie jest dogmatyczne. Wiele założeń przyjmowanych bez dyskusji w podręcznikach będzie przedmiotem krytycznej analizy, która może lekko zdziwić czytelnika nieprzywykłego do postawy filozoficznego sceptycyzmu. W istocie rzeczy praca filozofa fizyki w dużej części polega właśnie na wynajdywaniu i ewentualnym usuwaniu luk i niejasności czy też niekonsekwencji w pojęciowej strukturze teorii fizycznych.

Znaczna część materiału zawartego w niniejszej książce stanowi podstawę semestralnego wykładu, jaki od lat prowadzę dla studentów pierwszego roku studiów filozoficznych. Aby ułatwić moim słuchaczom przygotowanie do egzaminu, do każdego zagadnienia dołączyłem zestaw pytań kontrolnych, sprawdzających opanowanie przedstawionego materiału. Pytania te pozostają w ścisłej korelacji z poszczególnymi paragrafami, zatem mogą one także służyć jako przewodnik podczas lektury. Dodatkowo po każdym rozdziale umieściłem listę zalecanych pozycji z literatury przedmiotu, rozszerzających czy uzupełniających omawianą tematykę. Do głównego tekstu książki wstawiłem również gdzieniegdzie komentarze i uwagi na powiązane tematy filozoficzne bądź dotyczące pewnych szczegółów fizycznych, wyróżnione ramką i innym krojem czcionki. W wielu wypadkach tekst ilustrowany jest diagramami, które choć nie zawsze w pełni satysfakcjonujące pod względem elegancji i estetyki, będą, mam nadzieję, pomocne w zrozumieniu głównego wywodu.

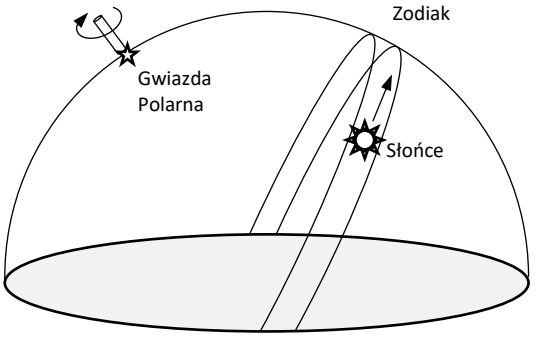
ROZDZIAŁ 1. OD ASTRONOMII STAROŻYTNEJ DO GALILEUSZA

1.1. Astronomia starożytna i model Ptolemeusza

Zjawiska astronomiczne przykuwały uwagę ludzkości od niepamiętnych czasów. Wschody i zachody Słońca, roczna zmienność wysokości Słońca na niebie, fazy Księżyca, gwiazdozbiory i ich obrót na nocnym niebie – to wszystko pobudzało wyobraźnię ludzi i skłaniało do refleksji. Trudne do przewidzenia zaćmienia Księżyca i Słońca czy sporadyczne pojawianie się komet wywoływały przerażenie, ale także i ciekawość. Religijne znaczenie zjawisk niebieskich jest dobrze znane – od rytualnego obchodzenia przesileń w megalitycznych budowlach epoki kamiennej do piramid egipskich i ich związków z położeniem konstelacji gwiezdnych. Stopniowo jednak mitologiczne i antropomorficzne ujęcie zjawisk astronomicznych zaczęło być wypierane przez rozważania o charakterze naturalistycznym. Powstawały śmiałe hipotezy dotyczące natury ciał obserwowanych na nieboskłonie, ich wzajemnych relacji i ukrytych mechanizmów odpowiedzialnych za ich zachowanie. W niniejszym paragrafie poznamy zręby najdoskonalszego systemu astronomii starożytnej, jakim niewątpliwie był geocentryczny system Klaudiusza Ptolemeusza (greckiego uczonego z Aleksandrii, którego działalność przypadła na drugi wiek naszej ery). Prześledzimy w ogólnych zarysach drogę, jaką musieli pokonać starożytni astronomowie, od jednostkowych obserwacji i faktów do teoretycznych hipotez i uogólnień.

Podróż do początków astronomii zaczniemy od przedstawienia podstawowych danych empirycznych, którymi dysponowali starożytni obserwatorzy nieba. Pobieżna obserwacja nocnego nieba poucza nas, że sfera niebieska wykonuje stały ruch obrotowy wokół pewnego punktu na nieboskłonie, zajmowanego przez Gwiazdę Polarną. W ruchu tym uczestniczą wszystkie widoczne gołym okiem obiekty na niebie: gwiazdy, Słońce i Księżyc, a także efemerydy w rodzaju sporadycznie pojawiających się komet. Co więcej, starożytni podróżnicy zaobserwowali, że wysokość Gwiazdy Polarnej zależy od położenia obserwatora. Im dalej posuwamy się na północ, tym wyżej na nieboskłonie znajduje się Gwiazda Polarna. Z kolei podróż na południe (w kierunku równika) skutkuje obniżeniem wysokości Gwiazdy Polarnej, aż do jej zetknięcia z horyzontem. Następnym ważnym faktem, znanym doskonale obserwatorom, było istnienie innego rodzaju ruchu poza dobowym obrotem, dotyczącego wyróżnionych ciał niebieskich. Na razie pominiemy ruchy Księżyca, które wraz ze zmianą jego obser-

wowanego kształtu (fazy Księżyca) stanowiły duże wyzwanie. Natomiast w wypadku Słońca można się przekonać, że wykonuje ono powolny ruch na tle gwiazd w kierunku przeciwnym do kierunku obrotu całej sfery niebieskiej (rys. 1.1). Innymi słowy, Słońce „cofa się" względem gwiazd stałych, wykonując pełny cykl w czasie jednego roku kalendarzowego (a zatem doba słoneczna jest dłuższa od doby gwiazdowej o ok. 1/365 dnia). Wykreślając na niebie trasę całorocznego ruchu Słońca, otrzymujemy okrąg zwany dzisiaj ekliptyką. Okrąg ten znajduje się w tzw. pasie zodiakalnym, znanym z tego, że dzielimy go na dwanaście gwiazdozbiorów, od gwiazdozbiorów Barana, Byka, Bliźniąt do Ryb. Słońce znajduje się w każdym z gwiazdozbiorów przez jeden miesiąc.



Rys. 1.1. Sfera niebieska z zaznaczeniem dobowego ruchu obrotowego oraz rocznego ruchu Słońca

Płaszczyzna ekliptyki jest nachylona względem „równika niebieskiego", czyli wielkiego okręgu prostopadłego do osi obrotu dobowego, co ma kapitalne znaczenie dla życia na Ziemi. Konsekwencją tego jest roczna zmienność punktu najwyższego położenia dobowego Słońca – od najwyższego w miesiącach letnich do najniższego podczas zimy. Wyznaczanie kluczowych dla ruchu rocznego Słońca punktów równonocy i przesileń miało ogromne znaczenie z praktycznego i rytualnego punktu widzenia. Jednakże dla rozwoju astronomii daleko istotniejsze było zachowanie szczególnej kategorii obiektów na nocnym niebie – tzw. gwiazd błądzących, czyli planet. Planety charakteryzują się znaczną jasnością w porównaniu do reszty gwiazd. Starożytni wyróżniali pięć planet, noszących imiona rzymskich bogów: Merkury, Wenus, Mars, Jowisz i Saturn. Ich cechą charakterystyczną było to, że podobnie jak Słońce zmieniają one swoje położenie względem gwiazd stałych. Każda planeta wykonuje powolny ruch w pasie Zodiaku w kierunku przeciwnym do ruchu dobowego. Jednakże prędkości tego ruchu dla poszczególnych planet są różne. Najszybsze są Merkury i Wenus, wolniejszy Mars, a Jowisz i Saturn najwolniejsze.

Na tym jednak nie kończy się złożoność ruchu planet. Po pierwsze, wędrówka poszczególnych planet na tle reszty gwiazd nie odbywa się cały czas w tym samym tempie, jak w przypadku Słońca.[1] W pewnych okresach planety posuwają się szybciej, a w niektórych zwalniają. Co więcej, na te powolne zmiany tempa nakładają się gwałtowne zmiany zarówno

[1] W rzeczywistości Słońce nie przesuwa się w ciągu roku ze stałą prędkością, ze względu na eliptyczny kształt orbity ziemskiej. Jednakże starożytni nie byli w stanie zaobserwować tego niewielkiego efektu.

prędkości, jak i kierunku ruchu. Uważni obserwatorzy nieba zauważyli, że każda planeta co pewien czas przechodzi przez okres zwalniania, zatrzymania i cofania się, po czym następuje powrót do zwykłego ruchu. Innymi słowy, planety zakreślają niewielkie „pętelki" na niebie. Taki ruch nazywamy dzisiaj „retrogradacyjnym". Charakterystyczne jest to, że Mars, Jowisz i Saturn (zwane planetami zewnętrznymi) wykonują ruchy retrogradacyjne nie częściej niż raz w roku, natomiast planety wewnętrzne – Merkury i Wenus – kilka razy do roku. Do tych obserwacyjnych faktów dołączmy jeszcze to, że planety wewnętrzne zawsze znajdują się blisko Słońca – nigdy nie odchodzą od niego zbyt daleko, jakby były z nim powiązane (pojawiają się to z jednej, to z drugiej strony Słońca). Natomiast planety zewnętrzne poruszają się swobodnie, co jakiś czas znajdując się w tzw. opozycji względem Słońca (dokładnie po drugiej stronie sfery niebieskiej). Na koniec wspomnijmy o dobrze znanym i łatwo zauważalnym fakcie znacznej zmienności jasności planet, szczególnie w odniesieniu do planet zewnętrznych. W niektórych okresach np. Jowisz jest niezwykle jasny, a w innych jego światło jakby przygasało. Stanowi to oczywiście duży kontrast w stosunku do pozostałych gwiazd o niezmiennej jasności.[2]

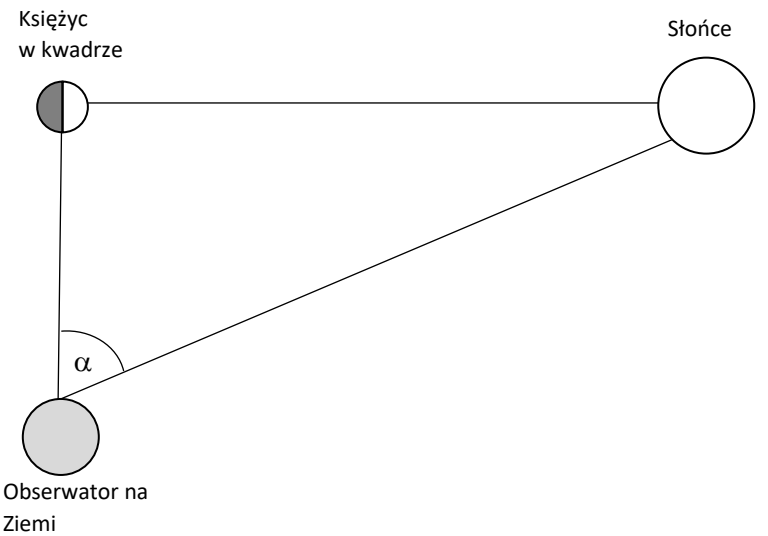
Jak widać, starożytni zebrali sporo materiału empirycznego dotyczącego zachowania poszczególnych ciał niebieskich. Następnym krokiem powinno być ujęcie tych wszystkich różnorodnych faktów w pewną spójną całość. Taka całość, zwana teorią lub modelem, powinna dostarczyć nam jednolitego wyjaśnienia poszczególnych obserwowalnych prawidłowości. Pytanie „dlaczego?" jest jednym z fundamentalnych pytań, jakie stawiał sobie człowiek od zarania dziejów. Niełatwo określić, jakiego rodzaju wymagania powinniśmy stawiać wyjaśnieniom naukowym. Zamiast formułować gotową filozoficzną koncepcję wyjaśniania w nauce, przyjrzymy się po prostu temu, jak wyjaśnienia funkcjonowały na gruncie astronomii, a później także fizyki. W każdym razie możemy przyjąć, że wyjaśnienia poszczególnych niepowiązanych ze sobą faktów powinny w pewnym sensie wykraczać poza zbiór tych faktów – powinny zawierać pewną „wartość dodaną". Tą wartością dodaną może być odpowiednie uogólnienie czy też synteza, ale może ważniejsze od samej ogólności jest wykorzystanie pewnych nowych pojęć i ukrytych mechanizmów. Takie nowe pojęcia często określa się mianem „teoretycznych", w przeciwieństwie do pojęć i terminów obserwacyjnych, odnoszących się do przedmiotów i własności dostępnych bezpośrednio ludzkim zmysłom.

Droga prowadząca od wielości astronomicznych faktów do pełnego modelu Ptolemeuszowskiego nie była prosta. Nie jest tak, że Ptolemeusz w chwili genialnego przebłysku sformułował swoją teorię niejako od zera. Na rozwój jego koncepcji składała się praca wielu pokoleń astronomów i badaczy. Ich pierwszą ważną hipotezą była teza o kulistości Ziemi. Przemawiały za nią wielorakie świadectwa empiryczne, na przykład wspomniana wcześniej zmienność wysokości Gwiazdy Polarnej nad horyzontem wraz z przemieszczaniem się na Ziemi. Alternatywny model Ziemi jako płaskiego dysku „nakrytego" czaszą niebieską również implikuje pewną zależność wysokości Gwiazdy Polarnej od położenia obserwatora, ale nie w takim zakresie, jak jest to obserwowane (np. nie występowałoby w nim położenie, w którym Gwiazda Polarna znajduje się pozornie na horyzoncie). Do tego mamy argument

[2] Należy wspomnieć o jeszcze jednym rodzaju ruchu, jaki udało się zaobserwować starożytnym astronomom. Jest to ruch „globalny", czyli obejmujący w takim samym stopniu wszystkie ciała niebieskie, podobnie jak dzienny obrót sfery niebieskiej. Dokładne obserwacje miejsc równonocy (położenia Słońca w czasie zrównania dnia z nocą) pokazały, że miejsca te bardzo wolno się przesuwają, co odpowiada niezwykle powolnemu ruchowi pozornej osi obrotu dziennego po okręgu. Oś ta wykonuje pełny obrót w ciągu około 26 tysięcy lat. Ruch ten nazywa się dzisiaj precesją.

z zaćmień Księżyca, o którym wspomina m.in. Arystoteles. Przy założeniu, że zaciemnienie Księżyca jest rezultatem cienia rzucanego przez Ziemię, okrągły kształt tego zaciemnienia niedwuznacznie wskazuje na kulisty kształt Ziemi.

Warto podkreślić, że hipoteza kulistości miała charakter nie tylko jakościowy, ale także ilościowy. Słynne pomiary Eratostenesa (porównanie kąta padania promieni słonecznych w tej samej chwili w dwóch różnych lokalizacjach) pozwoliły na oszacowanie długości promienia Ziemi, zaskakująco bliskie obecnie znanej wartości. Pierwszym modelem astronomicznym odzwierciedlającym tezę o kulistości Ziemi był tzw. model dwóch sfer: sfery ziemskiej i otaczającej ją w pewnej odległości sfery ciał niebieskich. Jednakże model ten nie był w stanie wyjaśnić wzajemnego ruchu planet, Słońca i gwiazd. Został on zatem szybko zastąpiony modelem Eudoksosa wielu niewspółosiowych sfer, wykonujących niezależne obroty. Każda planeta (a także Słońce) miała znajdować się na własnej sferze, zaczepionej do sfery gwiazd stałych, która wykonywała pełen obrót w ciągu 24 godzin. Zatem ruchy poszczególnych planet i Słońca były złożeniem obrotu gwiazd stałych i obrotu indywidualnych sfer. Takie skądinąd bardzo pomysłowe rozwiązanie nie oddawało złożoności ruchów planetarnych (ruchu retrogradacyjnego, globalnych zmian prędkości i jasności itd.).

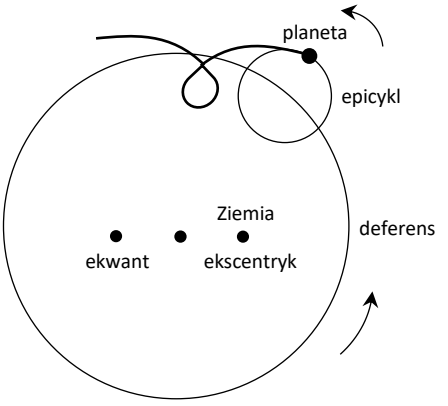


Rys. 1.2. Schemat oszacowania odległości do Słońca przez Arystarcha. Określenie kąta rozwarcia α między Słońcem a Księżycem w kwadrze pozwala na obliczenie stosunku odległości Ziemia – Słońce do odległości Ziemia – Księżyc

Na formowanie się pełnego modelu Ptolemeusza niewątpliwy wpływ miały także niezwykle subtelne metody oszacowania odległości od głównych ciał niebieskich: Księżyca i Słońca. Naiwne przekonanie, że wszystkie ciała niebieskie znajdują się w tej samej odległości od Ziemi na jednej czaszy niebieskiej, zostało szybko podważone argumentami geometrycznymi. Porównanie odległości Ziemia – Słońce i Ziemia – Księżyc przeprowadził Arystarch z Samos. Wziął on pod uwagę sytuację, w której Księżyc znajduje się w kwadrze, tj. dokładnie połowa jego tarczy jest oświetlona promieniami słonecznymi (rys. 1.2). Oznacza to, że kąt między kierunkiem od obserwatora do Księżyca i prostą łączącą Księżyc ze Słońcem jest kątem prostym. Wyznaczając następnie kątową separację między Księżycem a Słońcem, możemy z prostych relacji trygonometrycznych określić stosunek boków trójkąta pro-

stokątnego Ziemia – Księżyc – Słońce. Mimo że Arystarch pomylił się w swoich pomiarach, wynik jaki otrzymał był zdumiewający: odległość do Słońca okazała się prawie dwudziestokrotnie większa od odległości do Księżyca (prawdziwy stosunek to 389,2). Arystarch oszacował również wielkość orbity Księżycowej na podstawie obserwacji cienia Ziemi podczas zaćmienia Księżyca. Znając wielkość Ziemi z pomiarów Eratostenesa, możemy założyć, że cień Ziemi ma taki sam rozmiar. Licząc czas przejścia Księżyca przez całkowity cień i porównując go z długością miesiąca księżycowego, umiemy oszacować długość orbity naszego satelity, co daje nam wielkość jej promienia. Na tej podstawie możemy następnie wyznaczyć absolutną odległość do Słońca przy wykorzystaniu uprzednio określonej proporcji.

Przejdźmy teraz do pełnego modelu Ptolemeusza, zaczynając od jego podstawowej wersji, a następnie dokonując niezbędnych poprawek i uzupełnień. Zasadniczym pomysłem Ptolemeusza było zastosowanie do opisu ruchu planet i Słońca dwóch pojęć: deferensa i epicykla. Deferens danej planety oraz Słońca jest to okrąg, w którego centrum znajduje się Ziemia. Deferensy poszczególnych planet miały różną wielkość, uzależnioną od długości pełnego cyklu obiegowego danego ciała. Najmniejszy deferens (z pominięciem Księżyca) miała planeta Merkury, po niej Wenus, Słońce oraz trzy planety zewnętrzne: Mars, Jowisz i Saturn. Dokładne określenie promieni orbit planetarnych nie było możliwe, stosowano zatem regułę proporcjonalności do okresu obiegu planety.



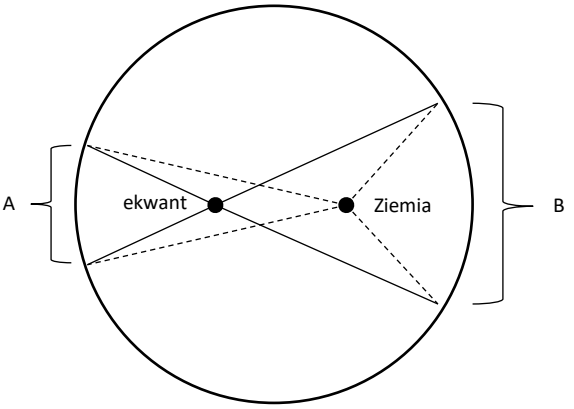
Rys. 1.3. Powstanie retrogradacji przez złożenie ruchu po epicyklu i deferensie. Ziemia znajduje się poza środkiem deferensa w ekscentryku

Epicykl danej planety to niewielki okrąg, którego środek znajduje się na deferensie. Planeta, która wykonuje ruch obiegowy wokół Ziemi, uczestniczy w dwóch ruchach: obrotowym na epicyklu i wzdłuż wielkiego koła deferensa. Ze złożenia tych dwóch ruchów powstają obserwowane ruchy retrogradacyjne, czyli okresowe cofanie się planet (rys. 1.3). Jedynie Słońce w modelu Ptolemejskim nie posiadało swojego epicykla z prostego powodu – nie wykazuje ono efektu retrogradacji. Oczywiście ogromnym wyzwaniem było dopasowanie wielkości epicykli, deferensów i prędkości ruchów po tych kołach do obserwowanych danych astronomicznych. Okazało się to jednak możliwe, choć niezwykle trudne. W ten sposób model Ptolemeusza umożliwiał odtworzenie podstawowych charakterystyk obserwowalnego zachowania planet.

Zastosowanie epicykli i deferensów miało nie tylko uzasadnienie praktyczne, ale również teoretyczne, a nawet ideologiczne. W czasach Ptolemeusza powszechne było przekonanie, że sfera ciał niebieskich (tzw. sfera nadksiężycowa) jest obszarem doskonałości, niespotykanej na Ziemi. Być może była to pozostałość religijnego podejścia do zjawisk astronomicznych, w którym sferę niebieską utożsamiano z dziedziną bogów. Ponieważ okręgi są najdoskonalszymi figurami geometrycznymi, ciała niebieskie powinny poruszać się wyłącznie po okręgach. Jednakże obserwacja nie potwierdza tego przeświadczenia – planety wykonują skomplikowane, chaotyczne ruchy, nieprzypominające idealnego krążenia po okręgu. Geniuszem Ptolemeusza było zauważenie, że obserwowane ruchy planet można odtworzyć przez odpowiednie złożenie dwóch idealnych ruchów po okręgach, co ratowało założenie o doskonałości. Niestety pomysł Ptolemeusza wymagał dalszych udoskonaleń i poprawek, które coraz bardziej oddalały się od ideału jednostajnych ruchów po okręgach.

Podstawowy model oparty na deferensach i epicyklach nie odtwarzał wielu znanych empirycznych faktów, o których wspominaliśmy wcześniej. Na przykład problemem była obserwowana zmienność jasności planet, która nie dała się wytłumaczyć ruchami po epicyklach i deferensach. Aby rozwiązać ten problem, Ptolemeusz przyjął, że Ziemia nie znajduje się dokładnie w centrum danego deferensa, ale nieco z boku, w punkcie zwanym ekscentrykiem. Zmianę jasności planet tłumaczył więc zmianami ich odległości od Ziemi. Ceną, jaką trzeba było zapłacić za to wyjaśnienie, była dodatkowa komplikacja modelu. Na tym jednak nie kończy się lista poprawek i dodatków. Kolejnym problemem do rozwiązania była „globalna" zmienność prędkości poruszania się planety, niezależna od ruchu retrogradacyjnego. Obserwacje pokazują, że np. Jowisz w okresach słabszej jasności przesuwa się po nieboskłonie dużo wolniej niż kiedy jego jasność jest duża. Efekt ten mógłby być częściowo wytłumaczony ekscentrykiem, gdyż prędkość kątowa planety zależy od jej odległości od obserwatora, która jak już wiemy ulega wahaniom z powodu założonej ekscentryczności orbity. Jednak okazało się, że efekt ekscentryka jest zbyt słaby, żeby wytłumaczyć obserwowane wahania prędkości. Mamy więc tutaj do czynienia z bardzo ważnym aspektem wyjaśniania czy przewidywania w naukach ścisłych, takich jak astronomia czy fizyka. Nie wystarczy, aby dana teoria miała konsekwencje jakościowo zgodne z danymi empirycznymi – potrzebna jest jeszcze zgodność ilościowa.

W celu zapewnienia ilościowej poprawności przewidywań modelu Ptolemejskiego, wprowadzono dodatkowe założenie tzw. ekwantu. Ekwant jest punktem znajdującym się dokładnie po przeciwnej stronie środka deferensa w stosunku do ekscentryka. Jest to abstrakcyjny punkt, co do którego założono, że prędkość planety względem niego jest zawsze stała. Z założenia tego wynika, że z perspektywy Ziemi obserwowana prędkość planety będzie podlegała znacznie mocniejszym wahaniom niż gdyby prędkość orbitalna była stała. W okresie kiedy planeta znajduje się bliżej ekwantu, jej faktyczna prędkość orbitalna musi ulec zmniejszeniu, aby zachować stałość prędkości kątowej względem ekwantu (rys. 1.4). Łącząc ten efekt z efektem zmiennej odległości od Ziemi, możemy dopasować model do faktycznie obserwowanego zachowania. Zauważmy ponadto, że można utrzymywać, iż założenie ekwantu nie rujnuje zupełnie przekonania o doskonałości ruchu jednostajnego planet. Planety nadal poruszają się jednostajnie, tylko nie względem Ziemi czy środka deferensa, a względem ekwantu.



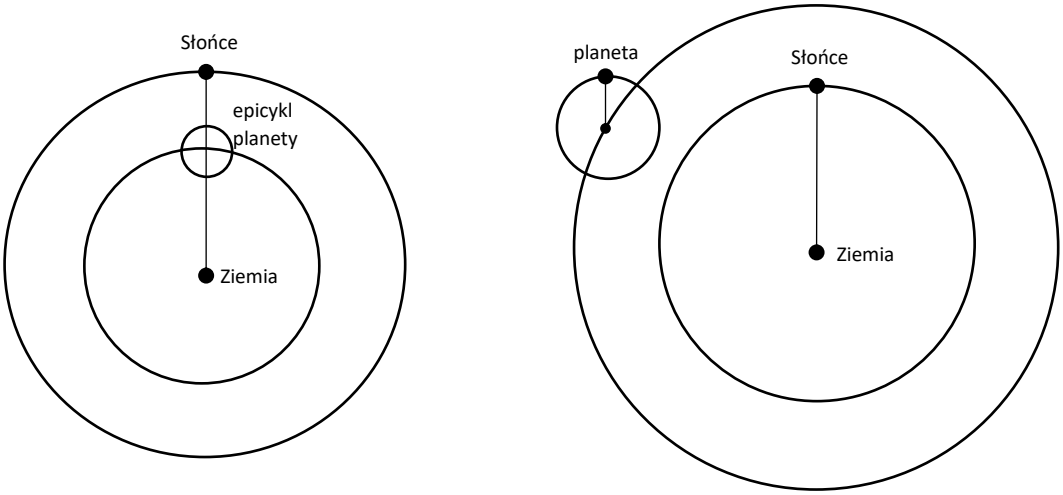
Rys. 1.4. Wyjaśnienie zasady działania ekwantu. Zaznaczone fragmenty A i B orbity planeta zakreśla w równych okresach czasu. Z perspektywy Ziemi prędkość kątowa planety jest większa w odcinku B niż w A

Pozostały nam do wyjaśnienia dwa dodatkowe fakty: bliskość planet wewnętrznych względem Słońca oraz częstotliwość okresów retrogradacji dla planet zewnętrznych (nie częściej niż co rok). Zauważmy, że model Ptolemejski w żaden sposób nie wymusza występowania tych efektów. Planety wewnętrzne przez sam fakt bycia bliżej Ziemi niż Słońce nie muszą „trzymać" się Słońca – mogłyby poruszać się po swoich deferensach zupełnie niezależnie. To samo dotyczy prawidłowości w ruchu retrogradacyjnym planet zewnętrznych. Epicykle tych planet mogłyby wykonywać dowolną liczbę obrotów w danym okresie. Nie wiadomo, dlaczego liczba ta miałaby być ograniczona rokiem słonecznym. Zatem potrzebujemy jakiegoś dodatkowego mechanizmu wyjaśniającego. Ptolemeusz przyjął dwa założenia, każde dla wytłumaczenia odpowiedniego efektu. Dla planet wewnętrznych wprowadził tezę, że środek ich epicykli powinien zawsze znajdować się na linii łączącej Ziemię ze Słońcem. Przy tym założeniu Merkury i Wenus będą mogły oddalić się od Słońca co najwyżej o odległość promienia swojego epicyklu. (Oczywiście nakłada to dodatkowe ograniczenie na postulowaną wielkość epicykli, ale ufamy, że Ptolemeuszowi udało się pozostać w zgodzie z wszystkimi obserwacyjnymi faktami.) Natomiast problem planet zewnętrznych został rozwiązany przez przyjęcie postulatu, iż promień epicyklu każdej z planet musi być równoległy do promienia łączącego Ziemię ze Słońcem (rys. 1.5). Wtedy na jeden pełny obrót Słońca wokół Ziemi przypada jeden obrót epicyklu, a zatem także jeden cykl retrogradacji.[3]

Podsumowując, możemy zauważyć, że model Ptolemeusza składa się z wielu niezależnych od siebie postulatów. Na podstawowe założenie o krążeniu planet i Słońca wokół Ziemi nakłada się szereg dodatkowych hipotez. Każda z nich wymuszona jest przez jedno konkretne zjawisko w celu jego wyjaśnienia. Poniższa tabelka przedstawia te hipotezy wraz z odpowiadającym im zjawiskom astronomicznym. Hipotezę powołaną do wyjaśnienia tylko jednego specyficznego faktu określa się często mianem *ad hoc* (dosłownie: „do tego"). Powszechnie

[3] W istocie dane obserwacyjne pokazują, że okres między kolejnymi retrogradacjami planet zewnętrznych jest nieco dłuższy od roku słonecznego. W modelu Ptolemejskim wynika to z tego, że w ciągu roku słonecznego planeta zewnętrzna przesunie się po swoim deferensie, a zatem okres ruchu retrogradacyjnego wypadnie po okresie odrobinie dłuższym od okresu obiegu po jej epicyklu.

uważa się, że obecność postulatów i wyjaśnień *ad hoc* w nauce jest oznaką słabości danej teorii. Dobre teorie powinny unikać „ręcznego" wprowadzania założeń po to jedynie, żeby się „zgadzało". Ideałem nauki jest poszukiwanie uniwersalnych praw i twierdzeń, które będą w stanie wyjaśnić szereg niezależnych od siebie faktów.



Rys. 1.5. Wyjaśnienie obserwowanego zachowania planet wewnętrznych (lewy diagram) i zewnętrznych (prawy diagram)

Hipoteza (element modelu Ptolemeusza)	**Wyjaśnione zjawisko**
Epicykl	Ruch retrogradacyjny
Ekscentryk	Zmienność jasności planet
Ekwant	Zmienność prędkości planet
Położenie środka epicyklu na linii Ziemia-Słońce	Bliskość planet wewnętrznych w stosunku do Słońca
Równoległość promieni epicykli i linii Ziemia-Słońce	Częstotliwość retrogradacji dla planet zewnętrznych (raz do roku)

Tab. 1.1. Elementy modelu Ptolemeusza i ich rola

Można zadać pytanie, dlaczego powinniśmy preferować teorie opierające się na małej liczbie niezależnych hipotez o szerszym zakresie stosowalności. Z pewnością odgrywają tutaj rolę trudno uchwytne kwestie elegancji, prostoty czy też ogólnie estetyki. Wielość hipotez, z których każda ma niewielką moc wyjaśniającą ograniczoną do wąskiego typu zjawisk, sprawia wrażenie chaosu i nieporządku, w przeciwieństwie do np. jednej hipotezy wyjaśniającej cały szereg różnorodnych zjawisk. Istnieje jednak argument odwołujący się do naczelnego celu nauki, jakim jest dążenie do prawdy. Hipotezy w nauce oceniamy pod kątem ich prawdopodobieństwa, czyli tego, jak duża jest szansa na to, że trafnie oddają one rzeczywistość. Z elementarnego rachunku prawdopodobieństwa wiemy,

że prawdopodobieństwo łącznego zajścia niezależnych zdarzeń jest iloczynem ich prawdopodobieństw. Zatem jeśli teoria liczy sobie wiele niezależnych hipotez, to prawdopodobieństwo ich łącznej prawdziwości będzie iloczynem poszczególnych prawdopodobieństw, a zatem będzie dużo mniejsze niż prawdopodobieństwa poszczególnych hipotez. Mniejsza liczba niezależnych hipotez oznacza więc większą szansę ich prawdziwości.

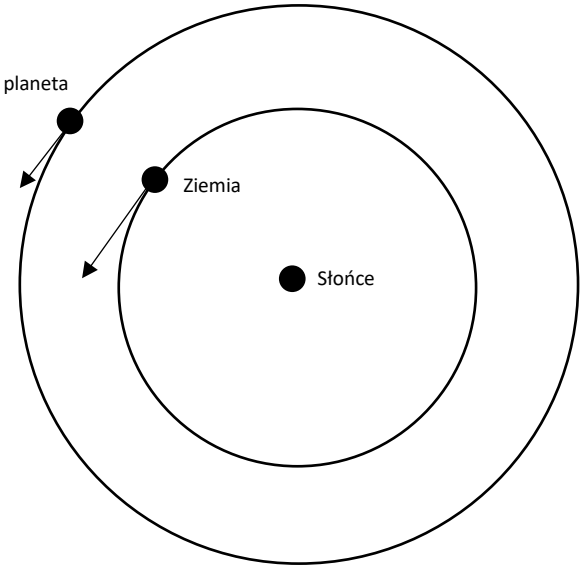
1.2. Teoria heliocentryczna Kopernika

Przez stulecia dzielące Ptolemeusza od Kopernika model greckiego uczonego był stale uzupełniany i poprawiany. Wraz z napływem nowych, bardziej dokładnych danych empirycznych dotyczących położeń ciał niebieskich (otrzymywanych dzięki udoskonalaniu technik obserwacyjnych i instrumentów) pierwotny model Ptolemeusza coraz bardziej odstawał od rzeczywistości. Rozwiązaniem pojawiających się niezgodności między teorią a doświadczeniem było wielokrotne stosowanie metody, która tak dobrze sprawdziła się w oryginalnym modelu: składania ruchów po niezależnych okręgach (epicyklach). Wprowadzając coraz to nowe kółka na kółkach, uczeni byli w stanie z niemałym trudem utrzymywać jaką-taką zgodność z danymi. Jednakże stopień komplikacji modelu, znaczny na samym jego początku, urósł do gargantuicznych rozmiarów. Stało się jasne, że potrzeba nowych, radykalnych rozwiązań.

Rozwiązanie zaproponowane przez Kopernika nie było zupełną nowością. Już w starożytności pojawiła się hipoteza, że Słońce zajmuje centralne miejsce, a Ziemia wraz z innymi planetami krąży wokół tego centrum. Ptolemeusz poświęcił nawet cały rozdział swojego dzieła *Almagest* na omówienie i skrytykowanie tej koncepcji. Podstawowym argumentem przeciwko heliocentryzmowi był brak obserwowanej zmienności względnych odległości między gwiazdami, która powinna być widoczna z powodu ruchu Ziemi (tak zwany problem paralaksy gwiezdnej). Wrócimy jeszcze do tej kwestii po omówieniu koncepcji Kopernika. Na razie przyjrzyjmy się niewątpliwym i natychmiastowym korzyściom, jakie niesie za sobą założenie o ruchu obiegowym Ziemi wokół Słońca.

Zasadnicze tezy modelu Kopernikańskiego są powszechnie znane. W podstawowej, uproszczonej wersji, model ten zakłada, że każda planeta porusza się po pewnym okręgu ze Słońcem w centrum, w następującej kolejności licząc od Słońca: Merkury, Wenus, Ziemia, Mars, Jowisz, Saturn. Jedynie Księżyc pozostaje w ruchu wokół Ziemi, zarazem razem z Ziemią uczestnicząc w obiegu dookoła Słońca. Dodatkowo Ziemia wykonuje ruch obrotowy wokół własnej osi, co tłumaczy obserwowany obrót dobowy całego firmamentu niebieskiego. Jednakże szkolna wersja modelu Kopernika nie uwzględnia faktu, że uczony z Fromborka nadal stosował epicykle i ekscentryki w celu lepszego dopasowania modelu do danych. (Liczba kół stosowanych w oryginalnym modelu Ptolemeusza i w modelu Kopernika była w przybliżeniu taka sama.) W przeciwieństwie do oryginalnego modelu Ptolemeusza model Kopernikański stosował epicykle nie w celu odtworzenia ruchów retrogradacyjnych planet, ale dla zapewnienia lepszej zgodności z danymi. Wynikało to z zasadniczo niepoprawnego założenia o kolistości orbit. Jak dzisiaj dobrze wiemy, orbity planetarne nie są okręgami, a elipsami. Okazuje się, że kształt elipsoidalny można odtworzyć, nakładając na siebie odpowiednio dobrane ruchy po deferensie i epicyklu. Wyobraźmy sobie ciało orbitujące po wielkim kole, którego środek z kolei wykonuje jeden mały obrót w przeciwnym kierunku. Łatwo

się przekonać, że otrzymamy w ten sposób wydłużoną orbitę przypominającą elipsę. Taką właśnie metodę przybliżania elips okręgami zastosował nieświadomie Kopernik, zachowując tym samym ideę doskonałości ruchów ciał niebieskich.

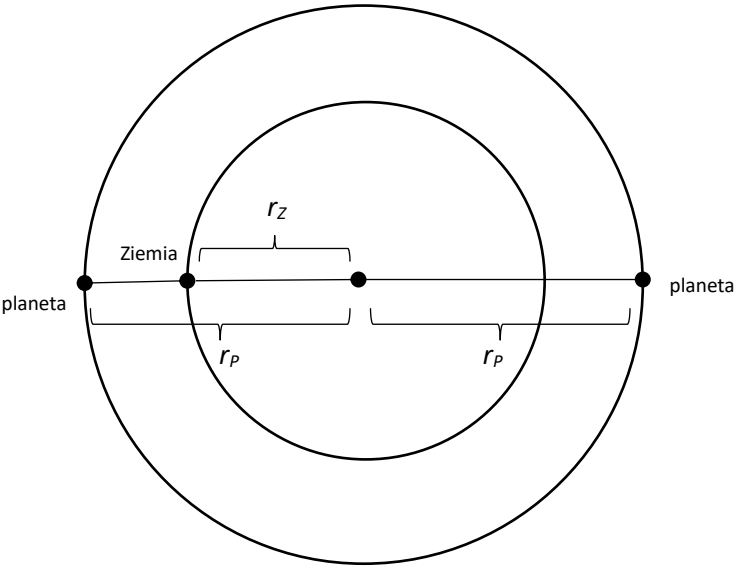


Rys. 1.6. Retrogradacja w modelu Kopernikańskim. Ziemia wyprzedza planetę w okresie największego zbliżenia, co daje efekt „cofania"

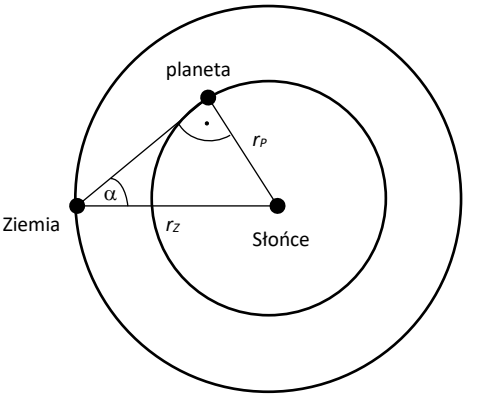
Pomińmy jednak ten szczegół oryginalnej konstrukcji Kopernika i omówmy pokrótce, jak jego model odtwarza znane dane empiryczne. Wyjaśnieniem dla większości obserwacyjnych faktów, które sprawiały niemałą trudność astronomom starożytnym, jest w systemie heliocentrycznym złożenie dwóch ruchów: „prawdziwego" ruchu danej planety oraz ruchu Ziemi i związanych z nią obserwatorów. Ruch retrogradacyjny danej planety na tle gwiazd stałych powstaje wtedy, kiedy Ziemia w swoim obiegu zaczyna „wyprzedzać" tę planetę (rys. 1.6). Jest to efekt analogiczny do efektu obserwowanego podczas wyprzedzania poruszającego się wolniej pojazdu, kiedy ten zdaje się cofać względem otoczenia. Nietrudno zauważyć, że wyprzedzanie planet zewnętrznych może odbyć się tylko raz w ciągu całego obiegu Ziemi dookoła Słońca, co automatycznie tłumaczy znaną prawidłowość bez konieczności wprowadzania dodatkowych założeń typu równoległości promienia epicykla i promienia deferensu Słońca. Z kolei w wypadku planet wewnętrznych efekt retrogradacyjny pojawia się raz w ciągu ich roku słonecznego, który jest krótszy od ziemskiego, a zatem obserwowana częstotliwość retrogradacji jest większa niż dla planet zewnętrznych.

Proste jest również wyjaśnienie zmienności jasności planet oraz ich globalnej obserwowanej prędkości. Bierze się ona z różnicy odległości między daną planetą a Ziemią. Dla planet zewnętrznych maksymalna odległość od Ziemi to $r_Z + r_P$, *gdzie* r_Z – promień orbity Ziemi, r_P – promień orbity danej planety. Minimalna odległość natomiast to $r_P - r_Z$ (rys. 1.7). Dodatkowo od razu mamy wyjaśnienie faktu, że minimalna jasność planety i minimalna prędkość kątowa przypadają na okres, kiedy znajduje się ona pozornie blisko Słońca (w „koniunkcji"), natomiast maksymalne wartości są wtedy, kiedy jest ona w opozycji. U Ptolemeusza trzeba było wprowadzić to ręcznie, odpowiednio ustawiając ekscentryki i ekwanty danej pla-

nety względem Słońca. Na koniec wspomnijmy, że pozostawanie planet wewnętrznych w pobliżu Słońca jest natychmiastowo tłumaczone tym, że ich orbity są w całości zawarte wewnątrz orbity ziemskiej (rys. 1.8).



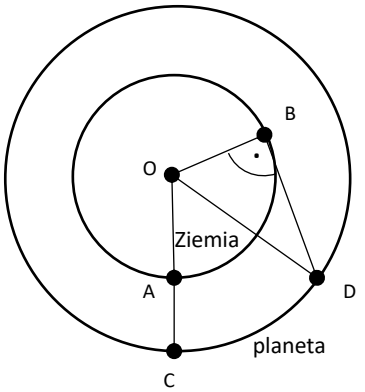
Rys. 1.7. Wyjaśnienie zmienności jasności planet zewnętrznych w systemie Kopernikańskim



Rys. 1.8. Wyjaśnienie bliskości planety wewnętrznej w stosunku do Słońca oraz obliczenie jej promienia orbity. Kąt α wyznaczamy podczas maksymalnej separacji planety i Słońca. Promień r_P będzie równy r_Z razy sinus kąta α

Mamy więc bezdyskusyjne korzyści teoretyczne modelu Kopernikańskiego: jest on zdecydowanie prostszy od Ptolemejskiego, nawet jeśli uwzględnimy obecność dodatkowych epicykli. Prostota ta ujawnia się nie tylko w mniejszej liczbie dodatkowych założeń, ale przede wszystkim w mechanizmie wyjaśniania. Jedno założenie (ruch Ziemi wokół Słońca) jest w stanie wyjaśnić wiele pozornie niezwiązanych ze sobą faktów, bez konieczności przyj-

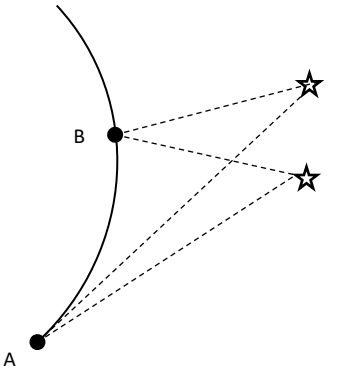
mowania rozstrzygnięć *ad hoc*. Mówimy w takiej sytuacji o unifikacyjnej roli danej hipotezy, gdyż podaje ona jeden zasadniczy mechanizm powstawania różnorodnych efektów (ich wspólną przyczynę). Koncepcja Kopernika miała jeszcze inne walory teoretyczne w stosunku do modelu Ptolemejskiego. Na przykład w modelu heliocentrycznym możliwe jest określenie promieni orbit poszczególnych planet za pomocą prostych obserwacji astronomicznych. W przypadku planet wewnętrznych określenie to opiera się na wyznaczeniu maksymalnej separacji kątowej danej planety od Słońca i rozwiązaniu odpowiedniego trójkąta prostokątnego. Dla planet zewnętrznych postępowanie jest nieco bardziej skomplikowane, gdyż wymaga dokonania dwóch obserwacji: jednej podczas opozycji danej planety względem Słońca, a drugiej, kiedy separacja kątowa planety i Słońca jest równa 90°. Szczegóły rozumowań geometrycznych w obu przypadkach podane są na diagramach (rys. 1.8 i 1.9). Model Ptolemeusza nie dawał takiej teoretycznej możliwości poza szacowaniem wielkości orbit za pomocą okresów obiegu (było to oczywiście z zasadniczych powodów niepoprawne, gdyż prędkości orbitalne poszczególnych planet nie są takie same).



Rys. 1.9. Wyznaczenie promienia orbity planety zewnętrznej. Mierzymy czas przejścia *t* planety zewnętrznej od położenia opozycji względem Słońca (punkt C) do momentu, w którym separacja kątowa między planetą a Słońcem wynosi 90° (punkt D). Ze stosunku *t* do odpowiednich okresów obiegu planety i Ziemi wyznaczamy kąty ∠COD i ∠AOB, a następnie ∠DOB, co pozwala nam na rozwiązanie trójkąta prostokątnego DOB

Należy jednak podkreślić, że nowa teoria nie była pozbawiona wad. Rozważając ją z punktu widzenia aktualnego w czasach Kopernika stanu wiedzy, możemy zrozumieć, dlaczego wielu ówczesnych naukowców podchodziło sceptycznie do proponowanego rozwiązania. Przede wszystkim uważano, że ruch Ziemi, zarówno obrotowy dookoła własnej osi, jak i obiegowy wokół Słońca, powinien być zauważalny dla obserwatorów znajdujących się na jej powierzchni. Szczególnie w przypadku ruchu obrotowego spodziewano się wystąpienia znacznych efektów. Wyobrażano sobie, że podobnie jak na kręcącej się karuzeli, byłoby nam niezwykle trudno ustać na nogach, gdyby Ziemia obracała się z niewiarygodną prędkością (która na równiku wynosi ok. 1600 km/h). Do tego odczuwalibyśmy ogromny pęd powietrza, przewracający drzewa i budowle. Spadające przedmioty powinny odchylać się od pionu w kierunku zachodnim, ze względu na „ucieczkę" powierzchni Ziemi w czasie spadku. Problemy te zostały zupełnie zignorowane przez Kopernika. Dopiero włoski uczony Galileo Galilei (Galileusz) znalazł właściwą odpowiedź na problem braku obserwowalnych efektów

ruchu obrotowego Ziemi, ale wymagała ona odrzucenia uznawanej za czasów Kopernika fizyki Arystotelesowskiej. Będziemy o tym szerzej mówić w następnych paragrafach.

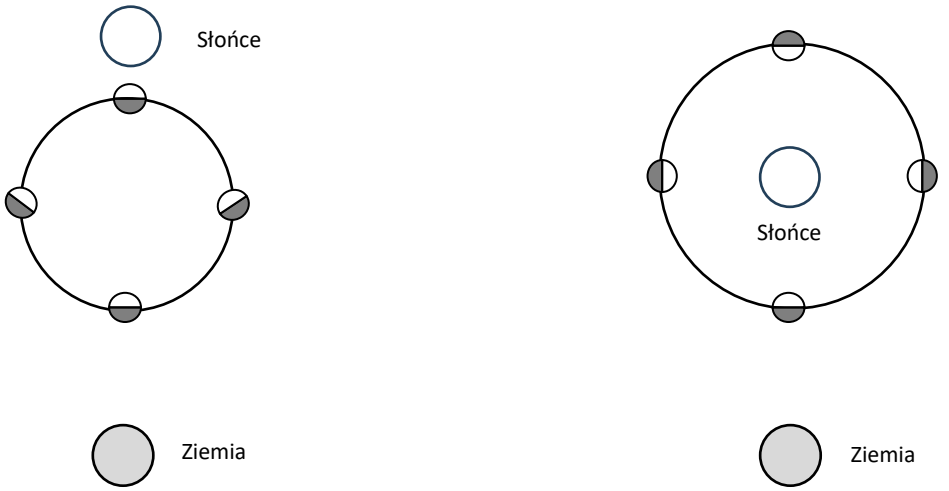


Rys. 1.10. Efekt paralaksy. W punkcie A gwiazdy są widoczne pod mniejszym kątem niż w punkcie B, zatem ich relatywne położenie na niebie ulega zmianie

Z kolei ruch obiegowy Ziemi wokół Słońca powinien być obserwowalny za pomocą efektu paralaksy gwiazdowej. Ogólnie paralaksa jest zjawiskiem polegającym na zmianie pozornej odległości między przedmiotami z perspektywy obserwatora w ruchu. Na przykład obserwując dwa drzewa z poruszającego się samochodu, zauważymy, że z jednego punktu drogi wydają się one bardzo blisko siebie, a z innego ich separacja kątowa jest większa. Efekt ten jest tym większy, im bliżej nas znajdują się obserwowane przedmioty. Wydawało się oczywiste, że zmiana położenia Ziemi wskutek ruchu rocznego powinna umożliwić nam obserwację paralaksy gwiazd na sferze niebieskiej (rys. 1.10). Jednakże żaden taki efekt nie został stwierdzony. Odpowiedź Kopernika na ten zarzut była bardzo prosta: najwidoczniej gwiazdy znajdują się zbyt daleko, aby efekt paralaksy był zauważalny. Jednakże zwróćmy uwagę, że Kopernik nie miał żadnych danych niezależnie potwierdzających jego hipotezę. W modelu Ptolemeusza gwiazdy stałe znajdowały się co prawda najdalej od Ziemi, ale na pewno nie w odległości, która uniemożliwiałaby zaobserwowanie efektu paralaksy, gdyby Ziemia naprawdę poruszała się po orbicie okołosłonecznej. Rozwiązanie Kopernika było zatem ewidentnie *ad hoc*. Nie ma przy tym znaczenia, że jego intuicja okazała się trafna – w świetle dostępnych mu danych była to czysta spekulacja, sformułowana wyłącznie po to, aby pozbyć się problemu. Na marginesie dodajmy, że efekt paralaksy dla gwiazd najbliższych w stosunku do Układu Słonecznego został zaobserwowany, ale dopiero w dziewiętnastym wieku. Dla większości gwiazd nawet najdoskonalsze instrumenty nie są w stanie wykryć ich pozornych zmian położenia, ze względu na ogromną odległość dzielącą nas od tych gwiazd. Zatem teoria Kopernika zyskała powszechną akceptację, na długo zanim problem paralaksy znalazł zadowalające rozwiązanie.

Sytuacja nowej teorii heliocentrycznej byłaby dużo klarowniejsza, gdyby istniało jej empiryczne potwierdzenie, jednocześnie podważające konkurencyjną teorię geocentryczną Ptolemeusza. Na możliwość takiego empirycznego faktu wskazał Kopernik. Teoria Kopernika

przewiduje, że planety wewnętrzne powinny przechodzić przez cztery fazy, podobne do faz księżycowych: nów, pierwszą kwadrę, pełnię i drugą kwadrę. Fazy te są rezultatem różnego ustawienia planety względem Słońca: nów występuje, kiedy planeta jest mniej więcej na linii łączącej Słońce z Ziemią, kwadry powstają, gdy planeta jest maksymalnie kątowo odseparowana od Słońca, a pełnia, kiedy planeta zachodzi za Słońce. Natomiast w modelu ptolemejskim mamy tylko trzy fazy: nów, pierwszą kwadrę, nów, drugą kwadrę, gdyż planety wewnętrzne nigdy nie znajdują się dalej od Ziemi niż Słońce (rys. 1.11). Weryfikacja empiryczna obu przewidywań, np. w wypadku planety Wenus, pozwoliłaby na rozstrzygnięcie, którą z teorii należy odrzucić. Niestety za czasów Kopernika dokładna obserwacja kształtu planet była niemożliwa. Sytuacja zmieniła się wraz z wprowadzeniem teleskopów. Pionierem wykorzystania teleskopu do obserwacji astronomicznych był Galileusz. On też poddał obserwacji fazy Wenus i jednoznacznie stwierdził istnienie fazy pełni, przechylając tym samym szalę na korzyść teorii Kopernikańskiej.



Rys. 1.11. Porównanie faz planety wewnętrznej w modelu geocentrycznym (po lewej) i heliocentrycznym (po prawej)

1.3. Filozoficzne aspekty przewrotu Kopernikańskiego

Przełom, jaki dokonał się w astronomii w szesnastym wieku, dostarcza bogatego materiału do refleksji filozoficznej. W niniejszym paragrafie wymienimy trzy zagadnienia filozoficzne inspirowane przewrotem Kopernikańskim. Są to: spór pomiędzy instrumentalistyczną a realistyczną interpretacją teorii, zagadnienie porównywania konkurencyjnych teorii w nauce oraz pojęcie rewolucji naukowych. Zaczynając od pierwszego z nich, przytoczmy może fragmenty ze wstępu do dzieła Kopernika *De Revolutionibus Orbium Coelestium* pióra protestanckiego teologa Andreasa Osiandra:

> Hipotezy astronomiczne „nie potrzebują […] być prawdziwe ani nawet zbliżone do prawdy, lecz wystarczy to jedno, że dają obliczenia zgadzające się z obserwacjami.”

Astronomia „nie zna dogłębnie i jasno przyczyn obserwowanych ruchów nieregularnych."

„Do jednego i tego samego ruchu różne się nieraz nastręczają hipotezy, jak np. przy ruchu Słońca mimośrodowość i epicykle. Astronom tej się przede wszystkim chwyta, która do zrozumienia jest najłatwiejsza."

Stanowisko Osiandra jest dzisiaj powszechnie znane jako instrumentalizm. Głosi ono, że podstawowym celem nauki nie jest odkrywanie prawd na temat rzeczywistości, ale zapewnienie zgodności z obserwacyjnymi danymi (tzw. „zachowanie zjawisk"). Nie ma sensu pytać, czy planety „naprawdę" poruszają się po orbitach okołosłonecznych, czy też okrążają Ziemię po deferensach i epicyklach. Ważne jest tylko to, która z konkurencyjnych teorii lepiej zgadza się z empirią lub też posiada inne walory, takie jak prostota. Współczesny zwolennik instrumentalizmu, Bas van Fraassen, wprowadził pojęcie „empirycznej adekwatności". Zamiast pytać, czy dana teoria naukowa jest prawdziwa, powinniśmy skupić się na ocenie tego, czy jest ona empirycznie adekwatna, czyli czy odtwarza możliwe do zaobserwowania fakty. Empiryczna adekwatność teorii oznacza, że jej część zawierająca wyłącznie terminy obserwacyjne jest prawdziwa. Pozostała część, obejmująca tzw. terminy teoretyczne, nie musi być natomiast zgodna z rzeczywistością.

Motywacja Osiandra była oczywiście teologiczna. Jako zagorzały zwolennik protestantyzmu uważał Biblię za jedyne źródło objawionej wiedzy prawdziwej. Skoro w Biblii nie ma żadnej wzmianki na temat konkretnego modelu astronomicznego (nie pojawiają się w niej pojęcia epicykli, deferensów czy kopernikańskie orbity okołosłoneczne), nie możemy swoimi skończonymi umysłami rozstrzygnąć kwestii prawdziwości tych modeli. Natomiast naturalistyczne wersje instrumentalizmu kładą przede wszystkim nacisk na zasadniczą nierozstrzygalność teoretycznych założeń modeli w nauce. W wypadku modeli astronomicznych nie jesteśmy w stanie wznieść się ponad Układ Słoneczny i z tej perspektywy ocenić, czy planety rzeczywiście orbitują naokoło Ziemi po deferensach i epicyklach, czy też, jak chce tego Kopernik, okrążają Słońce wraz Ziemią. Pozostaje nam jedynie analiza dostępnych danych na temat obserwowanego położenia ciał niebieskich. W wypadku innych teorii naukowych taką empirycznie niedostępną sferą może być według instrumentalistów dziedzina nieobserwowalnych przedmiotów, takich jak molekuły, atomy, geny czy cząstki elementarne i ich własności. Obiekty te postuluje się w celu wyjaśnienia złożoności obserwowanych zjawisk fizykochemicznych czy biologicznych, ale one same nie są dostępne naszej bezpośredniej obserwacji ze względu na ich zbyt małe rozmiary.

Przeciwne stanowisko w stosunku do instrumentalizmu zajmuje realizm naukowy. Realiści podkreślają, że głównym celem nauki jest i powinno być odkrywanie, jak się rzeczy naprawdę mają. Powołują się oni przy tym na praktykę naukowców, którzy niezmiernie rzadko przyjmują sceptyczną czy agnostyczną postawę w stosunku do teoretycznych założeń danej teorii. Zdarza się, że naukowcy jawnie wprowadzają pewne pomocnicze pojęcie jako instrument, ułatwiający czy upraszczający opis zjawisk. Do takich pojęć należy np. pojęcie dziur elektronowych, używane w teorii półprzewodników na oznaczenie braku elektronu w danym paśmie energetycznym półprzewodnika. Okazuje się, że wygodnie jest opisać przewodnictwo prądu tak, jakby występowały w nim dodatnio naładowane cząstki (dziury). W ostatecznym rachunku istnieją jednakże tylko elektrony. Status samych elektronów czy innych cząstek jest zupełnie inny niż instrumentalistycznie traktowanych dziur – elektronom przypisuje się obiektywne, niezależne istnienie.

Realiści uważają, że instrumentaliści przyjmują zbyt restrykcyjne epistemiczne standardy prawdziwości zdań w nauce. Dla instrumentalisty jedynie bezpośrednie zmysłowe postrzeganie jest gwarantem prawdziwości odpowiednich twierdzeń. Jednakże pojęcie bezpośredniej obserwacji jest samo w sobie nieostre. Czy obraz fatamorgany, zaburzony gorącym powietrzem na pustyni, jest bezpośrednim postrzeganiem zapewniającym prawdziwość odpowiedniego zdania postrzeżeniowego (w rodzaju zdania „Przede mną znajduje się oaza")? Aby wykluczyć takie przypadki, musimy tak zaostrzyć kryteria prawdziwości, że w rezultacie otrzymamy konsekwencję sceptycką – nic nie może być z całą pewnością znane. Realiści uważają, że jest to postawienie sprawy na głowie. Za prawdziwością danej hipotezy mogą przemawiać świadectwa pośrednie. Jeśli dana teoria dobrze oddaje empiryczne fakty, jest to argument za tym, że jest ona prawdziwa. Realiści odwołują się do tzw. argumentu z najlepszego wyjaśnienia. Najlepszym wyjaśnieniem sukcesu empirycznego danej teorii jest po prostu jej prawdziwość. Sukcesy empiryczne teorii Kopernika powinny nas przekonać, że Kopernik był bliżej prawdy niż Ptolemeusz.

Możemy teraz przejść do drugiego zagadnienia filozoficznego, jakim jest kwestia porównywania konkurencyjnych teorii. Jeśli mamy dwie niezgodne ze sobą teorie, które wyjaśniają podobny zestaw zjawisk, na jakiej podstawie powinniśmy dokonać wyboru jednej z nich? Ogólnie możemy wyróżnić dwa aspekty oceny teorii naukowych: aspekt teoretyczny i empiryczny. Do oceny teoretycznej zaliczamy takie cechy danej teorii jak elegancja, prostota, zakres stosowalności, sposób wyjaśniania znanych zjawisk oraz przewidywania nowych. Jak już zauważyliśmy, porównanie pod tym względem teorii Ptolemeusza i Kopernika jednoznacznie wskazuje na przewagę tej ostatniej. Prostota i elegancja oferowanych w teorii heliocentrycznej wyjaśnień jest bezsporna. Również szerszy zakres przewidywań, obejmujący takie parametry jak promienie orbit planetarnych, jest warty wzmianki. Jednakże powszechnie przyjmuje się, że najważniejszym aspektem porównania teorii jest ich testowanie empiryczne. Ideałem takiego testowania jest tzw. eksperyment krzyżowy (*experimentum crucis*), który zasadniczo powinien umożliwić definitywny wybór jednej spośród wielu konkurencyjnych teorii.

Struktura eksperymentu krzyżowego jest bardzo prosta. Dla dwóch teorii T_1 i T_2 musimy znaleźć pewne zdanie empiryczne *e* takie, że *e* wynika z T_1, natomiast jego negacja wynika z T_2. Następnie powinniśmy empirycznie stwierdzić zachodzenie bądź niezachodzenie faktu opisywanego przez *e*. W ten sposób fałszywość zdania *e* pozwala nam na wybór teorii T_2, a prawdziwość *e* – na wybór teorii T_1. Przykładem możliwego *experimentum crucis* dla przypadku teorii astronomicznych jest kwestia istnienia czterech faz planet wewnętrznych. Twierdzenie o istnieniu czterech faz wynika z teorii Kopernikańskiej, natomiast jego negacja – z teorii Ptolemeusza (gdyż przewiduje ona istnienie trzech faz). Dokonując odpowiedniej obserwacji np. faz planety Wenus, wybieramy właściwą teorię, którą okazuje się teoria Kopernika.

Potencjalnym eksperymentem krzyżowym dla obu teorii astronomicznych mogłoby być także zjawisko paralaksy gwiazdowej. Tu jednak ujawnia się problem związany z dodatkowymi milcząco przyjętymi założeniami. Na pozór wydaje się, że *experimentum crucis* oparte na zjawisku paralaksy sugeruje wybór teorii Ptolemeuszowskiej. Teoria heliocentryczna Kopernika implikuje istnienie paralaksy, natomiast geocentryczna jej brak. Ponieważ nie obserwujemy pozornego przemieszczania się gwiazd względem siebie w ciągu roku słonecznego, przemawia to przeciwko teorii Kopernika, a za teorią Ptolemeusza. Jednakże problemem jest to, że teoria heliocentryczna implikuje istnienie obserwowalnego efektu paralaksy tylko przy

dodatkowym założeniu, że gwiazdy stałe znajdują się wystarczająco blisko nas. Ponieważ nie wiemy, czy tak w istocie jest, eksperyment krzyżowy w tym wypadku nie daje jednoznacznego rozstrzygnięcia.

Wielu filozofów uważa, że problem dodatkowych założeń dotyka w istocie każdego przypadku empirycznego wyboru między teoriami. Nawet wydawałoby się niebudzący wątpliwości przypadek faz planet wewnętrznych opiera się na dodatkowych założeniach, takich jak prawa optyki geometrycznej (rozchodzenie się promieni świetlnych po liniach prostych, prawo odbicia itp.). Przy odrobinie inwencji zawsze można wymyślić argument, który podważy rezultat danego eksperymentu krzyżowego. Francuski fizyk i filozof Pierre Duhem sformułował naczelne hasło wyznawanego przez siebie poglądu, zwanego konwencjonalizmem, w postaci tezy „Nie istnieje *experimentum crucis*". Konwencjonaliści uważają, że w ostatecznym rachunku o wyborze danej teorii decydują nie tyle względy empiryczne, co wcześniej wspomniane zalety teoretyczne: prostota, elegancja itp.[4]

Zastąpienie teorii Ptolemeusza teorią Kopernikańską rzuca nowe światło na problem rozwoju nauki. Istnieje pogląd, zgodnie z którym nauka rozwija się stopniowo i liniowo przez gromadzenie coraz to większej liczby faktów, danych i wreszcie uogólnień. Taki model rozwoju nauki nazywa się często kumulatywnym, gdyż zakłada się w nim, że postęp naukowy dokonuje się poprzez poszerzanie istniejącego już zasobu wiedzy. Natomiast zmiany, jakie dokonały się w astronomii szesnastego wieku sugerują inny model. Kopernik nie uzupełniał teorii Ptolemeusza, ale ją całkowicie odrzucił i zastąpił zupełnie odmiennym podejściem, opartym na nowych pojęciach i założeniach. Thomas Kuhn, filozof i historyk nauki dwudziestego wieku, określa tego typu zmiany w nauce mianem „rewolucji naukowej". Według niego historię rozwoju każdej nauki można podzielić na dwa rodzaje okresów: tzw. okresy nauki normalnej oraz sytuacje kryzysowe, skutkujące gwałtownymi, rewolucyjnymi przemianami. Nauka normalna charakteryzuje się stopniowym rozwojem, polegającym na zbieraniu danych i ulepszaniu istniejących modeli i teorii oraz rozwiązywaniu pojawiających się trudności i problemów. W astronomii etap normalnego rozwoju przypada na stulecia dzielące Ptolemeusza od astronomii nowożytnej. W okresie tym model Ptolemeusza był poddawany zmianom i korektom, ale jego zasadniczy trzon pozostawał taki sam – opierał się na założeniu geocentryzmu i powszechnym stosowaniu deferensów i epicykli do odtworzenia obserwowanych ruchów planet.

Kryzys w astronomii Ptolemejskiej polegał na rosnącym stopniu komplikacji modelu przy jednoczesnych trudnościach z dopasowaniem do coraz dokładniejszych danych empirycznych. Kuhn zwraca uwagę, że kiedy sytuacja kryzysowa w danej dziedzinie osiąga punkt krytyczny, naukowcy dokonują gwałtownej przemiany stosowanych metod i tworzą nową teorię. Zmianę taką opisuje przy pomocy pojęcia paradygmatu. Obowiązujący w danym momencie paradygmat to zespół przyjętych założeń i metod rozwiązywania problemów, które nie są poddawane pod dyskusję przez środowisko naukowe. Paradygmat astronomii Ptolemeusza zawierał założenie o nieruchomości Ziemi, jak również metodę odtwarzania ruchów planetarnych za pomocą złożenia ruchów kołowych. Rewolucja Kopernikańska zmieniła ten paradygmat na heliocentryczny, zastępując złożenia ruchów planetarnych kombinacją praw-

[4] Jeszcze jednym problemem *experimentum crucis* jest to, że milcząco zakłada się nieistnienie innych opcji poza rozważanymi teoriami. Może się jednak zdarzyć, że będzie istnieć wiele teorii, które mają taką samą konsekwencję empiryczną, potwierdzoną w doświadczeniu. W takiej sytuacji eksperyment krzyżowy nie daje nam możliwości wyboru między tymi teoriami. Sytuacje takie nazywa się „niedookreśleniem teorii przez dane" (*underdetermination of theory by data*).

dziwego ruchu planety i ruchu obserwatora na powierzchni Ziemi. Kontrowersyjnym elementem doktryny Kuhna było założenie o tzw. niewspółmierności paradygmatów przed i po rewolucji w nauce. Niewspółmierność oznacza, że różne paradygmaty nie mogą być nawet ze sobą porównane w jednej siatce pojęciowej. Przyjęcie nowego paradygmatu powoduje, że stary staje się nie tyle fałszywy, co pozbawiony sensu.

1.4. Prawa Keplera ruchu planetarnego

Heliocentryczna koncepcja Kopernika nie została od razu przyjęta przez środowisko naukowe. Opór przeciwko doktrynie heliocentryzmu był częściowo spowodowany kwestiami politycznymi i światopoglądowymi (wpływ nauki Kościoła), ale jak zwracaliśmy uwagę w poprzednim paragrafie, istniały również poważne zastrzeżenia o charakterze naukowym. Duński astronom Tycho de Brahe zaproponował rozwiązanie kompromisowe, łączące elementy kopernikanizmu z modelem Ptolemeusza. Założył on, w zgodzie z modelem Ptolemejskim, że Ziemia spoczywa nieruchomo w centrum wszechświata. Jednakże przyjął, że pozostałe planety orbitują nie naokoło Ziemi, a wokół Słońca, a dopiero razem ze Słońcem krążą wokół Ziemi. Dzięki temu de Brahe uniknął problemów z brakiem odczuwalnych efektów ruchu Ziemi (w tym braku paralaksy gwiezdnej). Warto może zauważyć, że przy odpowiednio dobranych orbitach planet wokół Słońca, jego model jest w zasadzie „kinematycznie" równoważny modelowi Kopernika.[5] Powstaje on przez przetransformowanie się z układu, w którym spoczywa Słońce, do układu związanego z Ziemią. Oczywiście sam de Brahe nie dysponował jeszcze pojęciem różnych układów odniesienia ani ideą względności ruchu, która pojawiła się dopiero u Galileusza.

Prawdziwy postęp w modelu heliocentrycznym dokonał się za sprawą Johannesa Keplera. Astronom ten miał prawdziwą obsesję na punkcie matematycznego opisania prawidłowości w ruchu planet. Przyjmując stanowisko Kopernikańskie, próbował odkryć matematyczne regularności w ruchach poszczególnych planet wokół Słońca. Jednym z jego pomysłów, który okazał się kompletnie chybiony, była próba zastosowania doskonałych brył Platońskich do wyrażenia proporcji między orbitami kolejnych planet. Rozważając różne alternatywne możliwości, wpadł na pomysł zastosowania tzw. krzywych stożkowych do opisu orbit planetarnych. Krzywe stożkowe powstają z przecięcia stożka płaszczyznami pod różnym kątem. Jedną z takich krzywych jest oczywiście okrąg, ale zmieniając kąt płaszczyzny w stosunku do wysokości stożka otrzymamy elipsy. (Inne krzywe stożkowe to parabola i hiperbola). Kepler w przebłysku geniuszu spróbował przyjąć, że orbity planet mają kształt nie okręgów, a elips. Okazało się to przysłowiowym „strzałem w dziesiątkę". Dzięki temu założeniu udało mu się precyzyjnie odtworzyć obserwowane dane astronomiczne bez konieczności stosowania Kopernikańskich epicykli.

Na tym jednak nie kończy się wkład Keplera w astronomię. Opisanie kształtu orbit planetarnych to dopiero początek. Kepler chciał jeszcze odkryć prawidłowości dotyczące prędkości poruszających się planet, jak również prawidłowości opisujące wielkości poszczegól-

[5] Modele te nie są natomiast dynamicznie równoważne ze względu na występowanie sił pozornych (odśrodkowych, Coriolisa) w układach obracających się. Będziemy o tym mówić w następnych paragrafach.

nych orbit. Wyniki swoich badań sformułował w postaci trzech dobrze znanych praw ruchu planetarnego, które przytoczymy we współczesnym kształcie.

1. Planety poruszają się po elipsach ze Słońcem w jednym z ognisk.
2. Promień wodzący planety (zaczepiony w Słońcu) zakreśla równe pola powierzchni w równych odcinkach czasu.
3. Stosunek kwadratu okresu obiegu do sześcianu średniego promienia jest taki sam dla wszystkich planet.

Ogniska danej elipsy to dwa punkty, takie że suma odległości dowolnego punktu na elipsie od ognisk jest zawsze taka sama. Drugie prawo Keplera wykazuje pewne podobieństwo do zasady ekwantu w modelu Ptolemejskim. U Ptolemeusza planety w równych okresach czasu zakreślały równe kąty względem ekwantu. Kepler osiągnął lepszą zgodność z doświadczeniem, zastępując kąty polem powierzchni, a punkt ekwantu Słońcem. Warto zauważyć, że drugie prawo Keplera wynika z zasady zachowania momentu pędu w mechanice klasycznej. Natomiast trzecie prawo Keplera jest rezultatem jego prób znalezienia regularności w wielkościach orbit poszczególnych planet. Niepowodzeniem zakończyły się jego wysiłki wyrażenia prawidłowości za pomocą samych promieni. Odkrył natomiast, że chociaż zasadniczo każda planeta może przyjmować orbitę o dowolnej wielkości, to jednak ustalenie promienia orbity R ustala okres jej obiegu T, zgodnie ze sformułowaną proporcją $\frac{T^2}{R^3} = \text{const}$.

Prawa Keplera uporządkowały wiedzę astronomiczną, zapewniając doskonałą zgodność z doświadczeniem nieosiągalną w poprzednich modelach. Jednakże ich wkład w rozwój nauki wykracza poza samą astronomię. Prawa Keplera umożliwiły Newtonowi odkrycie matematycznej formuły opisującej uniwersalne oddziaływania grawitacyjne, zarówno między Słońcem i planetami, jak i między dowolnymi przedmiotami obdarzonymi masą. Będziemy o tym szerzej mówić w kolejnym rozdziale.

Następną wielką postacią w rozwoju astronomii był wspomniany już kilkukrotnie Galileusz. Osiągnięcia uczonego z Pizy wykraczają daleko poza systematyzację zjawisk astronomicznych. Galileusz utorował drogę do ogromnego przełomu w nauce, jakim było zastąpienie mechaniki Arystotelesa przez mechanikę newtonowską. Jego wkład do astronomii polegał przede wszystkim na zastosowaniu instrumentu optycznego, którym był teleskop, do szczegółowych obserwacji ciał niebieskich. Dzięki temu dokonał kilku spektakularnych odkryć, posuwając naprzód naszą wiedzę i zrywając definitywnie z obrazem wszechświata ukształtowanym przez Ptolemeusza, a nawet częściowo przejętym przez Kopernika. Galileusz podważył przekonanie o doskonałości ciał niebieskich, obserwując nieregularne kratery i zaciemnione „morza" na powierzchni Księżyca. Zaobserwował i prawidłowo zidentyfikował księżyce Jowisza, których istnienia nawet nie podejrzewali jego poprzednicy. Jak już wspominaliśmy, za pomocą swojego teleskopu potwierdził istnienie czterech faz Wenus, zgodnie z przewidywaniami teorii Kopernika.[6]

[6] Warto może wspomnieć, że Galileusz położył podwaliny pod metodologię naukową, podejmując m.in. zagadnienie wiarogodności stosowania instrumentów optycznych, takich jak lunety czy teleskopy. Aby odpowiedzieć na formułowane zarzuty, że obserwacje astronomiczne za pomocą teleskopów są niewiarygodne ze względu na możliwe tworzenie artefaktów i złudzeń optycznych, zaproponował prosty test. Skierował swój przyrząd optyczny na oddalony obiekt ziemski, a następnie porównał zaobserwowany przez teleskop obraz przedmiotu z jego faktycznym wyglądem, przekonując się, że w obrazie teleskopowym nie powstają żadne istotne zniekształcenia czy zaburzenia.

Galileusz był oczywiście zagorzałym zwolennikiem heliocentrycznego systemu Kopernikańskiego, udoskonalonego przez Keplera. Chociaż nie wprowadził do tego systemu jakichś dodatkowych teoretycznych udoskonaleń, to jednak podjął się niezwykle ważnego zadania odparcia zarzutów pod adresem heliocentryzmu, opartych na braku odczuwalnych efektów ruchu Ziemi. Był to niezwykle szczęśliwy traf, gdyż dzięki swojej determinacji Galileusz dostrzegł zasadnicze błędy w postrzeganiu zjawisk kinematycznych przekazane nam od starożytności. Jego obrona Kopernika przybrała postać frontalnego ataku na zasady fizyki Arystotelesowskiej, dzięki czemu został wyznaczony właściwy kierunek dalszego rozwoju fizyki nowożytnej. Aby jednak przedstawić te sprawy nieco dokładniej, musimy cofnąć się w czasie do rozważań fizycznych słynnego filozofa ze Stagiry.

1.5. Fizyka Arystotelesa a fizyka Galileusza

Za czasów Arystotelesa fizyka ograniczała się prawie wyłącznie do mechaniki. W mechanice wyróżniamy do dzisiaj trzy działy, których nazwy wskazują na ich greckie pochodzenie. Działem mechaniki zajmującym się ogólnym opisem ruchu jest kinematyka (od gr. *kinesis* – ruch). Kinematyka klasyfikuje i opisuje ruch przy pomocy takich pojęć jak położenie, prędkość czy przyspieszenie. W ramach klasyfikacji wyróżnia się ruchy jednostajne, prostoliniowe, krzywoliniowe (w tym po okręgu), przyspieszone. Z kolei dynamika (gr. *dinamis* – siła, moc) poszukuje przyczyn ruchu w postaci działających sił. Trzecią gałęzią mechaniki jest statyka (gr. *statos* – trwały, niezmienny), opisująca przedmioty w położeniu równowagi, z naciskiem na warunki równowagi opisane w kategoriach działających sił (prawo dźwigni Archimedesa).

Dynamika Arystotelesa posługiwała się rozróżnieniem na ruchy naturalne i wymuszone. Ruchy naturalne wynikają z istoty uczestniczących w nich przedmiotów i nie wymagają przyczyny zewnętrznej (sprawczej). Mogą one być co najwyżej opisane w kategoriach przyczyn celowych. Naturalnym ruchem ciał znajdujących się w pobliżu Ziemi jest ruch pionowy w górę lub dół. Ciała ciężkie opadają w dół, gdyż ich natura jest ziemska, a więc dążą do osiągnięcia swojego naturalnego miejsca na powierzchni Ziemi (przyczyna celowa). Z kolei ciała lekkie, takie jak dym czy para, unoszą się ku górze, gdyż dominuje w nich pierwiastek niebieski. Arystoteles ponadto rozróżniał dwie sfery świata: podksiężycową (ziemską) i nadksiężycową (niebieską). W sferze nadksiężycowej naturalnymi ruchami są ruchy po okręgach – jak pamiętamy, przekonanie to było podstawą dla teorii Ptolemeuszowskiej, a nawet Kopernikańskiej.

Ruchy wymuszone to takie, które wymagają działania siły dla ich podtrzymania. Z doświadczenia wiemy, że aby poruszyć dany przedmiot – np. wóz na płaskiej powierzchni – musimy go popchnąć albo pociągnąć. Kiedy przestaniemy na niego działać siłą, wóz wraca do stanu spoczynku. Arystoteles uogólnił tę obserwację na wszystkie możliwe sytuacje ruchów nienaturalnych. Zatem podstawą dynamiki Arystotelesa było przekonanie, że ilość ruchu jest proporcjonalna do przyłożonej siły – kiedy siła jest zerowa, ruch ustaje. Oczywiście Arystoteles nie posługiwał się współczesnym pojęciem siły, które zostało wprowadzone przez Newtona, ale nie będzie chyba wielkim anachronizmem zinterpretowanie jego stanowiska przy pomocy dzisiejszych pojęć. Proponowane prawo dynamiki Arystotelesa dla ruchów wymuszonych miałoby więc dzisiaj następującą postać:

$$v \sim F,$$

gdzie F jest siłą przyłożoną do ciała, v – jego prędkością, a symbol ~ oznacza proporcjonalność. Arystoteles ponadto zdawał sobie sprawę, że na ruch przedmiotów ma wpływ opór otoczenia, w tym powietrza. Im większy opór, tym mniejsza prędkość przedmiotu przy danej sile. Oznaczają miarę oporu przez R, mamy następującą wersję prawa dynamiki Arystotelesa:

$$v \sim \frac{F}{R}.$$

Arystoteles zauważył niepokojącą konsekwencję swojego prawa – zgodnie z powyższą proporcją, kiedy opór otoczenia maleje do zera, prędkość powinna rosnąć do nieskończoności. Jest to jawnie niezgodne z doświadczeniem, a zatem trzeba dokonać jakiejś poprawki. Filozof wyciągnął wniosek, że wyeliminowanie oporu jest z zasadniczych powodów niemożliwe. W szczególności niemożliwe jest całkowite usunięcie powietrza z danego obszaru. Arystoteles uważał, że próżnia nie może istnieć (stąd wzięło się średniowieczne określenie *horror vacui* – strach przed próżnią). Ciała niebieskie poruszają się w doskonałej eterycznej substancji (tzw. piąta substancja – *quinta essentia*).

Zauważmy, że gdyby Arystoteles posługiwał się współczesną algebrą zamiast zwykłych proporcji, mógłby bez trudu poradzić sobie z problemem „ucieczki do nieskończoności". Wystarczyłoby w tym celu dokonać następującej modyfikacji powyższego wzoru:

$$v \sim \frac{F}{R + c},$$

gdzie c jest pewną stałą. W takiej sytuacji prędkość ciała w ośrodku pozbawionym oporów (np. próżni) dążyłaby do wartości maksymalnej, ale skończonej: $\frac{F}{c}$. Niestety upłynęły stulecia, zanim rozwój matematyki umożliwił stosowanie podobnych chwytów.

Arystoteles jako krytyczny obserwator zdawał sobie sprawę z innej poważnej trudności swojej teorii. Był to problem pocisków: rzucanych kamieni, strzał wystrzeliwanych z łuku czy pocisków wyrzucanych z katapult. We wszystkich tych przypadkach siła działająca na pocisk ustaje z chwilą wystrzelenia, a jednak kontynuuje on swój lot do momentu trafienia w cel lub też upadku na ziemię. Zatem wydaje się, że możliwy jest ruch bez podtrzymującej siły. W celu poradzenia sobie z tą trudnością, Arystoteles wymyślił następujące rozwiązanie. Ciało znajdujące się w locie oddziałuje z powietrzem, a zatem być może to powietrze podtrzymuje je w ruchu. Rozważając dokładniej ten problem, Arystoteles zasugerował, że pocisk „wypycha" w locie powietrze i tworzy za sobą chwilową próżnię. Jednakże, jak już wiemy, przyroda nie toleruje próżni, więc w to miejsce natychmiast napływa powietrze, które wywiera nacisk na ciało, utrzymując je w locie.

Nie sposób odmówić pomysłowości Arystotelesowi, chociaż widać też luki w jego podejściu. Jego koncepcja jest niekonsekwentna – z jednej strony obecność powietrza tworzy opory ruchu, zmniejszając prędkość pocisku, ale z drugiej wywiera siłę, podtrzymując pocisk w ruchu. Zaczekajmy jednak z krytyką teorii Arystotelesa do przenikliwego argumentu Galileusza. Na razie wspomnijmy jeszcze o jego teorii spadku swobodnego. Jak już zauważyliśmy, spadek swobodny jest ruchem naturalnym, uzależnionym od ilości pierwiastka ziemskiego w danym ciele. Ten pierwiastek ziemski ma prostą interpretację ilościową w postaci ciężaru. Nic więc dziwnego, że Arystoteles przyjął bez wahania prawo spadku

swobodnego, zgodnie z którym ciała cięższe spadają szybciej niż lekkie.[7] Innymi słowy, zachodzi proporcjonalność między czasem spadku z danej wysokości a ciężarem ciała. Jak się wydaje, doświadczenie potwierdza taką zasadę – wystarczy porównać spadek kamienia i piórka ptasiego.

Galileusz poddał krytycznej analizie podstawowe tezy fizyki Arystotelesa. Charakterystyczne, że choć sam Galileusz głosił konieczność poddawania doświadczalnej weryfikacji wszystkich twierdzeń na temat rzeczywistości, nawet tych najbardziej oczywistych, to jednak wiele z jego argumentów ma charakter aprioryczny, ujawniając wewnętrzną niespójność założeń Arystotelesa. Zacznijmy od problemu spadku swobodnego. Pobieżna obserwacja uświadamia nam, że nie jest możliwe, aby tempo spadku swobodnego było proporcjonalne do ciężaru ciała. Ciało ważące dziesięć kilogramów na pewno nie spada dziesięć razy szybciej niż ciężar jednokilogramowy. Możliwe jednak, że zależność tempa spadku od ciężaru jest innego rodzaju niż prosta proporcjonalność, opisywana funkcją liniową. Galileusz użył pomysłowego argumentu w celu pokazania, że prędkość spadku nie może być żadną rosnącą funkcją ciężaru.

Rozważmy dwa ciała o ciężarach G_1 i G_2, gdzie $G_1 > G_2$. Z założenia ciało o numerze 1 spada szybciej niż ciało 2. Połączmy teraz te ciała np. sztywnym prętem. Z jednej strony, otrzymamy wtedy przedmiot o łącznym ciężarze $G_1 + G_2$, który zatem powinien spadać jeszcze szybciej niż każdy składnik z osobna. Z drugiej strony, ciało o ciężarze G_2, jako wolniejsze, będzie nieco spowalniać spadek ciała G_1, podczas gdy cięższe ciało będzie przyspieszać to wolniejsze, a więc „wypadkowa" prędkość powinna mieścić się pomiędzy dwiema prędkościami każdego z przedmiotów rozważanych osobno. Mamy zatem sprzeczność. Galileusz wyciągnął z tego zaskakujący wniosek, że czas spadku swobodnego nie może w ogóle zależeć od ciężaru. Wszystkie ciała spadają z tą samą prędkością.[8]

Jak zatem wytłumaczyć fakt istnienia niewątpliwych różnic w czasach spadku między niektórymi przedmiotami (przykład kamienia i piórka)? Galileusz podał odpowiedź, którą znamy ze szkolnych podręczników: przyczyną nie są wewnętrzne cechy spadających obiektów, takie jak ciężar, lecz ich oddziaływanie z otoczeniem, czyli opór powietrza. Doświadczenia z upuszczaniem przedmiotów w zamkniętych naczyniach opróżnionych z powietrza jednoznacznie potwierdzają to przypuszczenie. W sytuacji braku oporów powietrza istotnie czas spadku przedmiotów jest jednakowy. Zwróćmy jednak uwagę, że Galileusz nie miał możliwości empirycznej weryfikacji swojego prawa spadku swobodnego. Uznał zatem, że

[7] W każdym razie takie prawo spadku swobodnego przypisywali Arystotelesowi jego następcy, w tym sam Galileusz. Ciekawe jest jednak, że we fragmencie pism Arystotelesa (w księdze IV *Fizyki*), w którym formułuje on zasadę proporcjonalności prędkości ciała do jego ciężaru, nie wspomina w ogóle o spadku, a jedynie o dowolnym ruchu w powietrzu (zapewne wywołanym daną siłą początkową). Arystoteles tłumaczy całkiem przekonująco wprowadzoną przez siebie zależność wpływem oporu powietrza (im cięższe ciało, tym mniejszy opór będzie mu stawiać powietrze – dziś wiemy, że to raczej rozmiar czy kształt ciała odgrywa tutaj rolę). Co więcej, Stagiryta jednoznacznie stwierdza, że w przestrzeni pozbawionej wszelkiego oporu ciała poruszałyby się z jednakową prędkością! Zatem dwa tysiące lat przez Galileuszem sformułował on zasadniczo poprawne prawo spadku. Jednakże Arystoteles uznał, że taka konsekwencja jest absurdalna, z czego wyprowadził znany już wniosek o nieistnieniu próżni.

[8] Oczywiście dzisiaj wiemy, że to nie prędkość, a przyspieszenie jest stałe w spadku swobodnym. Galileusz uświadamiał sobie różnicę między tymi dwoma pojęciami, ale w argumencie posługiwał się nieprecyzyjnym terminem prędkości (zapewne rozumianej jako średnia prędkość na całym odcinku drogi). Można też posłużyć się innym parametrem w postaci czasu spadku ciała z zadanej wysokości.

teoretyczne zalety jego hipotezy są ważniejsze niż niezgodność z obserwowalnymi faktami. Wytłumaczenie owej niezgodności oporami powietrza miało wszelkie cechy rozwiązania *ad hoc*. Jak widać, argumenty *ad hoc* mogą odgrywać w nauce pozytywną rolę, a kurczowe trzymanie się empirii bez teoretycznych założeń może prowadzić na manowce.

Argument Galileusza można uogólnić w następujący sposób. Załóżmy, że tempo spadku swobodnego jest funkcją pewnego parametru m, który jest wielkością addytywną. Wielkość addytywna to taka, której wartość dla fizycznej sumy dwóch przedmiotów jest sumą wartości dla każdego z tych przedmiotów z osobna. Ciężar przedmiotu należy do tej kategorii (ciężar dwóch przedmiotów jest sumą ich ciężarów). Innymi przykładami wielkości addytywnych są masa, objętość czy ładunek elektryczny. Jeśli teraz przyjmiemy za Galileuszem, że funkcja tempa spadku f powinna spełniać następujący warunek:

$f(m_1) \leq f(m_1 + m_2) \leq f(m_2)$ dla $m_1 \leq m_2$,

to z założenia ciągłości można udowodnić, że f musi być funkcją stałą.

Zwróćmy jednak uwagę na niezmiernie istotne założenie zależności tempa spadku od dokładnie jednego parametru. Mogłoby się zdarzyć, że f byłoby funkcją większej liczby niezależnych parametrów. W szczególności rozważmy przypadek, kiedy f jest niemalejącą funkcją ilorazu dwóch niezależnych wielkości addytywnych: $f\left(\frac{q}{m}\right)$. Okazuje się, że w takim wypadku warunek Galileusza

$$f\left(\frac{q_1}{m_1}\right) \leq f\left(\frac{q_1+q_2}{m_1+m_2}\right) \leq f\left(\frac{q_2}{m_2}\right) \text{ dla } \frac{q_1}{m_1} \leq \frac{q_2}{m_2}$$

będzie automatycznie spełniony, niezależnie od konkretnej postaci funkcji f. Zatem stałość czasu spadku swobodnego nie jest prawdziwa na mocy konieczności – wynika ona z tego, że jest dokładnie jeden parametr addytywny, od którego ten czas zależy. Kryje się za tym głębsza prawda fizyczna. Jak się okazuje, przyspieszenie ciała w spadku swobodnym można przedstawić jako funkcję ilorazu dwóch wielkości: masy grawitacyjnej, charakteryzującej oddziaływania grawitacyjne, oraz masy inercjalnej (bezwładnościowej), występującej w prawie dynamiki Newtona. Te dwie wielkości okazują się równe, co implikuje stałość przyspieszenia w polu grawitacyjnym. Jednak równość masy grawitacyjnej i inercjalnej jest faktem empirycznym, który nie wynika z założeń o charakterze *a priori*.

Równie pomysłowy argument Galileusza pokazuje, że Arystotelesowskie wytłumaczenie ruchu pocisku ma konsekwencje jawnie niezgodne z doświadczeniem. Gdyby rzeczywiście podtrzymanie ciał w locie odbywało się za pośrednictwem nacisku powietrza wypychanego przez lecący obiekt, przedmioty tępo zakończone powinny poruszać się szybciej niż szpiczaste, gdyż te pierwsze wypychają więcej powietrza. Ironizując, Galileusz zauważa, że w takim razie łucznicy powinni układać strzałę prostopadle do kierunku lotu, a nie równolegle. Należy zatem poszukać innego wytłumaczenia dla istnienia ruchów bez widocznej podtrzymującej siły. Już w średniowieczu zauważono, że przedmioty mają zdolność do zachowywania nadanej im ilości ruchu, jeśli tylko nie napotkają na jakieś przeszkody. Tę własność nazwano inercją albo bezwładnością. Galileusz zaakceptował zasadę inercji w postaci tezy, że ciała mają skłonność do pozostawania w ruchu ze stałą prędkością, dopóki nie zadziała jakiś czynnik zewnętrzny. Takim czynnikiem może być znany nam już dobrze z analizy spadku swo-

bodnego opór otoczenia – powietrza czy też chropowatej powierzchni, po której przesuwamy przedmiot.

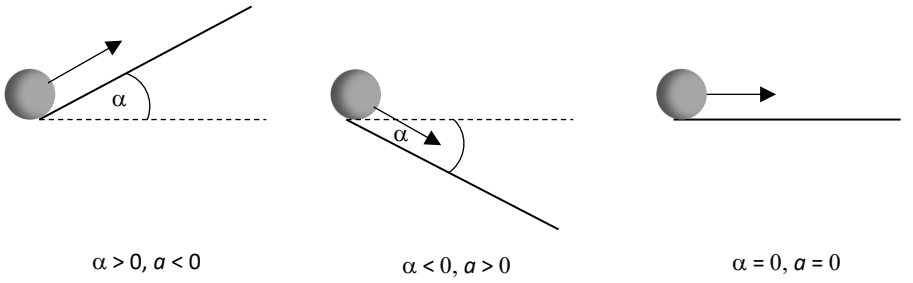
Znowu zatem mamy do czynienia z sytuacją, kiedy same obserwacje nie wystarczą do jednoznacznego określenia praw rządzących pewnymi zjawiskami. W rozważanym przypadku relacji między siłą a ruchem mamy dwie przeciwstawne obserwacje: z jednej strony doświadczenie z ciałami znajdującymi się na powierzchni ziemi, do których przesuwania potrzebna jest stała siła, a z drugiej strony obserwacja pocisków w locie. Arystoteles przyjął pierwszą sytuację jako podstawę swojego prawa dynamiki, a drugi przypadek potraktował jako anomalię, wymagającą dodatkowego wyjaśnienia. Natomiast Galileusz postąpił odwrotnie – uznał, że zachowanie pocisków w locie oddaje prawdziwą naturę zjawisk, a pierwsza, „anomalna" sytuacja powinna zostać wyjaśniona przez dodatkowe zaburzające czynniki, czyli tarcie. Sukces koncepcji Galileusza bierze się między innymi stąd, że wyjaśnienie anomalii proponowane przez Arystotelesa było niezadowalające. Pokazuje to, że teoretyzowanie i stawianie hipotez jest czynnością poznawczą, która wykracza poza prostą analizę danych empirycznych.

1.6. Zasada bezwładności i zasada względności Galileusza

Zasada bezwładności Galileusza miała kapitalne znaczenie dla jego polemiki z przeciwnikami heliocentrycznej teorii Kopernika. Jak pamiętamy, jednym z głównych zarzutów stawianych Kopernikowi był brak odczuwalnych efektów ruchu obrotowego Ziemi. Na przykład twierdzono, że gdyby Ziemia obracała się wokół własnej osi, przedmiot upuszczony z wieży odchyliłby się od podstawy wieży o odległość, jaką w tym czasie przebyłaby wieża wraz z całą powierzchnią. Galileusz w odpowiedzi zauważył po pierwsze, że sugerowany efekt nie pojawia się na przykład podczas upuszczenia przedmiotu z masztu statku płynącego z pewną stałą prędkością. Przedmioty upuszczane z masztu upadają u jego podnóża, niezależnie od ruchu statku (pod warunkiem, że pominiemy kołysania). Nie jest jasne, czy Galileusz wykonywał odpowiednie obserwacje, potwierdzające ten fakt empiryczny, który oczywiście ma miejsce. Natomiast teoretyczne wyjaśnienie opierało się właśnie na zasadzie zachowania stałej prędkości ciała w ruchu. Zarówno ciało znajdujące się na szczycie wieży, jak i ciało przymocowane do masztu poruszającego się statku, uczestniczą w ruchu (Ziemi lub statku) i posiadają pewną poziomą prędkość. Upuszczenie tych przedmiotów nie pozbawia ich tej prędkości, a zatem nawet w przypadku braku kontaktu z wieżą czy masztem ciała te nadal podążają w tym samym kierunku co pozostałe obiekty.

Galileusz nie poprzestał na prostym uogólnieniu empirycznych obserwacji w celu uzasadnienia zasady bezwładności. W zgodzie ze swoimi skłonnościami do teoretyzowania zaproponował ciekawy argument *a priori* z równi pochyłych za prawdziwością zasady bezwładności. Rozważmy płaszczyznę nachyloną do poziomu pod pewnym kątem oraz kulkę u podnóża tej płaszczyzny (rys. 1.12). Nadając tej kulce pewną prędkość początkową, spowodujemy, że zacznie się ona wtaczać na płaszczyznę, ale jednocześnie będzie tracić prędkość aż do całkowitego zatrzymania (i stoczenia się w dół). Kiedy zaczniemy zmniejszać kąt nachylenia płaszczyzny, kulka będzie wytracać swoją prędkość coraz wolniej (tj. będzie wznosić się na coraz większą odległość). Obracając jeszcze dalej płaszczyzną, osiągniemy kąt „ujemny" (nachylenie w dół), co spowoduje przyspieszanie kulki. Zatem dla pewnego kąta musimy mieć sytuację, w której kulka nie będzie ani przyspieszać, ani zwalniać, czyli

będzie poruszać się ze stałą prędkością.[9] Nietrudno zauważyć, że powinno się to zdarzyć dla kąta nachylenia równego zeru, a ruch kulki będzie w takim wypadku ruchem bezwładnym. Przedłużając poziomy tor jej ruchu bezwładnego, możemy przyjąć, że ruch ten jest prostoliniowy. Niestety Galileusz, zasugerowany kulistym kształtem Ziemi, błędnie wywnioskował ze swojego argumentu, że ruch bezwładny będzie w istocie ruchem po okręgu, w stałej odległości od centrum Ziemi. Wynikało to częściowo z faktu, że nie dysponował on jeszcze pojęciem siły grawitacji, które pojawiło się dopiero u Newtona. Galileuszowska koncepcja ciężkości była nadal bliższa Arystotelesowskiej koncepcji ruchów naturalnych bez przyczyny zewnętrznej.

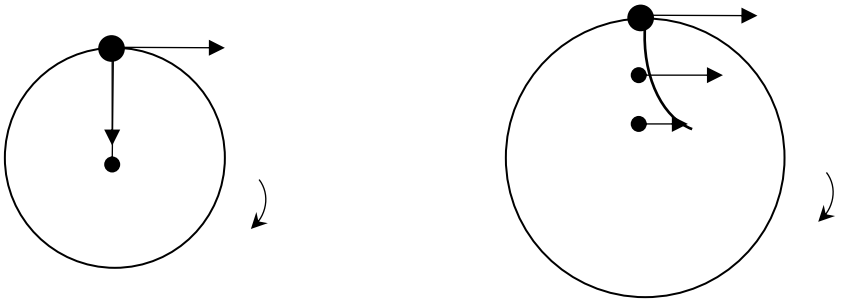


Rys. 1.12. Argument Galileusza z równi pochyłych za zasadą bezwładności

Brak pojęcia siły ciężkości wpłynął również na sposób, w jaki Galileusz usiłował odeprzeć zarzut z braku odczuwalnej siły odśrodkowej pochodzącej od założonego ruchu obrotowego Ziemi. W czasach Kopernika i Galileusza doskonale zdawano sobie sprawę z istnienia siły, która „wypycha" przedmioty znajdujące się w ruchu obrotowym w kierunku przeciwnym do osi obrotu. Zatem należałoby się spodziewać, że analogiczna siła powinna oddziaływać na wszystkie przedmioty znajdujące się na powierzchni Ziemi, które pod wpływem tej siły z wielką prędkością odlatywałyby w kosmos. Galileusz włożył wiele wysiłku w odparcie tego argumentu. Niestety jego rozwiązanie było całkowicie błędne – usiłował pokazać, że ponieważ ruch po okręgu można podzielić na bardzo małe odcinki zbliżone do prostych stycznych, ciało uczestniczące w ruchu obrotowym Ziemi nie będzie podlegało chwilowej odpychającej siły odśrodkowej. Trudno zrozumieć, dlaczego przenikliwy umysł Galileusza nie zauważył, że gdyby jego argument był poprawny, przeczyłoby to istnieniu wszelkich sił odśrodkowych, np. przy wyrzucaniu pocisków z procy. Poprawne rozwiązanie problemu siły odśrodkowej, jak dobrze wiemy, jest takie, że jest ona rekompensowana siłą przyciągania grawitacyjnego. Dokładniej, siła wypadkowa działająca na dany przedmiot na powierzchni Ziemi powstaje z wektorowego odjęcia siły odśrodkowej (w kierunku od osi obrotu) od siły ciężkości skierowanej ku centrum Ziemi. Powoduje to, że będąc na równiku ważymy nieco mniej niż na biegunie, ale różnica ta jest praktycznie niedostrzegalna.

[9] Argument Galileusza można zrekonstruować we współczesnym języku fizyki matematycznej. Przyjmujemy istnienie funkcyjnej zależności przyspieszenia kulki od kąta nachylenia płaszczyzny równi pochyłej: $a = f(\varphi)$; zakładamy ponadto, że funkcja f przyjmuje wartości ujemne dla dodatnich kątów φ oraz dodatnie dla wartości kątów ujemnych. Z założeń tych wynika, jeśli dodatkowo przyjmiemy ciągłość funkcji f, że dla kąta $\varphi = 0$ przyspieszenie wynosi 0.

Dodajmy w tym miejscu uzupełniającą informację na temat innego rodzaju siły, która powstaje w układach obracających się, a której istnienie można łatwo stwierdzić empirycznie. Jest to tzw. siła Coriolisa, która działa tylko na ciała poruszające się z pewną prędkością w obracającym się układzie. Na ciało, które zbliża się lub oddala od osi obrotu, zaczyna działać siła prostopadła do kierunku jego ruchu i zależna zarówno od prędkości obrotu, jak i prędkości poruszania się tego ciała. Istnienie sił Coriolisa jest odpowiedzialne za tworzenie się huraganów czy cyklonów, które powstają, gdy masy powietrza przemieszczają się z gorącego obszaru równikowego w kierunku północnym bądź południowym. Powstaje wtedy „wir", który zawsze obraca się w kierunku zgodnym z wskazówkami zegara na półkuli południowej, a przeciwnie na półkuli północnej. Innym świadectwem siły Coriolisa jest zachowanie tzw. wahadła Foucaulta, które w cyklu dobowym zmienia płaszczyznę swoich wahań. Efekty działania siły Coriolisa stanowią mocne świadectwo na rzecz obrotu Ziemi – niestety Galileusz nie znał tego efektu, więc nie mógł skorzystać z tego argumentu w swojej polemice z Arystotelikami.



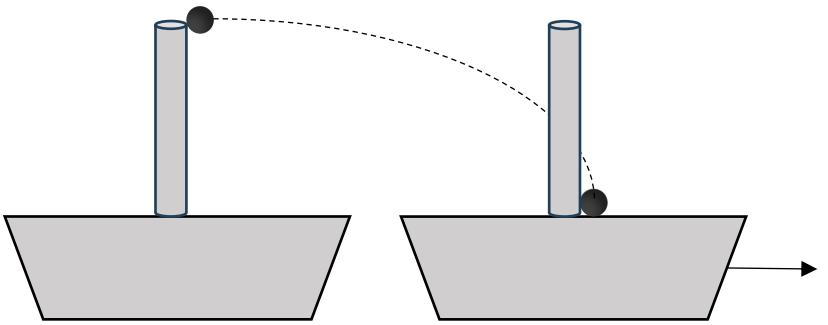
Rys. 1.13. Wyjaśnienie powstania siły odśrodkowej i Coriolisa w ruchu obrotowym za pomocą zasady bezwładności

Ciekawe jest, że istnienie zarówno siły odśrodkowej, jak i Coriolisa, może być jakościowo uzasadnione za pomocą Galileuszowskiej zasady bezwładności. Rozważmy przedmiot znajdujący się w ruchu obrotowym wokół pewnego punktu. W każdym momencie swego ruchu przedmiot ten ma określoną pewną prędkość liniową, skierowaną stycznie do okręgu, po którym się porusza. Ze względu na zasadę bezwładności przedmiot ten ma „skłonność" do kontynuowania swojego lotu bez zmiany prędkości (jej wartości i kierunku), zatem jeśli nie będzie on przytrzymywany, „wystrzeli" po linii stycznej do okręgu. Aby utrzymać przedmiot w ruchu po okręgu, musimy zadziałać na niego siłą dośrodkową, która będzie dokładnie równa co do wartości, lecz przeciwnie skierowana do siły odśrodkowej (rys. 1.13).

W celu wyprowadzenia siły Coriolisa rozważmy sytuację, w której przedmiot znajdujący się na obracającym się dysku zaczyna się przesuwać w kierunku osi obrotu. Znajdując się coraz bliżej osi obrotu, przedmiot ten napotyka punkty, które posiadają mniejszą prędkość liniową (prędkość liniowa punktów na obracającym się dysku zależy od ich odległości od centrum). Zgodnie z zasadą bezwładności prędkość liniowa przedmiotu styczna do okręgu pozostaje niezmieniona, a zatem zacznie on „wyprzedzać" napotkane punkty, co

obserwator znajdujący się na obracającym się dysku zinterpretuje jako przejaw siły działającej prostopadle do kierunku jego ruchu.

Wracając do ulubionego przykładu Galileusza ze statkiem, możemy zauważyć, że posłużył mu on jeszcze do omówienia dwóch związanych ze sobą zasad: zasady składania prędkości (niezależności ruchów) oraz zasady względności. Rozważmy ponownie przykład jednostajnie posuwającego się statku i ciężkiego przedmiotu zrzuconego z wierzchołka masztu. Jak argumentował Galileusz, z perspektywy statku spadek ciała nie będzie się niczym różnił od spadku z obiektu stacjonarnego. Natomiast z punktu widzenia obserwatora znajdującego się na brzegu, upuszczone ciało zakreśli pewną krzywą (dziś wiemy, że jest to parabola). Galileusz zauważył, że ruch ten można potraktować jako składający się z dwóch niezależnych ruchów, zachodzących równocześnie: pionowego spadku swobodnego i poziomego przesuwania się statku i wszystkich związanych z nim przedmiotów (rys. 1.14). Było to zupełnie nowatorskie odkrycie – nigdy przedtem nie „rozkładano" ruchów w ten sposób, traktując je jako całości.



Rys. 1.14. Zasada składania ruchów Galileusza

Galileusz zwrócił ponadto uwagę na fakt, że jedną ze składowych ruchu złożonego można w pewnym sensie „wyłączyć", nie zaburzając drugiej z nich, przez zmianę perspektywy (dziś powiedzielibyśmy: przez zmianę układu odniesienia). Tę obserwację zastosował m.in. do opisu ruchu pocisku armatniego, wystrzelonego poziomo w stosunku do ziemi. Argumentował, że przyjmując perspektywę obserwatora poruszającego się poziomo razem z pociskiem, otrzymamy nic innego jak spadek swobodny kuli armatniej. Zapewne rozważania tego rodzaju podsunęły mu uniwersalną myśl o kapitalnym znaczeniu dla późniejszego rozwoju fizyki. Doszedł bowiem do wniosku, że zjawiska mechaniczne wyglądają zasadniczo tak samo niezależnie od stanu ruchu obserwatora, jeśli tylko sam obserwator porusza się ruchem jednostajnym i prostoliniowym. Tezę tę, którą dzisiaj określamy mianem zasady względności, zilustrował Galileusz znów przy pomocy przykładu statku. Tym razem jednak wyobraźmy sobie, że obserwator został zamknięty pod pokładem w kajucie bez okien. Podobnie jak poprzednio załóżmy, że statek sunie po spokojnym morzu ze stałą prędkością, nie kołysząc się ani nie przyspieszając. Czy obserwator w kajucie będzie w stanie określić, czy statek jest w ruchu czy też stoi w porcie? Galileusz stwierdził, że żadne procesy mechaniczne

zachodzące wewnątrz kajuty nie będą zależeć od stanu ruchu statku. Wszystkie ruchy przedmiotów zamkniętych w kajucie, czy to swobodne czy wymuszone, będą wyglądały tak samo zarówno na statku spoczywającym, jak i poruszającym się. Aby stwierdzić, że statek znajduje się w ruchu, należy wyjść na pokład i porównać jego położenie z otoczeniem zewnętrznym. Innymi słowy, ruch jednostajny i prostoliniowy jest pojęciem względnym, a nie absolutnym.

W następnym rozdziale zobaczymy, jak zasada względności Galileusza zostanie doprecyzowana na gruncie mechaniki newtonowskiej w postaci tezy o niezmienniczości praw mechaniki względem pewnego rodzaju transformacji, zwanych transformacjami Galileusza. Będziemy musieli również wprowadzić ważne pojęcie inercjalnego układu odniesienia. Na razie jednak zakończmy przegląd niezwykłych osiągnięć Galileusza, wspominając jego badania nad ruchem jednostajnie przyspieszonym. Zauważył on, że ruch spadającego swobodnie ciała nie może być jednostajny, gdyż zaczyna się od zerowej prędkości, a kończy na pewnej niezerowej wartości. Zatem spadające ciało nie będzie pokonywać równych odcinków drogi w równych odstępach czasu. Aby zbadać tę sprawę dokładniej, uczony z Pizy obserwował kulki staczające się po równiach pochyłych (wbrew powszechnemu mitowi nie badał on czasów spadku przedmiotów z krzywej wieży w Pizie). Mierząc drogę, jaką kulki te pokonywały w jednakowych odcinkach czasu (np. w każdej kolejnej sekundzie), zauważył ciekawą prawidłowość: drogi te pozostawały do siebie w stosunkach odpowiadających kolejnym liczbom nieparzystym: 1 : 3 : 5 : 7 itd. Proporcję tę można wyrazić nieco inaczej, porównując kumulatywne drogi przebyte w czasie jednej, dwóch, trzech itd. sekund. W takim ujęciu drogi będą miały się do siebie jak kwadraty kolejnych liczb: $1^2 : 2^2 : 3^2 : 4^2$ … (Wynika to z algebraicznej zależności, że suma n kolejnych liczb nieparzystych jest kwadratem z liczby n.) Zatem Galileusz *de facto* odkrył, że droga w ruchu jednostajnie przyspieszonym rośnie proporcjonalnie do kwadratu czasu. W ten sposób rozszerzył on matematyczne metody stosowane do opisu zjawisk kinematycznych, wychodząc poza proste proporcje Arystotelesowe. Dalsze doskonalenie aparatu matematycznego fizyki dokona się za sprawę wielkiego Izaaka Newtona.

Pytania i problemy

1. Wymień podstawowe obserwacyjne fakty dotyczące ruchów ciał niebieskich (gwiazd, w tym Gwiazdy Polarnej, Słońca i planet) znane starożytnym.

2. W jaki sposób starożytni astronomowie byli w stanie oszacować absolutną odległość Ziemi do Słońca? Zwróć uwagę na pomocniczą rolę obliczenia promienia Ziemi przez Eratostenesa i oszacowania odległości od Ziemi do Księżyca.

3. Wymień elementy geocentrycznego modelu Ptolemeusza i ich rolę w wyjaśnieniu obserwowalnych prawidłowości ruchów planetarnych. Na czym polega charakter *ad hoc* tych elementów?

4. Jaka relacja musi zachodzić między prędkością obiegową planety po epicyklu i po deferensie, aby model Ptolemeusza poprawnie odtwarzał obserwowane zjawisko „cofania się" planety? Jakiego rodzaju założenie dotyczące prędkości orbitalnych Ziemi oraz planet zewnętrznych i wewnętrznych musi zostać przyjęte w teorii kopernikańskiej, aby uzyskać ten sam efekt? Które z założeń – ptolemejskie czy kopernikańskie – jest bardziej *ad hoc*?

5. Omów zasadnicze zalety i trudności modelu kopernikańskiego w porównaniu z geocentrycznym modelem Ptolemeusza. Jakie naukowe powody przemawiały za tym, aby teorię Kopernika traktować z pewną dozą sceptycyzmu?

6. Co to jest *experimentum crucis* (eksperyment krzyżowy)? Czy w nauce możemy spotkać eksperymenty krzyżowe w czystej postaci?

7. Na czym polega instrumentalistyczna interpretacja danej teorii naukowej? Jakiego rodzaju podział w obrębie języka danej teorii postuluje instrumentalizm? Zilustruj to zagadnienie na przykładzie teorii kopernikańskiej.

8. Jak poznać, że w nauce dokonała się rewolucja? Jakie są charakterystyczne cechy rewolucji naukowych?

9. Omów trzy prawa Keplera ruchu planetarnego. Czy prawa te stosują się do wszystkich obiektów we wszechświecie?

10. Sformułuj podstawowe prawo dynamiki Arystotelesa. Jak możemy przedstawić je w języku współczesnej algebry? Jaki jest w koncepcji Arystotelesa warunek niezbędny do tego, aby ciało kontynuowało swój ruch (nie zatrzymało się)?

11. Omów argument Galileusza przeciwko prawu spadku swobodnego przypisywanemu Arystotelesowi, a także argument przeciwko jego wytłumaczeniu ruchu pocisków.

12. Przedstaw argument Galileusza z równi pochyłych za zasadą bezwładności. Do jakiego mylnego wniosku doszedł sam Galileusz w kwestii kształtu toru ciała poruszającego się ruchem bezwładnym na powierzchni Ziemi?

13. W jaki sposób Galileusz wykorzystał swoją zasadę bezwładności, aby odeprzeć zarzuty w stosunku do teorii heliocentrycznej z nieobserwowalności efektów ruchu obrotowego Ziemi?

14. Omów zasadę składania prędkości na przykładzie ciała upuszczonego z wierzchołka poruszającego się statku.

15. Przedstaw argument Galileusza ze statkiem za zasadą względności. Jak dokładnie brzmi ta zasada? Jakie jest ograniczenie jej stosowalności?

Literatura uzupełniająca

Szczegółowe omówienie historycznego rozwoju starożytnej astronomii i mechaniki można znaleźć w monumentalnej publikacji: A.K. Wróblewski, *Historia fizyki*, PWN Warszawa 2007, rozdział 2.

Bardzo przystępne porównanie modelu ptolemejskiego i kopernikańskiego zawiera znany podręcznik: E.M. Rogers, *Fizyka dla dociekliwych. Tom 2 Astronomia*, PWN Warszawa 1974.

Filozoficzne aspekty przewrotu kopernikańskiego ze szczególnym uwzględnieniem problemu rewolucji w nauce omówione są w klasycznej książce: T. Kuhn, *Przewrót kopernikański: astronomia planetarna w dziejach myśli Zachodu*, wyd. 2, Prószyński i S-ka, Warszawa 2006.

Polecam lekturę pięknego dzieła, jakim są *Dialogi* Galileusza: Galileo Galilei, *Dialogi o dwu najważniejszych układach świata: Ptolemeuszowym i Kopernikowym*, wyd. 2, PWN, Warszawa 1962.

Rozwój astronomii od Ptolemeusza do Kopernika i Galileusza jest głównym tematem części II książki: J.T. Cushing, *Philosophical Concepts in Physics*, Cambridge University Press, Cambridge 1998.

ROZDZIAŁ 2. MECHANIKA KLASYCZNA

Mechanika klasyczna, zwana również newtonowską, stanowi fundament całej fizyki, także współczesnej. Bez pojęć analizowanych w mechanice, takich jak siła, pęd, energia, masa itd., nie sposób wyobrazić sobie rozwoju nawet najbardziej zaawansowanych teorii fizycznych. Mimo że oficjalnie mechanika newtonowska została obalona i zastąpiona bardziej adekwatnymi empirycznie teoriami (obie teorie względności i mechanika kwantowa), to jednak jej obecność zaznacza się wyraźnie w każdym niemal dziale fizyki. Matematyczne metody mające swoje źródło w mechanice nadal stosuje się z powodzeniem do opisu różnorodnych zjawisk, włącznie z procesami na poziomie kwantowym, obejmującymi zarówno cząstki, jak i pola kwantowe. Z filozoficznego punktu widzenia mechanika newtonowska oferuje bogactwo problemów metodologicznych, ontologicznych i epistemologicznych. Naszą filozoficzną analizę mechaniki klasycznej rozpoczniemy od przedstawienia podstawowych praw dynamiki w formie zbliżonej do oryginalnego sformułowania Newtona. Będziemy zastanawiać się nad statusem metodologicznym i treścią empiryczną tych praw, podejmując kontrowersyjną kwestię, czy przypadkiem nie są one ukrytymi definicjami (konwencjami). Następnie pokażemy, jak zasady dynamiki działają w praktyce, umożliwiając nam wyznaczenie przyszłego zachowania mechanicznego (ruchu) ciał poddanych działaniom różnych sił. Prowadzi to wprost do niezwykle ważnego zagadnienia ontologicznego i epistemologicznego, jakim jest problem determinizmu. Powszechnie uważa się, że mechanika klasyczna stanowi wzorcowy przykład teorii deterministycznej, w przeciwieństwie np. do indeterministycznej mechaniki kwantowej. Zbadamy ten pogląd dokładniej, kładąc nacisk na precyzyjne scharakteryzowanie różnych wariantów stanowiska deterministycznego. Okaże się, że teza o deterministycznym charakterze zjawisk mechanicznych nie jest wcale tak bezsporna, jak by się mogło wydawać.

Istnieje bliski związek między mechaniką klasyczną a zagadnieniem natury czasu i przestrzeni. Opis zachowań mechanicznych ciał wymaga wprowadzenia parametrów charakteryzujących ich położenie w przestrzeni i czasie (współrzędnych przestrzennych i czasowych). Do czego jednak odnoszą się owe parametry? Czy czas i przestrzeń są substancjami istniejącymi niezależnie od wypełniających je przedmiotów fizycznych? Spór na ten temat toczył się pomiędzy dwoma wybitnymi umysłami epoki, Newtonem i Leibnizem, z których pierwszy reprezentował stanowisko absolutyzmu, a drugi relacjonizmu. Będziemy analizować dwa typowe argumenty wykorzystane w tym sporze, a także formalny aspekt tego zagadnienia

w postaci tzw. transformacji Galileusza i niezmienników galileuszowskich. Poruszymy również zagadnienie wyprowadzenia prawa grawitacji z obserwacji astronomicznych i opartych na nich praw ruchu planetarnego Keplera, a także filozoficzny problem tzw. działania na odległość. Wreszcie wspomnimy o nowszych i bardziej zaawansowanych matematycznie podejściach do mechaniki klasycznej: mechanice Hamiltonowskiej, zasadzie najmniejszego działania i mechanice Lagrange'a. Szczególnie zasada najmniejszego działania zasługuje na uwagę filozofa, jako że wydaje się ona rehabilitować zdyskredytowane w naukach przyrodniczych pojęcie przyczyny celowej.

2.1. Prawa dynamiki Newtona

Podwaliny nowożytnej mechaniki klasycznej zostały położone przez Izaaka Newtona w jego przełomowym dziele *Principia Mathematica* (jego pełny tytuł brzmi *Philosophiae Naturalis Principia Mathematica* – Matematyczne zasady filozofii naturalnej), opublikowanym w 1687 r. Życiorys Newtona zasługuje na osobny rozdział, na co jednak nie mamy tutaj miejsca. Urodzony w ubogiej rodzinie, wychowywany przez matkę i ojczyma, z którym nie łączyły go specjalnie bliskie uczucia, szybko wykazał się nieprzeciętnymi zdolnościami naukowymi i ambicją, jak również trudnym charakterem. Jego działalność naukowa obejmowała prace nad podstawami mechaniki, matematyki (rachunek różniczkowy i całkowy), teorii grawitacji, optyki. Wykazywał również zainteresowania problemami o charakterze bardziej ezoterycznym czy teologicznym. Miał ponadto spore talenty organizacyjne, pełniąc funkcję nadzorcy mennicy królewskiej oraz przewodniczącego Królewskiego Towarzystwa Naukowego. Był uwikłany w ostre konflikty personalne, między innymi z Robertem Hookiem i oczywiście z Leibnizem w kwestii pierwszeństwa w odkryciu rachunku różniczkowego. Jednakże ogromne osiągnięcia Newtona na polu naukowym, które wyznaczyły kierunek rozwoju fizyki na kolejne stulecia, pozwalają na łagodniejsze potraktowanie niedostatków jego charakteru.

Principia napisane zostały w sposób odzwierciedlający preferowaną przez niego metodę naukową, opartą na ścisłych definicjach, aksjomatach i twierdzeniach. Na początku dzieła podaje on osiem definicji fundamentalnych pojęć, w tym pojęcia masy, ilości ruchu (dzisiaj pędu) i siły. Następnie we fragmencie pod nazwą *Scholium* przedstawia swoje poglądy na temat czasu i przestrzeni. Wreszcie kończąc pierwszą część *Principiów* formułuje słynne trzy prawa (aksjomaty) dynamiki, od przypomnienia których zaczniemy nasze spotkanie z mechaniką klasyczną. Oto w jakiej postaci znajdujemy je u Newtona:

1. Każde ciało pozostaje w stanie spoczynku lub w stanie jednostajnego ruchu po linii prostej, dopóki nie zostanie zmuszone do zmiany przez działanie siły.
2. Zmiana ilości ruchu jest proporcjonalna do przyłożonej siły i dokonuje się w kierunku działania tej siły.
3. Wzajemne oddziaływania dwóch ciał na siebie są zawsze równe i przeciwnie skierowane.

Pierwsze prawo dynamiki rozpoznajemy jako zasadę bezwładności, antycypowaną już przez Galileusza i przyjętą przez kartezjan. Zrywając definitywnie z paradygmatem arystotelesowskim, Newton podkreśla w tym prawie, że ruch jest stanem naturalnym, niewymagającym żadnej przyczyny, jeśli tylko odbywa się w sposób jednostajny i po prostej linii. Sfor-

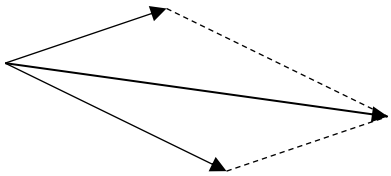
mułowanie Newtona zawiera ponadto dodatkową wskazówkę na temat roli sił w mechanice – skutkiem działania sił nie jest „podtrzymywanie" ruchu, ale jego zmiana. Ta myśl zostaje rozwinięta w ilościowy sposób w drugim prawie, które było oryginalnym i nowatorskim wkładem Newtona. Zgodnie z tym prawem, zmiana ilości ruchu, definiowanej jako iloczyn masy ciała i jego prędkości, jest proporcjonalna do wielkości przyłożonej do ciała siły. Należy do tego dodać, że proporcjonalność obejmuje również czas działania siły – im dłuższy ten czas, tym większa zmiana. W rezultacie otrzymujemy następującą prostą zależność matematyczną:

$$\Delta(mv) = F\Delta t\,,$$

z czego wyprowadzamy, przy założeniu niezmienności masy, znane nam wszystkim sformułowanie drugiej zasady:

$$F = m\frac{\Delta v}{\Delta t} = ma\,,$$

gdzie *a* jest przyspieszeniem ciała.[1]



Rys. 2.1. Reguła równoległoboku zastosowana do obliczenia wypadkowej siły

Trzecie prawo dynamiki to dobrze znana zasada akcji i reakcji. Warto może pamiętać, że siły akcji i reakcji wzajemnie się nie „kasują", chociaż są sobie równe i przeciwnie skierowane. Są one bowiem przyłożone do różnych obiektów. Jeśli natomiast przyłożylibyśmy takie siły do tego samego ciała, ich rezultat byłby zerowy. Prowadzi to nas do kolejnej zasady przyjętej przez Newtona, którą niektórzy nazywają czwartym prawem. Jest to znana reguła

[1] Warto zauważyć, że Newton nigdy nie posługiwał się równaniem $F = ma$. Pierwszym uczonym, który wprowadził to sformułowanie drugiej zasady, był Leonhard Euler. Dodatkowo można postawić pytanie, czy pierwsze prawo nie zawiera się już w drugim. Podstawiając $F = 0$, mamy wniosek, że przyspieszenie ciała jest zerowe, co można zinterpretować jako twierdzenie, że ciało porusza się ruchem jednostajnym i prostoliniowym. Jednakże sprawa nie jest oczywista. Po pierwsze, druga zasada dynamiki w sformułowaniu Newtona nic nie mówi o tym, jak wygląda ruch ciała w przypadku braku sił. Podstawienie wartości zerowej siły jest ściśle rzecz biorąc wykroczeniem poza zakres stosowalności drugiego prawa (chociaż można ten krok uzasadniać założeniem ciągłości przy rozważaniu coraz to mniejszych wartości siły). Po drugie, założenie, iż przyspieszenie ciała jest zerowe, jeszcze nie implikuje, że ciało może mieć niezerową prędkość. Zagorzały arystotelik mógłby utrzymywać, że brak siły skutkuje natychmiastowym zatrzymaniem ciała, w rezultacie czego zarówno prędkość, jak i przyspieszenie są zerowe, zgodnie z drugą zasadą. Oczywiście możliwe są dalsze argumenty pokazujące błędność takiego rozumowania – jeśli w chwili, w której na ciało przestaje działać siła, miało ono niezerową prędkość, to zgodnie z drugim prawem prędkość ta nie może ulec zmianie, a więc ciało musi poruszać się ruchem jednostajnym. W każdym razie z powodów co najmniej historycznych formułuje się pierwsze prawo jako osobną zasadę.

równoległoboku, umożliwiająca nam dodawanie sił działających w różnych kierunkach. Newton uświadamiał sobie, że wielkości takie jak siła, prędkość czy przyspieszenie charakteryzują się nie tylko wartościami liczbowymi, ale także kierunkowością. Dzisiaj wielkości takie nazywamy wektorowymi. Aby dodać do siebie dwa wektory, nie możemy po prostu dodać liczbowo ich długości – musimy wykreślić równoległobok o bokach równych wartościom tych wektorów i kierunkach zgodnych z kierunkami wektorów, a ich suma będzie wektorem łączącym przeciwległe wierzchołki równoległoboku (rys. 2.1).

Obecnie zajmiemy się bardziej szczegółowo problemem metodologicznego statusu praw dynamiki. Pojawiają się czasem sugestie, że prawa dynamiki Newtona są w istocie ukrytymi definicjami występujących w nich terminów, takich jak siła czy masa, a zatem są one prawdziwe na mocy znaczenia, a nie faktów pozajęzykowych. Jeśli nawet aksjomaty dynamiki nie są jedynie definicjami, czyli posiadają pewną treść empiryczną, to pozostaje pytanie, czy możliwe jest definitywne empiryczne sprawdzenie ich prawdziwości. Sam Newton raczej na pewno nie przeprowadzał eksperymentów mających na celu weryfikację sformułowanych przez siebie praw. Można mieć wątpliwości, czy taka weryfikacja (lub też falsyfikacja, czyli obalenie) jest w ogóle wykonalna. Gdyby okazało się, że żadna eksperymentalna procedura nie jest w stanie wykazać fałszywości praw dynamiki, ich przyjęcie byłoby kwestią konwencji, a nie badań empirycznych.

Może budzić zdziwienie, dlaczego niemożliwość okazania fałszywości jakiegoś prawa stanowi problem. Czy nie jest tak, że w nauce zależy nam na tym, aby uznawać tylko prawdziwe twierdzenia? Filozof Karl Popper zwrócił uwagę, że możliwość obalenia (falsyfikacji) danego twierdzenia jest równie ważna, a może nawet ważniejsza od jego prawdziwości. Zdania, których nie da się w żaden sposób obalić, są nieciekawe z punktu widzenia empiryzmu, gdyż nie mówią niczego na temat rzeczywistości (nie ograniczają w żaden sposób dopuszczalnych możliwości). Falsyfikacjonizm Poppera to stanowisko, zgodnie z którym podstawowym celem naukowca powinno być usiłowanie obalenia uznanych praw poprzez ich stałe testowanie. Dopiero jeśli dana teza przejdzie pomyślnie takie testy, zostaje ona zaakceptowana.

Zacznijmy jednak od wprowadzenia pewnych podstawowych pojęć epistemologicznych, które ułatwią nam rozważenie powyższych problemów. Filozofowie dokonują rozróżnienia pomiędzy sądami (zdaniami) empirycznymi i apriorycznymi, a także między sądami analitycznymi i syntetycznymi. Sąd empiryczny (inaczej zwany *a posteriori*) to taki, którego prawdziwość (czy fałszywość) zależy od pewnych faktów danych nam w doświadczeniu. Na przykład zdanie „Ta róża jest czerwona" jest ewidentnie empiryczne, gdyż stwierdzamy jego prawdziwość na podstawie bezpośredniej obserwacji. Mogą istnieć oczywiście sądy empiryczne, których prawdziwość nie opiera się na bezpośrednich obserwacjach, jak np. twierdzenie „Elektrony są obdarzone ujemnym ładunkiem elektrycznym". Jednakże nawet w tym wypadku weryfikacja takiego zdania następuje w drodze nierzadko skomplikowanych obserwacji przy pomocy odpowiednich urządzeń pomiarowych. Natomiast sądy aprioryczne nie posiadają „weryfikatorów" w formie obserwacji. Do stwierdzenia ich prawdziwości nie jest potrzebne żadne doświadczenie zmysłowe. Najbardziej typowymi przykładami sądów *a priori* są twierdzenia matematyki, takie jak „Suma wszystkich kątów w trójkącie jest równa kątowi półpełnemu". Twierdzenie to weryfikujemy podając jego dowód, a nie przez odpowiednie obserwacje.

Rozróżnienie na zdania analityczne i syntetyczne jest nieco bardziej subtelne. Wprowadził je słynny niemiecki filozof epoki Oświecenia, Immanuel Kant. Zdanie analityczne, w ujęciu Kanta, „rozkłada" (czyli analizuje) podmiot w orzeczeniu. Innymi słowy, w zdaniu analitycznym stwierdzamy o podmiocie to, co jest już w nim zawarte. Współcześnie bardziej powszechna jest interpretacja sądów analitycznych jako prawdziwych na mocy znaczenia wyrazów, które z kolei zwykle nadajemy za pomocą definicji. Zdanie „Trójkąt jest figurą o trzech kątach" jest analityczne, gdyż wyrażenie „trójkąt" definiujemy właśnie jako figurę o trzech kątach. Z kolei sądy syntetyczne to po prostu te, które nie są analityczne. Ich prawdziwość musi być oparta na czymś innym niż nadane przez nas znaczenia słów.

Z powyższych charakterystyk wynika, że zdanie analitycznie prawdziwe musi być *a priori*, gdyż nadawanie znaczeń słowom nie jest procedurą opartą na obserwacjach czy szerzej na doświadczeniu zmysłowym. Sprawą dyskusyjną natomiast jest kwestia istnienia sądów syntetycznych *a priori*. Kant uważał, że sądy takie można znaleźć m.in. wśród twierdzeń matematyki, które nie są definicjami, ale jest to stanowisko kontrowersyjne (por. poniższą tabelkę). Nie rozstrzygając tego problemu filozoficznego, zajmijmy się teraz kwestią tego, do jakiej kategorii sądów należy zaliczyć prawa dynamiki. Możliwości są dwie: albo są to zdania syntetyczne *a posteriori*, czyli niewynikające ze znaczeń, ale oparte na doświadczeniu, albo analityczne *a priori*. Aby wykluczyć tę drugą możliwość, musimy przyjrzeć się, w jaki sposób interpretuje się podstawowe terminy wchodzące w ich skład.

	analityczne	**syntetyczne**
a priori	„Trójkąt ma trzy kąty"	?
a posteriori	X	„Trawa jest zielona"

Tab. 2.1. Podział sądów

Zacznijmy od pierwszego prawa, czyli zasady bezwładności. Występuje w nim pojęcie siły, które odłóżmy na razie do czasu analizy drugiego prawa. Ponadto mówi się tam o ruchu jednostajnym. Ruch jednostajny to taki, w którym ciało pokonuje równe odcinki drogi w równych okresach czasu. Pojęcie równości odcinków przestrzennych nie nastręcza specjalnych trudności – pomiary odległości przestrzennych są powszechnie znane i zrozumiałe. Natomiast pomiary czasu są obciążone pewnym problemem. Chodzi o to, że zasadą działania dobrego zegara jest jego regularność – każdy kolejny cykl zegara powinien mieć dokładnie takie samo trwanie, co poprzedni. Skąd jednak możemy wiedzieć, że nasz zegar np. nie spowalnia w niezauważalny sposób? Aby to ustalić, potrzebujemy drugiego zegara w celu skontrolowania pierwszego. Ale to samo pytanie można postawić w odniesieniu do zegara kontrolnego, a zatem problem powtarza się w nieskończoność.

Większość filozofów nauki zgadza się, że rozwiązanie tego problemu musi opierać się na przyjęciu pewnej konwencji. Jedna z możliwości, jaka się nasuwa, związana jest bezpośrednio z pierwszą zasadą dynamiki. Czy nie można po prostu przyjąć za wzorcowy zegar ciała niepoddanego działaniu żadnych sił? Równe odcinki czasu byłyby wtedy zdefiniowane jako takie, w których owe ciało pokonuje równe drogi. Oczywiście nie jest to pomysł do praktycznej realizacji z powodów, o których będziemy mówić za chwilę. Na razie jednak rozważmy pytanie, czy przyjęcie takiej konwencji pozbawia pierwsze prawo charakteru em-

pirycznego, sprowadzając je do analitycznego zdania *a priori*, prawdziwego na mocy definicji. Na pozór wydaje się, że tak jest – z przyjętego określenia równości odcinków czasowych wynika automatycznie, że ciało pozbawione działania sił będzie poruszać się ruchem jednostajnym. Jednakże pominęliśmy pewien istotny szczegół: pierwsze prawo dynamiki odnosi się do wszystkich „swobodnych" obiektów we wszechświecie, a nie tylko jednego wzorcowego ciała. Nadal istnieje możliwość, że zegar oparty na ruchu wybranego ciała pokaże, iż inne obiekty w analogicznych sytuacjach będą przyspieszać. Treść empiryczna pierwszego prawa ujawnia się w twierdzeniu, że ciała, na które nie działają żadne siły, nie przyspieszają *względem siebie*. Wybierając jedno konkretne ciało swobodne, możemy w arbitralny sposób ustalić, czy jego ruch nazwiemy jednostajnym czy nie. Natomiast nie potrafimy zapewnić, żeby wszystkie swobodne ciała we wszechświecie zachowywały się tak samo jak to wzorcowe.

Przy okazji naszych dyskusji na temat statusu praw dynamiki warto więcej uwagi poświęcić zagadnieniu definiowania terminów. Filozofowie wyróżniają wiele rodzajów definicji, z których najbardziej typowa jest definicja identycznościowa: definiujemy termin przez podanie jego równoważnika (np. woda to substancja o składzie chemicznym H_2O). Jednakże w ogólności definicją terminu może być dowolne zdanie zawierające ten termin, o którym zakładamy, że jest prawdziwe na mocy znaczenia definiowanego terminu (czasem takie definicje nazywa się postulatami znaczeniowymi). Postulaty znaczeniowe muszą spełniać pewne warunki poprawności, z których bodaj najważniejszym jest tzw. warunek nietwórczości. Głosi on, że postulat zawierający definiowany termin *t* nie powinien implikować żadnych konsekwencji niezawierających terminu *t*, poza tautologiami, czyli zdaniami prawdziwymi na mocy logiki. Chodzi o to, aby prawdziwość naszego postulatu nie była uzależniona od przygodnych faktów niezwiązanych ze znaczeniem nadawanym danemu terminowi – faktów, które mogłyby okazać się fałszywe. Gdybyśmy na przykład zdecydowali jako postulat znaczeniowy ustalający znaczenie terminu „równe odcinki czasowe" przyjąć zdanie „Każde ciało, na które nie działa żadna siła, pokonuje równe odcinki drogi w równych odstępach czasu", to byłby to przykład postulatu twórczego, czyli niepoprawnego. Wynika z niego bowiem, że dla każdych dwóch ciał niepoddanych działaniu sił, w czasach wyznaczonych przez równe drogi pokonywane przez pierwsze z nich, drugie ciało również będzie przebywać równe drogi (ciała nie przyspieszają wzajemnie). Jest to jednak, jak wspomnieliśmy, fakt empiryczny, który mógłby nie mieć miejsca. Natomiast postulat „Ciało A, na które nie działa żadna siła, przebywa równe odcinki drogi w równych czasach" jest już nietwórczy. Nie wynika z niego nic na temat zachowania innych ciał w analogicznej sytuacji.

Jak się zatem wydaje, obroniliśmy tezę o empiryczności pierwszego prawa. Pojawia się jednak następna trudność. Zasada bezwładności głosi coś na temat zachowania ciał w sytuacji całkowitego braku działających sił. Jest to jednak sytuacja trudna czy wręcz niemożliwa do realizacji, choćby z powodu wszechobecnej siły grawitacji, przed którą nie ma możliwości ucieczki. Zasadniczo nie możemy więc empirycznie sprawdzić, czy pierwsze prawo obowiązuje, ponieważ nie wyeliminujemy wszystkich działających na dane ciało sił. Innymi słowy, niemożliwe jest empiryczne obalenie pierwszej zasady. Niektórzy wyciągają stąd wniosek, że pierwsze prawo jest rodzajem konwencji, nie tyle znaczeniowej, co metodologicznej, podyktowanej względami elegancji czy prostoty. Po prostu jest nam wygodnie opisywać obser-

wowane zjawiska mechaniczne, przyjmując jako pewnik, że gdyby pozbawić ciało wszelkich działających na nie sił, poruszałoby się ono w idealnie jednostajny, prostoliniowy sposób.

Jednakże wniosek o nieempirycznym charakterze pierwszego prawa wydaje się zbyt pochopny. Po pierwsze, choć niemożliwe jest wyeliminowanie wszystkich działających sił, można je w dobrym przybliżeniu równoważyć, tworząc sytuację, w której wypadkowa siła jest równa lub bliska zeru. W przypadku siły grawitacji jej idealne równoważenie z pozorną siłą inercji występuje w układach znajdujących się w spadku swobodnym (np. na stacji kosmicznej orbitującej wokół Ziemi). To prawda, że nie będziemy w stanie wyeliminować innych sił (np. wpływu nierównomiernie rozłożonego ciśnienia otaczającego powietrza itp.), ale w granicach błędu, z dobrym przybliżeniem możemy poddać pierwsze prawo testowi. Po drugie, nawet w zwykłych warunkach ziemskich możemy pominąć siłę grawitacji, jeśli rozważamy ruch prostopadły do kierunku jej działania.[2] Wreszcie na koniec możemy powiedzieć, że realistyczne testowanie pierwszego prawa ma miejsce w każdej sytuacji, w której stosujemy równania Newtona do wyznaczenia ruchu danego obiektu (np. planety czy rakiety kosmicznej). Pierwsze prawo jest w pewnym sensie częścią każdego modelu w mechanice klasycznej opisującego dane zjawisko. Pokazuje to, że testowanie danej hipotezy w nauce odbywa się w praktyce niemal zawsze „zbiorowo" we współpracy z innymi założeniami.

Przejdźmy teraz do drugiego prawa dynamiki. Występują w nim dwa kluczowe terminy: siła i masa. O masie powiemy więcej przy okazji trzeciego prawa, a na razie zajmijmy się siłą. Czy istnieje sposób scharakteryzowania tego pojęcia niezależny od równania $F = ma$? Newton w *Principiach* zdefiniował siłę ogólnie jako działanie wywierane na ciało w celu zmiany jego stanu spoczynku lub ruchu jednostajnego prostoliniowego. Taka charakterystyka jest niebezpiecznie bliska sfomułowania drugiego prawa, co może prowadzić do wniosku, że w istocie prawo to jest ilościową definicją siły: wypadkowa siła działająca na ciało jest z definicji iloczynem jego masy (inercji) i tempa zmiany prędkości. Jednakże w dalszej części swoich wyjaśnień definicyjnych Newton wymienia szczególny rodzaj sił: siłę dośrodkową, siły ciężkości i magnetyczne. Jest zatem prawdopodobne, że jego intencją było wskazanie na możliwość niezależnych charakterystyk sił. Istotnie, możemy scharakteryzować wiele sił mechanicznych przez podanie odpowiednich formuł. Na przykład korzystając z prawa Hooke'a, które łączy rozszerzanie metali z przyłożoną siłą, możemy obliczyć siłę pochodząca od rozciągniętej sprężyny. Z kolei sam Newton wyprowadził słynny wzór występujący w prawie powszechnej grawitacji (powiemy o nim w jednym z następnych paragrafów), określający siłę grawitacji działającą na dane ciało i pochodzącą od innego ciała. Zatem możliwe jest wstawienie odpowiedniej formuły w miejsce F i eksperymentalne sprawdzenie, czy równość $F = ma$ istotnie zachodzi.

Jednakże problemem jest to, że lista znanych sił (obejmująca także siły elektryczne i magnetyczne) nie jest zamknięta. Możliwe, że nauka nie poznała jeszcze wszystkich istniejących w świecie oddziaływań. W takim razie mamy problem z testowaniem drugiego prawa. Wyobraźmy sobie, że obserwujemy pewne ciało, które nagle z nieznanych powodów zmienia swoją prędkość (zaczyna przyspieszać albo skręca). Żadna ze znanych sił nie działa w rozważanym przypadku. W takiej sytuacji mamy dwie możliwości – albo uznać, że drugie prawo nie obowiązuje powszechnie (jest fałszywe), albo przyjąć istnienie nowej nieznanej siły (czy

[2] W istocie rzeczy neutralizacja siły grawitacji inną siłą jest sprawą dziecinnie prostą – kładziemy przedmiot na stole, a siła grawitacji działająca na ten przedmiot będzie idealnie zrekompensowana siłą reakcji stołu.

sił), której wartość jest dokładnie równa iloczynowi masy i chwilowego przyspieszenia ciała. To drugie rozwiązanie chroni drugie prawo przed obaleniem – zawsze będziemy mogli wytłumaczyć obserwowane zachowanie ciała (zmianę jego prędkości) przez wprowadzenie dodatkowej siły. Jednak ceną, jaką musimy zapłacić za uratowanie drugiej zasady dynamiki, jest pozbawienie jej treści empirycznej i sprowadzenie do konwencji terminologicznej – siła staje się po prostu definicyjnie tożsama z wielkością *ma*.

Istnieje rozwiązanie pośrednie, proponowane przez wielu filozofów, które zachowuje pewną treść empiryczną drugiego prawa, a zarazem dopuszcza możliwość odkrycia nowych sił mechanicznych. Propozycja polega na tym, aby traktować drugie prawo jako pewną regułę metodologiczną, nakazującą poszukiwanie *prostych* funkcji siły w zależności od czasu i położenia, które tłumaczyłyby obserwowalne zachowanie ciał (ich przyspieszenia).[3] Jeśli takie funkcje jesteśmy w stanie odtworzyć na podstawie obserwacji zachowań kinematycznych ciał, drugie prawo uznajemy za spełnione. Możliwe jednak, że ruch obserwowanych ciał będzie tak chaotyczny i nieprzewidywalny, że wprowadzenie „rozsądnej" funkcji siły stanie się praktycznie niewykonalne. W takiej sytuacji będziemy musieli zrewidować nasze prawo łączące przyspieszenie z działającą siłą.

Zajmijmy się teraz pojęciem masy. Newtonowska definicja tego ważnego pojęcia pozostawia wiele do życzenia – w jego ujęciu masa jest miarą ilości materii, obliczoną jako iloczyn gęstości i objętości. Jednakże takie określenie masy jest niezadowalające, gdyż gęstość z kolei definiujemy jako masę ciała na jednostkę objętości. Niektórzy uczeni (Ernst Mach, Henri Poincaré) proponowali, aby do zdefiniowania masy zastosować trzecie prawo Newtona. Rozważmy dwa izolowane ciała A i B, które oddziałują na siebie wzajemnie bez kontaktu z otoczeniem. Zgodnie z trzecim prawem, siła działająca na ciało A jest równa i przeciwnie skierowana w stosunku do siły działającej na ciało B: $F_A = -F_B$. Wynika stąd następująca zależność między przyspieszeniami i masami ciał:

$$\frac{m_A}{m_B} = -\frac{a_B}{a_A}.$$

Równanie to może posłużyć do zdefiniowania relatywnej masy jednego przedmiotu względem drugiego jako stosunku odwrotności ich przyspieszeń. Powstaje znowu pytanie, co w takim razie ze statusem trzeciego prawa dynamiki. Czy staje się ono zdaniem analitycznym, prawdziwym na mocy definicji? Odpowiedź jest podobna, jak w wypadku pierwszego prawa i pojęcia równych odcinków czasowych. Wybierając jedno wzorcowe ciało, możemy określić masy wszystkich innych ciał przez porównanie ich przyspieszeń przy oddziaływaniu z wzorcem. Ta definicja zapewnia, że trzecia zasada będzie spełniona w sytuacji oddziaływania ciała wzorcowego z każdym innym ciałem. Natomiast nie jest sprawą przesądzoną, że podczas oddziaływania dwóch ciał niebędących wzorcami ich wzajemne siły obliczone przy pomocy wyznaczonych wcześniej mas i zarejestrowanych przyspieszeń będą się równoważyć. Zatem trzecie prawo nadal posiada dobrze określoną treść empiryczną.

Powyższa analiza pokazuje, że prawa dynamiki Newtona pełnią (lub ostrożniej, mogą pełnić) podwójną rolę. Można wykorzystać je do scharakteryzowania podstawowych terminów mechaniki, takich jak masa, siła i jednostajność czasowa, ale nie pozbawia to ich roli

[3] Jednym z możliwych warunków prostoty, o którym będziemy jeszcze szczegółowo mówić, jest to, aby funkcja siły dała się przedstawić w postaci tempa zmiany z odległością pewnej wielkości skalarnej (w postaci pochodnej po odległości lub ogólnie w postaci gradientu). Pole takich sił nazywamy zachowawczym.

syntetycznych uogólnień empirycznych, które mówią nam coś o świecie. Pojęcie masy może być zinterpretowane kinematycznie przy pomocy trzeciego prawa dynamiki zastosowanego do oddziaływania danego ciała z wzorcem, którego treść empiryczna będzie wtedy wyrażalna w stwierdzeniu, że dla każdych dwóch ciał oddziałujących ze sobą wzajemnie iloczyn ich uprzednio określonej masy i przyspieszenia będzie taki sam. Z kolei siłę wolno nam scharakteryzować ogólnie jako każdą „rozsądnie" zachowującą się funkcję F, która spełnia równanie $F = ma$. Drugie prawo dynamiki będzie wtedy głosiło, że dla każdego systemu mechanicznego istnieje taka rozsądna funkcja. Wreszcie pierwsze prawo może nam służyć do zdefiniowania dobrego zegara, a jego treść empiryczna sprowadzi się do tezy, że wszystkie ciała swobodne, których ruch mierzony jest tym zegarem, poruszają się ruchem jednostajnym.

2.2. Wyznaczanie trajektorii ruchu

Podstawowym zastosowaniem praw Newtona jest obliczanie przyszłego zachowania mechanicznego ciał, czyli wyznaczanie ich trajektorii. Przyjrzymy się dokładniej sposobowi, w jaki mechanika klasyczna osiąga ten cel, zaczynając od prostego przykładu pojedynczego ciała o pomijalnych rozmiarach (czyli tzw. punktu materialnego). Dla uproszczenia przyjmijmy ponadto, że ewolucja układu w czasie następuje „skokowo", czyli dyskretnie, począwszy od momentu zerowego, przez moment numer jeden i tak dalej. Możemy np. założyć, że ciało zmienia swoje parametry i położenie co sekundę. W chwili zerowej t_0 ciało znajduje się w pewnym położeniu x_0, a działająca na nie siła wynosi F_0. Czy ta informacja wystarcza, aby przewidzieć zachowanie ciała w chwili t_1? Łatwo zauważyć, że nie. Zgodnie z drugim prawem dynamiki, przyłożona siła wyznacza zmianę prędkości, ale aby określić, jaka będzie prędkość ciała w momencie t_1, musimy wiedzieć, jaka była prędkość początkowa. Oznaczmy ją literą v_0. Prędkość v_0 określi nam, gdzie ciało będzie się znajdować w następnej sekundzie, a działająca siła F_0 – jaka będzie jego następna prędkość. Można to wyrazić symbolicznie w następujący sposób:

$$x_1 = x_0 + v_0 \Delta t$$

$$v_1 = v_0 + \frac{F_0 \Delta t}{m},$$

gdzie $\Delta t = t_1 - t_0$ jest interwałem między dwiema kolejnymi chwilami. Pamiętajmy, że w ogólności zarówno położenie, jak prędkość i siła są wektorami, a nie liczbami. Dla uproszczenia będziemy zapisywać nasze równania w postaci liczbowej, porównajcie jednak rysunek 2.2, na którym zobrazowana jest opisana wyżej procedura przy pomocy wektorów.

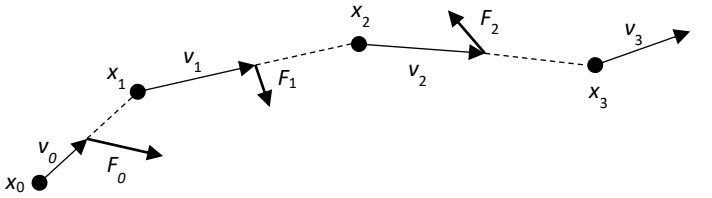
W chwili t_1 sytuacja się powtarza. Znów mamy zadaną prędkość v_1, a do tego nową siłę F_1. Wyznaczą one ewolucję obiektu w kolejnej chwili t_2 – prędkość określi jego położenie, a siła nową prędkość:

$$x_2 = x_1 + v_1 \Delta t$$

$$v_2 = v_1 + \frac{F_1 \Delta t}{m}.$$

Kontynuując tę procedurę dla kolejnych momentów, otrzymamy w rezultacie to, na czym nam zależało, czyli trajektorię przedmiotu, wyznaczoną kolejnymi położeniami $x_0, x_1, x_2, \ldots$

Przedstawiona powyżej procedura nosi nazwę numerycznego obliczania trajektorii i jest z powodzeniem stosowana przy pomocy komputerów do wyznaczania ewolucji skomplikowanych układów mechanicznych, składających się np. z ciał niebieskich – planet, asteroid, komet. Warto zwrócić uwagę, że każdy kolejny krok w ewolucji ciała poddanego działaniu sił jest jednoznacznie wyznaczony przez krok poprzedzający. Dostarcza to nam wskazówki, że prawa mechaniki Newtona mogą mieć deterministyczny charakter.



Rys. 2.2. Wyznaczanie trajektorii metodą numeryczną. Położenie ciała w każdym kolejnym momencie jest zdeterminowane jego prędkością w poprzedzającej chwili, natomiast nowa prędkość powstaje w wyniku dodania wektora siły do poprzedniej prędkości (przyjmujemy upraszczające założenie, że masa ciała, jak również odstępy między kolejnymi chwilami, są jednostkowe)

Numeryczne obliczanie trajektorii krok po kroku można w pewnych okolicznościach zastąpić bardziej elegancką, „globalną" procedurą matematyczną. Przyjrzyjmy się bliżej temu podejściu. Niech $F(t, x)$ oznacza pewną matematyczną funkcję, która określa wartość siły działającej w czasie t na ciało znajdujące się w miejscu x. Drugie prawo Newtona, jak pamiętamy, będzie miało postać równania

$$F(t,x) = ma(t).$$

Musimy teraz skorzystać ze ścisłej matematycznie definicji przyspieszenia. Do tej pory definiowaliśmy przyspieszenie jako stosunek przyrostu prędkości do czasu, w jakim ten przyrost zachodzi. Jednak w ten sposób nie określimy przyspieszenia w danym punkcie czasowym t, a raczej jego uśrednioną wartość w danym przedziale. Aby mówić o tzw. wartościach chwilowych, musimy obliczyć granicę odpowiednich ilorazów przy wartości przedziału czasowego zbiegającego do zera. Jak zapewne wielu z was doskonale wie, otrzymana liczba nazywa się pochodną funkcji prędkości $v(t)$ w punkcie (pojęcie to wprowadzili po raz pierwszy Newton i Leibniz). Stosując powszechnie używaną symbolikę, zapisujemy wzór na przyspieszenie jako funkcję czasu t następująco:

$$a(t) = \frac{dv(t)}{dt}.$$

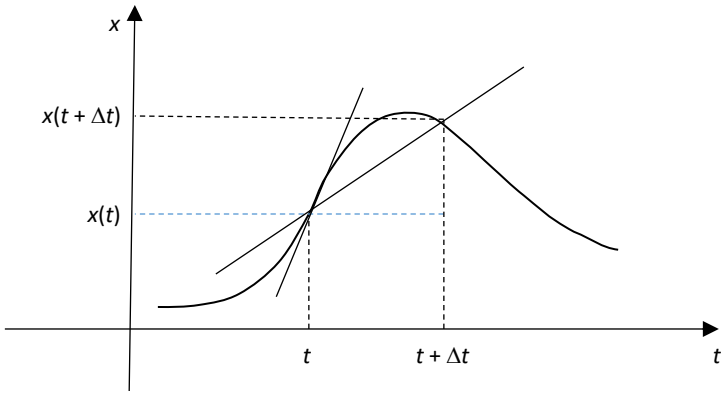
Z kolei prędkość chwilowa $v(t)$ jest w analogiczny sposób definiowana jako pochodna funkcji położenia $x(t)$ po czasie (por. wpis w ramce). Wynika z tego, że w celu obliczenia przyspieszenia musimy podwójnie zróżniczkować funkcję położenia. Taką podwójną pochodną zapisuje się w trochę dziwny sposób:

$$a(t) = \frac{d^2x(t)}{dt^2}.$$

Ostatecznie więc równanie Newtona przyjmuje następującą formę:

$$F(t,x) = m\frac{d^2x(t)}{dt^2}. \tag{2.1}$$

Szukaną wartością jest tutaj postać funkcji położenia $x(t)$, czyli trajektoria.[4] Równania takie nazywamy równaniami różniczkowymi, a ich rozwiązywaniem zajmuje się osobny dział matematyki, należący do tzw. analizy matematycznej. Dla wielu funkcji siły $F(t,x)$ rozwiązanie znajdujemy stosunkowo łatwo, dla innych natomiast jest to analitycznie niemożliwe i wtedy trzeba wrócić do metod numerycznych. Ci, którzy mieli styczność z elementarnym rachunkiem różniczkowym i potrafią obliczyć pochodne z prostych funkcji, takich jak funkcje potęgowe, mogą łatwo sprawdzić, że np. dla stałej funkcji F rozwiązaniem jest funkcja kwadratowa $x(t) = \frac{F}{2m}t^2$. Jest to znany wzór na drogę w ruchu jednostajnie przyspieszonym, antycypowany już przez Galileusza.



Rys. 2.3. Graficzna reprezentacja pochodnej funkcji

Zapewne wielu czytelników zetknęło się wcześniej z pojęciem pochodnej i jej zastosowaniem do wyrażenia pojęcia prędkości chwilowej. Dla tych, którzy nie mieli takiej okazji lub chcieliby odświeżyć nieco pamięć, przygotowałem ten krótki wpis. Niech $x(t)$ będzie funkcją, opisującą położenie ciała w czasie. Prędkością średnią w interwale czasowym od t do $t + \Delta t$ nazwiemy iloraz drogi pokonanej w tym czasie i interwału:

$$\frac{x(t+\Delta t) - x(t)}{\Delta t}.$$

Chcąc obliczyć prędkość chwilową w punkcie t, musimy brać coraz to mniejsze interwały Δt i obliczać odpowiednie ilorazy. Formalnie nazywamy to przechodzeniem w granicy z Δt zbliżającym się do zera. Poszukiwaną granicą będzie liczba, w pobliżu której „gromadzą się" powyższe ilorazy dla coraz mniejszych interwałów. Oczywiście istnieje formalnie poprawna definicja takiej granicy, ale nie będziemy jej tutaj przytaczać. Granicę tę nazwiemy pochodną funkcji $x(t)$ w punkcie t i oznaczymy ją przez $\frac{dx(t)}{dt}$. Istnieją pewne

[4] Oczywiście trzeba pamiętać, że w ogólności równanie to powinno zostać zapisane wektorowo. Innymi słowy, musimy napisać trzy równania, każde osobno dla jednej z współrzędnych x, y i z.

reguły i wzory pokazujące, jak obliczać pochodne znanych funkcji. Na przykład pochodna funkcji potęgowej t^n będzie dana przez nt^{n-1}.

Zwróćmy teraz uwagę na prostą geometryczną interpretację pochodnych. Na diagramie 2.3 przedstawiony został wykres funkcji $x(t)$. Iloraz $\frac{x(t+\Delta t)-x(t)}{\Delta t}$ określa nam kąt nachylenia (dokładnie jego tangens) cięciwy łączącej punkty $x(t)$ i $x(t + \Delta t)$. Zbiegając z Δt do zera, otrzymamy prostą styczną do wykresu w punkcie $x(t)$. Zatem pochodna definiuje nachylenie krzywej $x(t)$ w danym punkcie.

Zadajmy sobie teraz pytanie, czy równanie (2.1) ma dokładnie jedno rozwiązanie. Otóż łatwo pokazać, że odpowiedź jest przecząca. Dla danego rozwiązania równania $x(t)$ istnieje nieskończenie wiele rozwiązań o następującej postaci:

$$\tilde{x}(t) = x(t) + \alpha t + \beta,$$

gdzie α i β są dowolnymi liczbami. Wynika to stąd, że dwukrotne zróżniczkowanie funkcji $\alpha t + \beta$ daje zero, a zatem druga pochodna funkcji $\tilde{x}(t)$ jest taka sama jak dla $x(t)$. Zakładając, że funkcja $x(t)$ i jej pochodna przyjmują wartość 0 dla $t = 0$, otrzymujemy interpretację stałych α i β: stała β jest położeniem x_0 w chwili zerowej, a α – prędkością v_0. Zatem rozwiązania różnią się tym, że zakładają inne wartości początkowe położenia i prędkości. Ustalając początkowe wartości położenia i prędkości, możemy wyeliminować wszystkie funkcje o powyższej postaci z wyjątkiem jednej.

Istnieje matematyczne twierdzenie, które mówi, że dla każdej funkcji siły F spełniającej pewne warunki istnieje dokładnie jedno rozwiązanie $x(t)$ równania Newtona, które w punkcie zerowym przyjmuje daną wartość x_0, a jej pochodna w tym punkcie jest równa v_0. Jest to matematyczna podstawa dla sformułowania wersji determinizmu obowiązującej w mechanice klasycznej. W przypadku jednego ciała determinizm równania Newtona polega na tym, że ustalając położenie i prędkość początkową, ustalamy tym samym całą jednoznacznie wyznaczoną trajektorię w przyszłości. Dla układów wielu ciał oczywiście będziemy potrzebowali położenia i prędkości początkowych każdego z nich z osobna.

Pojawia się jednak tutaj drobny problem, który jest często pomijany w popularnych ujęciach determinizmu. Co z funkcją siły? Czy nie potrzebujemy także ustalić wszystkich sił działających na każde ciało w chwili początkowej? I czy takie ustalenie początkowej siły wystarcza? W istocie potrzebujemy znacznie więcej – musimy wiedzieć z wyprzedzeniem, jakie siły będą działały na nasz układ w przyszłości. Powstaje zatem wątpliwość, czy możemy w ogóle mówić o determinizmie, skoro tak ważny element, jak działające w przyszłości siły, musi być dodany „ręcznie". Odpowiedzią jest tutaj przyjęcie dwóch dodatkowych założeń, z których drugie jest zwykle ignorowane w filozoficznych ujęciach determinizmu mechaniki klasycznej. Pierwszym założeniem jest teza o izolowaniu rozważanego układu od wszelkich wpływów zewnętrznych. Determinizm w odniesieniu do wybranego układu wielu ciał musi być opatrzony tym założeniem, w przeciwnym razie będziemy musieli uwzględnić możliwość zupełnie nieprzewidywanego wpływu otoczenia, niewynikającego ze stanu początkowego układu. W przypadku determinizmu globalnego, czyli tezy deterministycznej dotyczącej całego wszechświata, założenie o izolacji jest zbędne, gdyż dla wszechświata nie istnieje zewnętrzne otoczenie, które mogłoby wpłynąć na jego ewolucję.

Drugie założenie jest takie, że w izolowanym układzie składającym się z n ciał, wszystkie siły działające na poszczególne ciała są funkcjami położeń i (ewentualnie) prędkości tych

ciał. Innymi słowy, znając chwilowe położenia i prędkości ciał, możemy obliczyć wszystkie działające w tym momencie siły. Przy tym założeniu, położenia i prędkości początkowe ciał pełnią podwójną rolę w wyznaczaniu przyszłego zachowania układu. Po pierwsze, jak już powiedzieliśmy, stanowią one niezbędny element pozwalający na wybór jednego spośród nieskończenie wielu rozwiązań równania ruchu Newtona. Po drugie, determinują one również początkowe siły działające na ciała. Zauważmy, że to dodatkowe założenie nie wynika w żadnym razie z praw dynamiki Newtona – mogło by się tak zdarzyć, że siły w przyrodzie brałyby się „znikąd", a drugie prawo byłoby nadal spełnione.[5] Sprowadzalność sił do wzajemnych położeń i prędkości ciał jest częścią szerszej doktryny, znanej pod nazwą mechanicyzmu.

2.3. Dwie wersje determinizmu

Na razie nie sprecyzowaliśmy jeszcze, co głosi filozoficzne stanowisko determinizmu. Będziemy musieli podjąć to zadanie, aby móc odpowiedzieć na pytanie, czy mechanika klasyczna jest teorią deterministyczną, jak się powszechnie uważa. Jak zwykle bywa z kluczowymi terminami filozofii, istnieje wiele nierównoważnych interpretacji pojęcia determinizmu. Zacznijmy od najbardziej rozpowszechnionego ujęcia, odwołującego się do kwestii przewidywalności. Nazwiemy tę wersję determinizmem epistemologicznym lub też laplasjańskim, od nazwiska Pierra Simone'a de Laplace'a. Laplace był wybitnym matematykiem, fizykiem i filozofem francuskim przełomu osiemnastego i dziewiętnastego stulecia, Jego prace w znacznym stopniu przyczyniły się do rozwoju mechaniki teoretycznej. Był on zafascynowany możliwościami mechaniki przy wyznaczaniu przyszłego zachowania ciał niebieskich – planet czy komet. Ekstrapolując sukcesy mechaniki niebieskiej na inne działy nauki, ogłosił, że zasadniczo nie ma przeszkód, aby jej metody zastosować do przewidywania wszelkiego typu przyszłych zdarzeń we wszechświecie. Jedynym warunkiem takiego przewidywania jest dokładne poznanie stanu świata w danej chwili oraz wykonanie odpowiednich obliczeń. Zatem determinizm epistemologiczny może być przedstawiony w formie następującego twierdzenia:

> Znajomość stanu początkowego danego układu fizycznego oraz wszystkich praw rządzących tym układem umożliwia obliczenie jego wszystkich przyszłych stanów.

Kluczowym pojęciem zastosowanym w powyższym sformułowaniu jest pojęcie „możliwości". Jak należy rozumieć możliwość obliczenia stanów przyszłych na podstawie teraźniejszych? Laplace miał świadomość, że praktyczne zrealizowanie predykcji przyszłych stanów może być dla nas niewykonalne. Są ku temu dwa powody. Po pierwsze, nie możemy poznać w najdrobniejszych szczegółach momentalnych stanów początkowych dla skomplikowanych układów fizycznych. Gdyby nawet to się nam udało, obliczenia konieczne do sformułowania przewidywań mogą przekroczyć możliwości nawet najpotężniejszych maszyn liczących. Stąd też Laplace w swoim słynnym sformułowaniu zasady determinizmu odwołał się do hipotetycznej „inteligencji", której zdolności percepcyjne i obliczeniowe są nieporów-

[5] Wprowadzenie sił, które zależą od innych parametrów poza konfiguracjami ciał, nie jest wcale rzadkością w fizyce współczesnej. Na przykład interpretacja mechaniki kwantowej, znana pod nazwą mechaniki Bohmowskiej, uzależnia trajektorie cząstek od pola sił, reprezentowanego przez tzw. funkcję falową. Wspomnimy o tym w rozdziale poświęconym mechanice kwantowej.

nanie większe od zdolności ludzkich. Utarło się określać tę inteligencję mianem „demona Laplace'a", chociaż sam francuski uczony nigdy nie używał tego terminu. Jednakże należy podkreślić, że wprowadzenie „demona" o bliżej niesprecyzowanych nadnaturalnych zdolnościach pozbawia zasadę determinizmu jasnej treści empirycznej. Sprowadzając problem do absurdu, możemy powiedzieć, że demon o nadludzkich zdolnościach mógłby po prostu mieć bezpośredni wgląd w przyszłe zdarzenia. Tak czy inaczej, epistemologiczna interpretacja determinizmu wikła nas w trudny problem obliczalności. Pojęcie obliczalności jest intensywnie analizowane przez logików i matematyków. W matematyce dowodzi się interesujących twierdzeń limitacyjnych, które pokazują niemożliwość rozstrzygnięcia pewnych kwestii matematycznych za pomocą algorytmicznych procedur obliczeniowych.[6] Nie jest wykluczone, choć nie zostało to zbadane, że podobne ograniczenia stosują się do przewidywań opartych na równaniach mechaniki klasycznej.

Najsłabsza interpretacja pojęcia możliwości w epistemologicznej wersji determinizmu głosi, że w świecie zachodzą minimalne warunki konieczne do tego, aby dokonać poprawnej predykcji. Takie minimalne warunki konieczne (choć z całą pewnością niewystarczające) udanej predykcji mogą przyjąć postać determinizmu w wersji ontologicznej. Jest dość oczywiste, że gdyby niezależny od podmiotu świat nie był deterministyczny, przewidzenie przyszłego toku zdarzeń byłoby z zasadniczych, a nie tylko praktycznych powodów niemożliwe. Determinizm w odniesieniu do świata, a nie podmiotów poznających, może być sformułowany w następujący sposób:

> Dla każdego kompletnego stanu układu fizycznego w chwili początkowej istnieje dokładnie jeden ciąg przyszłych stanów układu zgodny z tym stanem początkowym i prawami przyrody.

Gdyby warunek powyższy nie był spełniony, tj. gdyby przy zadanych warunkach początkowych i prawach istniały różne alternatywne ciągi przyszłych zdarzeń, przewidywalność by upadła. Z taką sytuacją zetkniemy się później przy okazji omawiania mechaniki kwantowej, natomiast na razie pozostańmy na gruncie fizyki Newtonowskiej. Uporządkowany czasowo ciąg stanów możemy w skrócie nazwać ewolucją układu, zatem powyższe sformułowanie determinizmu głosi, że dla danego stanu początkowego istnieje dokładnie jedna, z góry określona przez prawa przyrody ewolucja.

Jak pamiętamy z poprzedniego paragrafu, układy fizyczne opisywane równaniami Newtona zasadniczo spełniają warunek determinizmu ontologicznego (przy dodatkowych założeniach, o których mówiliśmy wcześniej). Jest to konsekwencją twierdzenia o jedyności rozwiązań równań ruchu – jedyna funkcja $x(t)$, która przyjmuje wartości początkowe x_0 i v_0, określa jednoznacznie przyszłą ewolucję układu. Dodajmy, że mechanika klasyczna w żadnym stopniu nie wyróżnia przyszłości względem przeszłości. Tradycyjnie przyjęło się formułować tezę determinizmu w odniesieniu do przyszłości – ma to oczywiście uzasadnienie praktyczne, gdyż na ogół jesteśmy zainteresowani tym, co ma się dopiero wydarzyć, a nie tym, co już miało miejsce. Jednakże równania mechaniki Newtona są całkowicie syme-

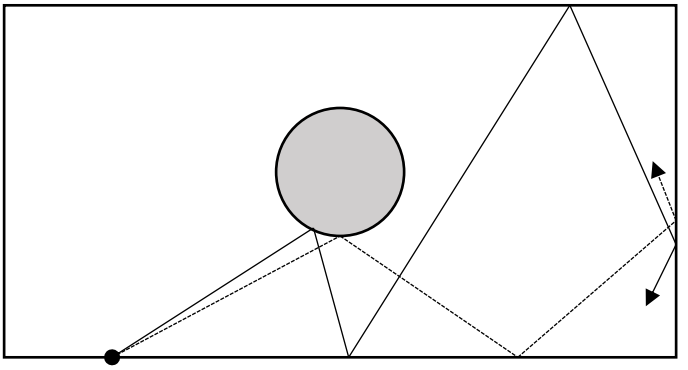
[6] Najsłynniejszym twierdzeniem tego typu jest twierdzenie Gödla, które stwierdza w najogólniejszych zarysach, że dla każdej odpowiednio skomplikowanej teorii matematycznej istnieją twierdzenia, których nie da się wyprowadzić z aksjomatów tej teorii. Przykładem nieobliczalności matematycznej jest np. zagadnienie istnienia całkowitych pierwiastków wielomianów o współczynnikach całkowitych.

tryczne względem czasu.[7] Oznacza to, że ustalenie wartości x_0 i v_0 w pewnej chwili jednoznacznie wyznacza trajektorię obiektu nie tylko w przyszłości, ale także w przeszłości.

Pojęcie stanu fizycznego układu w danej chwili (stanu momentalnego) jest niezmiernie ważnym elementem doktryny deterministycznej. Ogólnie stanem układu możemy nazwać każdą jego kompletną i nieredundantną charakterystykę w danym momencie. Kompletność oznacza, że z informacji na temat stanu w chwili *t* można wyprowadzić wszystkie fakty na temat układu w *t*, natomiast warunek nieredundantności zapewnia, że żaden element stanu nie jest zbędny. Jak pamiętamy, w mechanice klasycznej stanem układu *n* cząstek punktowych są dokładne wartości ich położeń i prędkości chwilowych. Pojawia się jednak pytanie, czy te charakterystyki są istotnie momentalne, czy też może zawierają ukrytą informację o zachowaniu układu przed i po danej chwili. Kwestia ta jest dość istotna z punktu widzenia tezy deterministycznej, gdyż gdyby okazało się, że momentalny stan zawiera w jakiejś formie taką informację, determinizm mógłby być trywialnie spełniony (jeśli np. stan układu w chwili początkowej zawiera informację, że za pięć sekund stan będzie taki a taki, nic dziwnego, że stan początkowy determinuje jednoznacznie stan po pięciu sekundach). W tym kontekście problematyczne jest użycie prędkości chwilowych jako elementów stanów momentalnych. Już Zenon z Elei w swoim słynnym paradoksie strzały zadał pytanie, czy ruch (a zatem niezerowa prędkość) może być przypisany nierozciągłym momentom. Wyobraźmy sobie „zamrożenie" poruszającego się ciała w pewnej chwili, jak na pojedynczej zastopowanej klatce filmowej. Oczywiste jest, że ciało to będzie miało dobrze określone położenie, ale czy jest sens mówić o jego prędkości? Obserwacyjnie ciało to nie różni się od spoczywającego, zatem aby stwierdzić, że mamy do czynienia z ruchem, należy „puścić" zatrzymany film.

Matematyczne pojęcie prędkości chwilowej, jak wytłumaczyliśmy w poprzedniej ramce, definiowane jest za pomocą pochodnej funkcji położenia. Może to sprawiać wrażenie, że prędkość zależy od położenia, zatem stan układu byłby redundantny. Jednak tak nie jest, jeśli ograniczymy się do jednego punktu. Wartość funkcji położenia $x(t)$ w danej chwili nie nakłada żadnych ograniczeń na wartość prędkości w tej chwili. Pochodna funkcji jest określona przez styczną do wykresu, a zatem aby ją obliczyć, musimy wziąć pod uwagę położenia ciała w chwilach innych niż wybrany moment t_0. Wydaje się, że implikuje to, iż prędkość chwilowa zawiera informację na temat przyszłych stanów! Jednak sprawa nie jest taka prosta. Informacja na temat prędkości chwilowej nie pozwala nam na wyprowadzenie żadnej konkretnej wartości położenia w żadnym późniejszym momencie *t*. Zawsze możemy bowiem wziąć funkcję położenia w mniejszym przedziale wokół chwili t_0, niezawierającym późniejszej chwili *t,* co wystarcza do określenia prędkości chwilowej. Prędkość chwilowa należy do kategorii wielkości charakteryzowanych infinitezymalnie – do jej określenia nie wystarcza informacja o położeniu w chwili t_0, ale każda informacja o położeniu w innej konkretnej chwili jest redundantna, gdyż można ją odrzucić.

[7] Formalnie znaczy to, że jeśli funkcja $x(t)$ jest rozwiązaniem równania Newtona, to funkcja $x(-t)$ jest również jego rozwiązaniem. Oczywiście nadal pozostaje prawdą, że dla danych wartości x_0 i v_0 tylko jedna funkcja $x(t)$ jest rozwiązaniem. Aby wprowadzić funkcję „odwrotną", należy odwrócić znak prędkości początkowej v_0 zamieniając ją na $-v_0$. Wtedy przyszła ewolucja układu od chwili zerowej będzie dokładnym odwróceniem jego przeszłej ewolucji do momentu zero, jak na filmie puszczonym do tyłu.



Rys. 2.4. Przykład układu chaotycznego („stół bilardowy"). Dwie trajektorie, które różnią się minimalnie w chwili początkowej, prowadzą do zupełnie odmiennych ewolucji

Jak wygląda kwestia przewidywalności na gruncie mechaniki klasycznej? Okazuje się, że determinizm epistemologiczny popada w jeszcze poważniejsze kłopoty niż te omówione na początku paragrafu. Związane to jest ze zjawiskiem tzw. deterministycznego chaosu, który stał się przedmiotem intensywnych badań w ostatnich dziesięcioleciach. Na pierwszy rzut oka pojęcie deterministycznego chaosu wydaje się wewnętrznie niespójne – jak można mówić o zjawiskach, które są deterministyczne, czyli uporządkowane, a mimo to chaotyczne? Jednakże terminologia ta ma dobre uzasadnienie. W najogólniejszym ujęciu, chaotyczny układ mechaniczny to taki układ, którego deterministyczna ewolucja jest „wrażliwa" na znikome różnice w warunkach początkowych. Oznacza to, że niewielka, nawet niezauważalna zmiana warunków początkowych skutkuje ogromną różnicą w późniejszej ewolucji układu. Efekt ten nazywany jest często „efektem motyla": trzepot skrzydeł motyla w dżungli amazońskiej może wywołać później huragan u wybrzeży Ameryki Północnej. Typowym przykładem chaotycznego systemu jest oczywiście atmosfera ziemska wraz ze wszystkimi skomplikowanymi zjawiskami pogodowymi. Są jednak dużo prostsze przykłady ilustrujące to zjawisko. Na przykład jeśli ustawimy na stole bilardowym przeszkodę w kształcie walca, jak na rysunku, to okaże się, że tor pojedynczej kuli bilardowej będzie w niesłychanie silny sposób zależał od niewielkich różnic w kącie, pod jakim została wystrzelona bila (rys. 2.4). Zatem układy chaotyczne wcale nie muszą być złożone z wielu elementów – mogą składać się z pojedynczego ciała poddanego działaniu standardowych sił mechanicznych.

Przewidywalność w systemach chaotycznych załamuje się z prostego powodu: znajomość stanu początkowego jest zawsze ograniczona dokładnością urządzeń pomiarowych. Błędy pomiarowe są nieusuwalnym elementem każdej procedury eksperymentalnej. Jeśli zatem stan początkowy jest znany tylko z pewnym marginesem błędu, to należałoby oczekiwać, że stany przyszłe będą podobnie wyznaczone z pewnym przybliżeniem. Jednakże systemy chaotyczne rozszerzają ten margines błędu w stopniu, który uniemożliwia przyszłe przewidywanie. Mimo to ewolucja systemu jest ściśle rzecz biorąc deterministyczna – każdy dokładnie określony stan początkowy jednoznacznie wyznacza jego ewolucję.

Pozostaje nam zatem determinizm w wersji ontologicznej, który wydaje się zagwarantowany prawami mechaniki klasycznej. Czy jednak na pewno nie ma możliwości jego złamania przez układ mechaniczny rządzony prawami Newtona? Okazuje się, że takie przypadki są

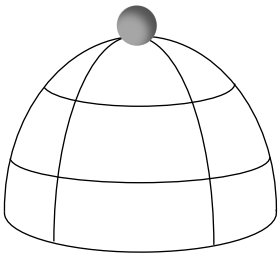
znane. Pamiętamy, że matematyczne twierdzenie o jedyności rozwiązania dla równania różniczkowego (2.1) opatrzone było komentarzem „dla funkcji siły spełniającej pewien warunek". Może się wydawać, że siły wyrażalne prostymi matematycznymi formułami powinny spełniać ten warunek bez większego problemu. Są jednak wyjątki. Jest prosta funkcja siły, dla której istnieje nieskończenie wiele rozwiązań równań ruchu o zadanych warunkach początkowych. Funkcja taka wygląda następująco:

$$F(x) = C\sqrt{x},$$

gdzie *C* jest pewną stałą. Można stosunkowo łatwo udowodnić, że rozwiązaniami równania Newtona dla tak dobranej siły są wszystkie funkcje czasu o postaci

$$x(t) = \begin{cases} 0 \text{ dla } t < \tau \\ \dfrac{C^2}{144m^2}(t-\tau)^4 \text{dla } t \geq \tau, \end{cases}$$

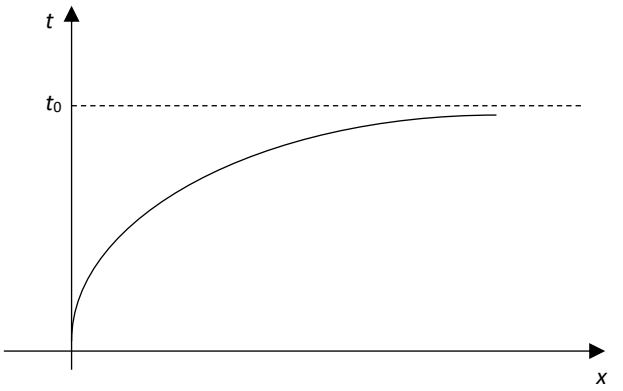
gdzie τ to pewna ustalona liczba, a *m* – masa obiektu. Zachęcam Czytelników zaznajomionych z regułami różniczkowania funkcji potęgowych do sprawdzenia, że w wyniku dwukrotnego zróżniczkowania tak określonej funkcji otrzymamy dokładnie *C* razy pierwiastek z tej funkcji podzielony przez *m*, czyli spełnione jest równanie ruchu Newtona $F = ma$. Sens fizyczny powyższej funkcji $x(t)$ jest taki, że opisuje ona ciało, które do chwili τ pozostaje nieruchome, po czym nagle zaczyna się poruszać z coraz to większą prędkością (przebywając drogę proporcjonalnie do czwartej potęgi czasu). Istotne jest, że parametr τ może przybierać dowolną wartość, a zatem czas, w którym ciało zacznie się poruszać, nie jest zdeterminowany zasadami mechaniki newtonowskiej. Jest to niewątpliwy przykład zachowania indeterministycznego.



Rys. 2.5. Kopuła Nortona

Można zadać pytanie, czy siły takie jak powyższa mają dobrze określony sens fizyczny. Na to pytanie twierdzącej odpowiedzi udzielił John Norton, współczesny filozof fizyki. Pokazał on, że istnieje prosty fizyczny model realizujący podaną wyżej matematyczną formę funkcji siły. Model ten to tak zwana kopuła Nortona, czyli odpowiednio zakrzywiona powierzchnia o kształcie kopuły, umieszczona w polu grawitacyjnym (zakładamy ponadto, że kopuła jest idealnie gładka, pozbawiona wszelkiego tarcia). Jej kształt opisuje precyzyjnie dobrana funkcja, która gwarantuje, że składowa siły grawitacji równoległa do powierzchni

będzie miała dokładnie wartość $a\sqrt{x}$, gdzie x jest odległością od szczytu mierzoną po krzywiźnie (rys. 2.5). Zatem z równań mechaniki klasycznej wynika, że dla ciała umieszczonego w wierzchołku istnieje nieskończenie wiele możliwych ewolucji. Ciało to może pozostać w punkcie wyjścia, ale może zacząć się zsuwać w dowolnym momencie – za minutę lub za sto lat.[8]



Rys. 2.6. Przypadek „najeźdźców z kosmosu". Przez odwrócenie czasowe przedstawionej powyżej sytuacji otrzymujemy niedeterministyczny proces

Rzecz jasna, w praktyce nie jesteśmy w stanie skonstruować idealnej kopuły Nortona, gdyż nawet niewielkie odstępstwa od matematycznie zdefiniowanej krzywizny spowodują, że powyższe rozumowanie przestanie obowiązywać. Jednakże taka sytuacja nie jest wykluczona żadną ogólną zasadą, a zatem determinizm teorii newtonowskiej nie jest sprawą przesądzoną. Istnieją inne przykłady systemów dopuszczalnych przez teorię Newtona, których zachowanie ma charakter niedeterministyczny. Jednym z nich jest przypadek tzw. kosmicznych najeźdźców. Wyobraźmy sobie ciało poddane działaniu siły, która rośnie wraz z czasem w takim tempie, że prędkość przyspieszanego ciała wzrasta do nieskończoności w skończonym przedziale czasowym (rys. 2.6). Oznacza to, że wykres drogi $x(t)$ w funkcji czasu będzie asymptotycznie zbiegać do linii o współrzędnej czasowej t_0 (ciało tak przyspieszone „zniknie" na zawsze z horyzontu zdarzeń, nigdy nie osiągając chwili t_0). Odwracając ten proces w czasie, co jest zawsze dozwolone w mechanice klasycznej, otrzymujemy sytuację, w której w chwili t_0 z nieskończoności nadleci obiekt (np. statek kosmiczny z obcymi najeźdźcami). Ponieważ chwila t_0 może być wybrana zupełnie arbitralnie, przylot najeźdźców nie jest w żaden sposób zdeterminowany przeszłym stanem wszechświata.

Zakończmy ten przegląd niedeterministycznych sytuacji w mechanice klasycznej innym, bardziej przyziemnym przykładem. Nie wszystkie sytuacje, z którymi się stykamy w mechanice, dają się opisać za pomocą równań Newtona z odpowiednio dobranymi siłami. Dobrze

[8] Jak to możliwe, że ruch w ogóle nastąpi, skoro w chwili początkowej prędkość ciała umieszczonego na szczycie kopuły wynosi zero, a działająca chwilowa siła jest również zerowa? Na to pytanie nie ma innej odpowiedzi jak ta, że równania „wiedzą lepiej". Nasze intuicje oparte na zdrowym rozsądku muszą ustąpić przed argumentami matematycznymi – skoro istnieje rozwiązanie równania Newtona opisujące przedmiot, który zaczyna się zsuwać z kopuły, to widocznie taka sytuacja jest możliwa.

znanym zjawiskiem tego typu są sprężyste zderzenia ciał, wymagające zastosowania nieskończonych sił, aby opisać momentalną zmianę pędu w wyniku zderzenia. Aby przewidzieć zachowanie ciał po takim zderzeniu, korzystamy z innych praw mechaniki, takich jak zasada zachowania pędu (momentu pędu) czy energii. Okazuje się jednak, że zasady te nie pozwalają nam na jednoznaczne przewidzenie, jak zachowają się trzy ciała (np. kule bilardowe) po idealnie symetrycznym, jednoczesnym zderzeniu. Oczywiście musimy podkreślić, że wszystkie rozważane powyżej przykłady opierają się na mocno idealizacyjnych założeniach, które nie są ściśle rzecz biorąc możliwe do spełnienia. Zatem można nadal utrzymywać, że realny świat mechaniki newtonowskiej jest *de facto* deterministyczny. Jednakże jeśli przymiotnik „deterministyczny" chcemy zastosować do teorii, to musimy się mocno zastanowić, zanim orzekniemy, że teoria Newtona jest teorią deterministyczną. W najlepszym razie będziemy musieli opatrzeć takie twierdzenie dodatkowymi zastrzeżeniami czy ograniczeniami.

2.4. Układy odniesienia i transformacja Galileusza

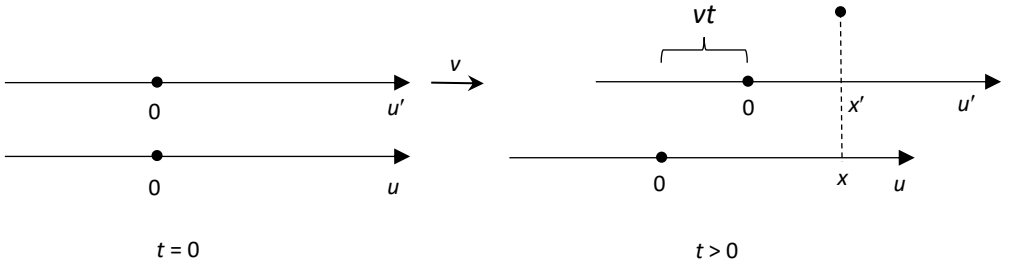
Przejdźmy teraz do kolejnego ważnego problemu mechaniki klasycznej, jakim jest zagadnienie czasu i przestrzeni. Wiemy, że podstawowe równania teorii Newtona zawierają dwa kluczowe parametry: położenie w przestrzeni x i lokalizację czasową t. Obecnie zastanowimy się nad ich dokładniejszą interpretacją, rozważając tę kwestię najpierw z punktu widzenia fizycznego, a potem przechodząc na grunt bardziej filozoficzny. Dla fizyka jest rzeczą oczywistą, że lokalizacje w czasie i przestrzeni wymagają pewnego punktu odniesienia – pytamy, jaka jest lokalizacja danego obiektu względem innych przedmiotów, a nie „sama w sobie". Aby uniknąć jawnego odwołania do konkretnych przedmiotów, wprowadza się abstrakcyjne pojęcie „układu odniesienia". Układem odniesienia może być zasadniczo każdy zespół obiektów umożliwiający dokonywanie pomiarów odległości przestrzennych i interwałów czasowych. Na ogół jednak wyobrażamy sobie układy odniesienia analogicznie do kartezjańskich układów współrzędnych, w postaci trzech prostych prostopadłych do siebie i zaczepionych w pewnym punkcie. Takie proste, czyli osie współrzędnych x, y i z, umożliwiają lokalizację przestrzenną każdego chwilowego obiektu przez podanie jego współrzędnych. Do tego należy oczywiście dodać zegar, za pomocą którego możemy również określić współrzędną czasową zdarzeń.

Z powyższego określenia widać wyraźnie, że istnieje wiele układów odniesienia, w których możemy opisywać procesy zachodzące w czasie i przestrzeni. Niektóre z tych układów będą się różnić w nieistotny z punktu widzenia fizyka sposób – na przykład osie jednego układu mogą być przesunięte względem drugiego albo obrócone. Najciekawsze są jednak układy, które różnią się od siebie stanem ruchu, tj. jeden porusza się względem drugiego. Na przykład jeden układ może być na stałe powiązany z powierzchnią Ziemi, a drugi z lecącym samolotem. Pytanie, jakie sobie postawimy, jest następujące: jak powiązać lokalizacje przedmiotów określone w dwóch układach odniesienia poruszających się względem siebie? Odpowiedź na to pytanie zawarta jest w tzw. transformacjach układowych, czyli formułach, które łączą współrzędne „nowego" układu x', y', z' i t' ze współrzędnymi układu starego x, y, z, t. Wyprowadzimy te transformacje dla najprostszego przypadku, kiedy układ odniesienia u' porusza się względem układu u ze stałą prędkością v wzdłuż osi x (w kierunku rosnących wartości x). Dla uproszczenia dodatkowo przyjmijmy, że kierunki osi układu u' pokrywają się z kierunkami układu u, oraz że w chwili $t = 0$ początki układów się pokrywają. Łatwo

zauważyć, że przy przyjętych założeniach jedynie współrzędna x' będzie się różnić od x (pozostałe dwie współrzędne przestrzenne pozostaną niezmienione). Z poniższego diagramu (rys. 2.7) wnosimy, że formuła łącząca x' z x będzie następująca:

$$x' = x - vt.$$

Dodajmy do tego intuicyjne założenie, że czas nie ulega zmianie przy przejściu z jednego układu do drugiego (wskazania zegara pozostają takie same niezależnie od tego, czy znajdujemy się na powierzchni Ziemi, czy w lecącym samolocie). Pełny zestaw równań obejmujący powyższe równanie oraz proste tożsamości $y' = y$, $z' = z$, $t' = t$ nazywa się transformacją Galileusza.



Rys. 2.7. Wyprowadzenie transformacji Galileusza

Mając do dyspozycji transformację Galileusza, możemy teraz wprowadzić fundamentalne pojęcie *inwariantu*. Inwariantem (niezmiennikiem) danej transformacji nazwiemy każde pojęcie, które nie ulega zmianie przy przejściu z jednego układu do drugiego. Pojęcie niezmiennika jest ważne z filozoficznego punktu widzenia, gdyż umożliwia odróżnienie realnych własności świata i przedmiotów fizycznych od własności „pozornych". Ogólnie przyjmuje się, że jeśli jakaś wielkość zmienia się przy przejściu od jednego układu odniesienia do drugiego, znaczy to, że nie odpowiada jej obiektywnie istniejąca rzeczywistość. Różne układy odniesienia traktujemy podobnie jak różne jednostki wyrażające tę samą wielkość, np. metry czy stopy do mierzenia odległości. Obiektywne fakty na temat świata nie powinny zależeć od tego, jakimi jednostkami się posługujemy, czyli muszą pozostać takie same przy zmianie jednostek.

Rozważmy dwa ważne przykłady ilustrujące problem niezmienniczości względem transformacji Galileusza. Pierwszym z nich będzie prędkość danego obiektu. Czy jest ona niezmiennikiem transformacji Galileusza? Łatwo się przekonać, że nie. Niech ciało o porusza się ze stałą prędkością v względem pewnego układu odniesienia w kierunku osi x. Znaczy to, że prędkość tego ciała wyrażona będzie formułą $v = \frac{x}{t}$. Jaka będzie prędkość ciała o względem drugiego układu poruszającego się w stosunku do pierwszego z prędkością V? Aby to policzyć, musimy zastosować transformację Galileusza do współrzędnych x' i t'. W rezultacie otrzymamy dobrze znany wzór na nową prędkość v':

$$v' = \frac{x'}{t'} = \frac{x - Vt}{t} = v - V.$$

Jest to wzór na składanie prędkości: aby obliczyć nową prędkość ciała, należy od starej odjąć prędkość drugiego układu względem pierwszego. Zatem prędkość nie jest wielkością absolutną (zauważył to już Galileusz).

Omówmy teraz drugi przykład, którym będzie przyspieszenie. Skorzystamy w tym celu z uniwersalnej formuły na przyspieszenie chwilowe, które jest dane przy pomocy drugiej pochodnej z funkcji położenia: $a = \frac{d^2x}{dt^2}$. W układzie „primowanym" będzie to oczywiście formuła wyrażona nowymi współrzędnymi: $a' = \frac{d^2x'}{dt'^2}$. Wstawiając w miejsce x' formułę $x - vt$ zauważamy, że podwójne zróżniczkowanie tej funkcji daje nam dokładnie drugą pochodną z x (druga pochodna funkcji vt po czasie wynosi zero). Zatem przyspieszenie nie zmienia się podczas transformacji Galileusza. Jest ono inwariantem. Wynika z tego również ważna konsekwencja dotycząca drugiego prawa Newtona. Ponieważ prawo to ma formę $F = ma$, czyli wykorzystuje przyspieszenie, a nie np. prędkość, jego prawdziwość nie zmieni się przy zastosowaniu transformacji Galileusza, o ile nie zmienią się działające siły. Zatem prawo dynamiki opisuje realną regularność, zachodzącą niezależnie od tego, w jakim układzie jest ono opisywane.

Należy jednak zwrócić uwagę na istotne ograniczenie powyższych analiz. Transformacja Galileusza łączy współrzędne układów poruszających się względem siebie ruchem jednostajnym i prostoliniowym. Co jednak, kiedy rozważymy zupełnie dowolny ruch, na przykład obrotowy albo prostoliniowy, lecz przyspieszający? Okazuje się, że sprawa się komplikuje. Dla dowolnego ruchu względnego można oczywiście wyprowadzić odpowiednie wzory transformacyjne, ale będą one dużo bardziej złożone niż prosta transformacja Galileusza. Jednakże ważniejsza od stopnia komplikacji jest kwestia inwariantów. W ogólnym wypadku przyspieszenie nie będzie inwariantem transformacji międzyukładowych, a zatem także i drugie prawo Newtona przestanie obowiązywać uniwersalnie. Aby temu zaradzić, wprowadza się ważne pojęcie układów inercjalnych. Zgodnie z popularnym ujęciem, układ inercjalny to taki, który nie przyspiesza. Nie jest to jednak poprawna definicja, gdyż przyspieszenie jest zasadniczo pojęciem względnym (np. żaden układ nie przyspiesza względem siebie). Bardziej poprawne, choć nieco abstrakcyjne, jest określenie układów inercjalnych jako pewnej wyróżnionej klasy układów, z których każde dwa pozostają względem siebie w ruchu jednostajnym i prostoliniowym. Zatem transformacje Galileusza stosują się bez wyjątku do wszystkich układów inercjalnych.

Na czym jednak polega owo „wyróżnienie" układów inercjalnych? Kryterium inercjalności oparte jest na pojęciu sił pozornych (zwanych również, choć nieco myląco, inercjalnymi), a ogólniej na kwestii stosowalności drugiego prawa dynamiki. Generalnie idea jest taka, że jeśli uwzględnimy wszystkie „normalne" siły działające na dane ciało i obliczymy wynikające z ich działania przyspieszenie, a okaże się, że ciało ma pewne dodatkowe przyspieszenie niewynikające z drugiej zasady dynamiki, to przypiszemy dodatkowe przyspieszenie właśnie działaniu sił pozornych. Dobrze znanymi z codziennego życia siłami pozornymi są siły występujące w pojazdach przyspieszających lub hamujących, a także siły odśrodkowe i wspomniana w poprzednim rozdziale siła Coriolisa. Siły pozorne działają tak samo na każde ciało (tj. nadają każdemu ciału dokładnie takie samo przyspieszenie). Układy, w których siły pozorne nie występują, a zatem prawo Newtona przyjmuje w nich najprostszą możliwą formę, nazywamy właśnie inercjalnymi.

Rozważmy najprostszy przykład układu nieinercjalnego, jakim jest układ poruszający się ze stałym przyspieszeniem *a* względem innego układu inercjalnego. W takim wypadku odpowiednia transformacja współrzędnej przestrzennej będzie wyglądała następująco:

$$x' = x - \frac{at^2}{2}.$$

Przyspieszenie danego ciała względem układu *u*′ wynosi (po dwukrotnym zróżniczkowaniu powyższej formuły):

$$\frac{d^2x'}{dt'^2} = \frac{d^2x}{dt^2} - a\,,$$

a zatem pojawi się nowy składnik przyspieszenia równy i przeciwnie skierowany do przyspieszenia całego układu. Aby zachować poprawność równania Newtona $F = m\frac{d^2x}{dt^2}$, musimy w układzie *u*′ dodać do zwykłej siły *F* siłę pozorną wynoszącą $-ma$.

2.5. Czas i przestrzeń w mechanice klasycznej

Problem czasu i przestrzeni, który będziemy rozważać w niniejszym paragrafie, sprowadza się do pytania, czy czas i przestrzeń są absolutne czy relatywne. Pytanie to można rozumieć na dwa sposoby. W ujęciu typowym dla fizyki przez absolutność rozumiemy niezmienniczość względem wyboru układu odniesienia. Czas absolutny jest taki sam w każdym układzie odniesienia, czas relatywny natomiast zmienia się w zależności od tego, w jakim układzie jest mierzony. Podobnie rzecz się ma z przestrzenią. Natomiast drugie rozumienie sporu absolutyzmu z relatywizmem (zwanym również relacjonizmem) ma charakter bardziej filozoficzny. Jest to pytanie o ontologiczną naturę czasu i przestrzeni – czy są one odrębnymi bytami, niezależnymi od innych substancji, czy też mają charakter wtórny, zależny np. od istnienia i własności przedmiotów materialnych. Absolutny charakter czasu i przestrzeni w pierwszym sensie nie gwarantuje jeszcze absolutności w sensie drugim. Może być tak, że w każdym układzie i czas, i przestrzeń wyglądają jednakowo, natomiast ontologicznie są zależne od jakichś innych bytów. Natomiast uzależnienie czasu i przestrzeni od wyboru układu odniesienia, jak się wydaje, pozbawia je charakteru ontologicznie odrębnych bytów.

Stosunkowo łatwo odpowiedzieć na pytanie, czy według mechaniki klasycznej czas zależy od układu odniesienia. Ze względu na przyjętą transformację Galileusza $t' = t$, czas jest mierzony jednakowo we wszystkich układach, a zatem jest absolutny w sensie fizycznym. Natomiast status przestrzeni nie jest oczywisty. Transformacja Galileusza zmienia wartość współrzędnych przestrzennych, ale czy znaczy to, że sama przestrzeń również zmienia się od układu do układu? Zależy to od tego, co dokładnie rozumiemy przez przestrzeń. W kontekście rozważań fizycznych w naturalny sposób możemy utożsamić przestrzeń z funkcją odległości zdefiniowaną na zbiorze punktów. Przestrzeń nazwiemy niezmienniczą, jeśli odległość przestrzenna punktów czy zdarzeń jest taka sama w każdym układzie odniesienia.

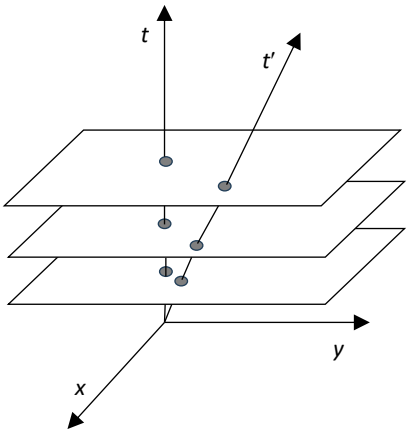
Zbadajmy tę sprawę dokładniej. Niech będą dane dwa zdarzenia A_1 i A_2 o współrzędnych (x_1, t_1) i (x_2, t_2). (Dla uproszczenia pomijamy dwie pozostałe współrzędne przestrzenne *y* i *z* – zawsze można wybrać układ współrzędnych, w którym współrzędne *y* i *z* obu zdarzeń wy-

noszą 0.) Odległość przestrzenna między A_1 i A_2 to oczywiście $|x_1 - x_2|$. Natomiast po transformacji Galileusza odległość ta generalnie ulegnie zmianie:

$$|x'_1 - x'_2| = |x_1 - x_2 - vt_1 + vt_2|.$$

Odległość przestrzenna między dwoma zdarzeniami nie jest w ogólnym wypadku inwariantna względem zmiany układu odniesienia. Można to zilustrować prostym przykładem – wyobraźmy sobie pasażera siedzącego w pociągu i bębniącego palcami po stoliku. Dwa kolejne uderzenia palcami w układzie związanym z pociągiem mają zerową odległość przestrzenną, natomiast z punktu widzenia obserwatora stojącego na ziemi ich odległość jest równa drodze, jaką w czasie między uderzeniami palców przebył pociąg.

Zauważmy jednak, że odległość w układzie u' będzie równa odległości w u, jeśli tylko czas zajścia zdarzeń A_1 i A_2 jest ten sam: $t_1 = t_2$. Pokazuje to, że odległość między zdarzeniami równoczesnymi jest inwariantem transformacji Galileusza (w istocie każdej transformacji współrzędnych łączącej dwa dowolnie poruszające się układy). Zatem przestrzeń może być uznana za niezmienniczą względem wyboru układu odniesienia, jeśli jest ona określona przez wszystkie odległości przestrzenne wzięte w tym samym momencie. Zbiór wszystkich zdarzeń (punktów) zachodzących w tym samym momencie i powiązanych funkcją odległości nazywamy przestrzenią momentalną (migawkową). Z naszej analizy wnioskujemy, że przestrzeń momentalna jest inwariantem transformacji Galileusza, a zatem jest ona absolutna w sensie fizycznym. Natomiast przestrzeń „globalna", obejmująca obiekty trwające w czasie, absolutna nie jest, gdyż jej charakterystyka przy pomocy funkcji odległości zależy od wyboru układu odniesienia.



Rys. 2.8. Dwa układy odniesienia i płaszczyzny równoczesności (momentalne przestrzenie)

Rozważania powyższe można zilustrować przy pomocy diagramów czasoprzestrzennych. Diagramy takie uwzględniają zarówno osie przestrzenne, jak i jedną oś czasową, na której odznaczamy współrzędną czasową danego obiektu czy zdarzenia. Z oczywistych względów nie możemy przedstawić graficznie ogólnego przypadku trzech osi przestrzennych i jednej czasowej, zazwyczaj więc redukujemy liczbę osi przestrzennych do dwóch, a nawet niekiedy do jednej osi x. Na diagramie (rys. 2.8) przedstawiona została linia ukośna, która reprezentuje tor ciała poruszającego się ze stałą prędkością względem danego układu. Linia

ta może również przedstawiać oś czasową t' nowego układu odniesienia, będącego w ruchu w stosunku do układu pierwotnego. Oś czasowa bowiem łączy ze sobą zdarzenia, które w danym układzie mają współrzędne przestrzenne równe zeru (zachodzą w początku układu), a początek układu u' porusza się względem układu u. Poziome płaszczyzny reprezentują przestrzenie chwilowe (momentalne), które wyglądają tak samo w każdym układzie odniesienia. Natomiast brak przestrzeni globalnej ujawnia się w fakcie, że nie można w absolutny sposób połączyć ze sobą punktów należących do różnych płaszczyzn relacją znajdowania się w tym samym miejscu. Dwa nierównoczesne zdarzenia, które zachodzą w tym samym miejscu względem jednego układu odniesienia, nie zachodzą w tym samym miejscu według innego układu. Innymi słowy, pojęcie tego samego miejsca jest zrelatywizowane do układu odniesienia. Powiązane ze sobą pojęcia ruchu, spoczynku i jednakowego umiejscowienia okazują się jedynie relatywne, a nie absolutne.

Przejdźmy teraz do zagadnienia filozoficznego. Czym naprawdę są czas czy przestrzeń? Czy powinniśmy sobie wyobrażać przestrzeń jako rodzaj pustego pojemnika wypełnionego miejscami, gotowymi do przyjęcia odpowiednich przedmiotów fizycznych? Czy czas jest rozciągłością składającą się z obiektów – momentów – które następują jeden po drugim, nawet wtedy, gdy nic się w nich nie dzieje? W filozofii Arystotelesa pojawia się pojęcie substancji, rozumiane nieco inaczej niż na gruncie nauk przyrodniczych. Substancja w ujęciu arystotelesowskim to byt zdolny do samodzielnego istnienia, czyli taki, który nie jest zależny w swoim istnieniu od innych bytów. Na przykład konkretne przedmioty fizyczne – kamienie czy drzewa – są substancjami, natomiast ich własności już nie są. Możemy zatem zadać pytanie, czy czas i przestrzeń są substancjami w wyżej określonym sensie. Czy przestrzeń istniałaby, gdyby nie było żadnych wypełniających ją przedmiotów? Czy czas może istnieć bez zachodzących w nim zdarzeń?

Twierdzącej odpowiedzi na powyższe pytania udziela stanowisko absolutyzmu (zwane również substancjalizmem). Zwolennikiem filozoficznego absolutyzmu w odniesieniu do czasu i przestrzeni był sam Newton. Powszechnie znane i cytowane fragmenty z jego *Principiów* nie pozostawiają co do tego żadnych wątpliwości:

> Czas absolutny, prawdziwy i matematyczny, sam z siebie i przez swą naturę upływa równomiernie bez związku z czymkolwiek zewnętrznym i inaczej nazywa się trwaniem.
>
> Przestrzeń absolutna, przez swą naturę, bez związku z czymkolwiek zewnętrznym, pozostaje zawsze taka sama i nieruchoma.
>
> Miejsce jest częścią przestrzeni, którą zajmuje ciało; zależnie od przestrzeni jest absolutne albo względne.
>
> Ruch absolutny jest przemieszczeniem ciała z jednego miejsca absolutnego w drugie.

Zauważmy, że Newton łączy stanowisko absolutyzmu w sensie filozoficznym z jego interpretacją fizyczną, utrzymując, że istnieje absolutne pojęcie miejsca, ruchu i spoczynku niezależne od stanu ruchu obserwatora (Newton nie korzystał jeszcze z pojęcia układu odniesienia). Twórca mechaniki klasycznej przedstawił argumenty teologiczne za swoim stanowiskiem, jednakże najbardziej znany jego argument odwołuje się do eksperymentu fizycznego. Jest to słynny argument z wiadrem.

Wyobraźmy sobie wiadro napełnione wodą i zawieszone na sznurze. Sznur zostaje skręcony, a następnie puszczony swobodnie, w wyniku czego wiadro zacznie się obracać. W pierwszej fazie ruchu obraca się wiadro, natomiast woda pozostaje w spoczynku. Obser-

wując powierzchnię wody, nie zauważamy żadnego efektu – powierzchnia jest płaska. Następnie woda zaczyna przejmować ruch obrotowy wiadra, w wyniku czego jej powierzchnia staje się wklęsła. W ostatnim etapie wiadro zostaje gwałtownie zatrzymane, natomiast woda nadal wiruje, a jej powierzchnia pozostaje wklęsła do czasu zatrzymania obrotu. Porównując ze sobą pierwszy i trzeci etap eksperymentu możemy zauważyć, że w obu wypadkach ruch względny wody i wiadra jest taki sam – woda obraca się względem wiadra lub wiadro względem wody, co na jedno wychodzi (rys. 2.9). Jednakże ewidentnie istnieje obserwacyjna różnica między obiema sytuacjami – w jednym przypadku powierzchnia wody jest płaska, a w drugim wklęsła. Newton wyprowadził stąd wniosek, że tylko w jednej z dwóch sytuacji woda „naprawdę" się obraca, a w drugiej naprawdę (czyli absolutnie) spoczywa. Kryterium absolutnego obrotu jest oczywiście pojawienie się sił pozornych, w tym wypadku siły odśrodkowej, która zmienia kształt powierzchni wody.

Rys. 2.9. Dwa etapy doświadczenia z wiadrem i wodą. Obserwowane ruchy obrotowe dokonują się względem zaznaczonych gwiazd stałych

Względem czego zachodzi prawdziwy ruch obrotowy? Newton nie miał wątpliwości, że jedyną możliwą odpowiedzią na to pytanie jest: względem przestrzeni absolutnej. Uczestniczące w ruchu cząsteczki wody zmieniają swoje położenie w stosunku do absolutnych miejsc w przestrzeni, i ten absolutny, a nie relatywny ruch wody objawia się realnym efektem fizycznym. Jednakże Newton miał świadomość, że nie każdy rodzaj ruchu daje obserwowalne efekty. Ruch jednostajny, jak już zauważył Galileusz, nie wywołuje skutków w postaci deformacji poruszającego się ciała. Dlaczego przestrzeń absolutna objawia się tylko wtedy, kiedy ciało przyspiesza względem niej, a nie kiedy po prostu zmienia ono swoje położenie? To pytanie pozostaje bez odpowiedzi. W konsekwencji z zasadniczych powodów nie możemy empirycznie stwierdzić istnienia i tożsamości absolutnych położeń w przestrzeni. Twierdzimy, że przyspieszające ciało musi zmieniać swoje położenie w absolutnej przestrzeni, ale nie wiemy, z jakiego na jakie. Może być tak, że dana cząsteczka wody w wiadrze przemieszcza się absolutnie z punktu A do punktu B na obwodzie okręgu, ale równie dobrze na ten ruch może się nałożyć nieobserwowalny ruch jednostajny całego wszechświata z niewiarygodną prędkością w jednym kierunku, a zatem absolutne położenie rozważanej cząstki po upływie sekundy może być wiele tysięcy kilometrów stąd. Należy jedynie wykluczyć możliwość (jeśli przyjmiemy Newtonowski punkt widzenia), że cząstka wirującej próbki wody mogłaby jednak znajdować się cały czas w tym samym absolutnym miejscu. Występowanie sił pozornych nie daje nam jednoznacznego kryterium identyfikacji absolutnych

miejsc, natomiast założenie istnienia tych miejsc jest najlepszym wyjaśnieniem dla obserwowalnego zachowania ciał poddanych przyspieszeniu.

Założenie, że istnienie przestrzeni absolutnej jest jedynym (lub też, ostrożniej, najlepszym) możliwym wyjaśnieniem efektów inercjalnych zostało jednak podważone. Ernst Mach, austriacki fizyk i filozof końca dziewiętnastego wieku, zauważył, że Newton zupełnie pominął istnienie innych ciał poza wiadrem i wodą, tak jakby wiadro z wodą było jedynym obiektem we wszechświecie. Jednak tak nie jest: wiadro i woda wirują względem pozostałych części wszechświata. Etapy pierwszy i trzeci eksperymentu z wiadrem różnią się empirycznie: w pierwszym to wiadro, a nie woda wiruje względem Ziemi, Słońca i gwiazd stałych; w trzecim wiadro spoczywa, a woda zaczyna wirować w stosunku do otoczenia. Aby uzyskać pełną symetrię, konieczną do wyprowadzenia Newtonowskiego wniosku, należałoby porównać następujące dwie sytuacje: jedną, w której woda wiruje względem wiadra i całej reszty wszechświata, a drugą, w której woda spoczywa, a wiadro wraz z całym wszechświatem wiruje wokół niej. To jednak z oczywistych względów jest niemożliwe – nie umiemy „obrócić" całym wszechświatem.

Argument Macha przeciw Newtonowskiej koncepcji przestrzeni absolutnej może być rozumiany dwojako. W rozumieniu słabszym Mach po prostu pokazuje lukę w rozumowaniu Newtona, która blokuje drogę do konkluzji. Jednakże wielu komentatorów interpretuje Macha jako stwierdzającego autorytatywnie, że ruch obrotowy całego wszechświata względem wody wywołałby dokładnie taki sam efekt jak obrót samej wody – pojawiłaby się siła odśrodkowa, zmieniająca kształt powierzchni wody. Sam Mach był dość ostrożny w formułowaniu hipotez „co by było, gdyby", natomiast jego nazwisko łączy się z konkretną doktryną. Doktryna ta, znana jako machizm, głosi, że efekty inercjalne należy tłumaczyć fizycznym wpływem całego wszechświata na dany przedmiot. W szczególności masa inercjalna jest rezultatem wpływu reszty wszechświata na każde ciało. Samotne ciało w pustym wszechświecie nie posiadałoby żadnych własności inercyjnych. Jest to z oczywistych względów spekulacja, której nie możemy poddać bezpośredniej weryfikacji, natomiast pewne konsekwencje machizmu można próbować testować. Na przykład można badać wpływ obrotu supercięż-kiego wydrążonego walca na masę obiektów znajdujących się w jego środku. Eksperymenty tego typu były przeprowadzane, choć ich wyniki nie są konkluzywne. W każdym razie idee Macha miały wielki wpływ na młodego Alberta Einsteina i jego pierwsze próby stworzenia ogólnej teorii względności, w której nie byłoby miejsca na absolutną przestrzeń, a wszystkie ruchy bez wyjątku byłyby względne.

Argument Newtona – jeśli uznamy go za przekonujący – rehabilituje pojęcie globalnej przestrzeni, odrzucone przez nas powyżej na korzyść przestrzeni migawkowej. Z tego względu teorię czasu i przestrzeni przyjętą przez Newtona nazywa się neo-Arystotelesowską, jako że przywraca ona Arystotelesowe pojęcie absolutnego ruchu i spoczynku, choć nadal dynamika Newtonowska pozostaje nie-Arystotelesowska w tym sensie, że siły wywołują nie sam ruch, a zmianę ruchu. Z kolei teorię czasu i przestrzeni bez absolutnej przestrzeni globalnej, a jedynie z przestrzenią momentalną, nazywa się Galileuszowską. Zatem fizyka Newtonowska może być zasadniczo połączona z dwiema koncepcjami czasu i przestrzeni.

Zagorzałym przeciwnikiem absolutystycznej koncepcji czasu i przestrzeni był Leibniz. Uważał on za absurdalne, że przestrzeń mogłaby istnieć bez wypełniających ją przedmiotów. Relacje geometryczne, takie jak odległość, zachodzą między przedmiotami, a dopiero wtórnie między punktami przestrzennymi. Punkty i miejsca przestrzenne i czasowe zawdzięczają swoją tożsamość przedmiotom i zdarzeniom w nich zlokalizowanym. Koncepcję Leibniza

nazywa się relacjonizmem, jako że kładzie ona nacisk na relacyjny charakter czasowych i przestrzennych własności obiektów. Jest sens mówić, że ciało A w chwili *t* jest odległe od ciała B o 100 metrów, ale nie można zakładać, że ciało A jest absolutnie zlokalizowane w pewnym miejscu, odległym o 100 metrów od lokalizacji ciała B. Leibniz przedstawił słynny argument przeciwko absolutystycznej koncepcji przestrzeni, znany jako argument z przesunięcia. Załóżmy, jak chcą tego absolutyści, że przestrzeń istnieje samodzielnie i niezależnie od zajmujących ją przedmiotów fizycznych. W takiej sytuacji można rozważyć przesunięcie wszystkich materialnych ciał we wszechświecie w pewnym kierunku o stałą odległość, np. pięciu metrów, bez zmiany ich relatywnych położeń. Sytuacje przed i po przesunięciu byłyby odmienne, ponieważ przedmioty zajmowałyby w nich numerycznie inne miejsca. Natomiast różnica ta w żaden sposób nie ujawniałaby się empirycznie czy też jakościowo.

Dwa stany rzeczy, różniące się jedynie położeniem całego świata względem przestrzeni absolutnej, stanowiłyby przykład czystej różnicy numerycznej bez żadnej różnicy w jakościowych własnościach. Leibniz był zwolennikiem (i *de facto* autorem) metafizycznej zasady tożsamości przedmiotów nieodróżnialnych. Głosi ona, że całkowita nieodróżnialność jakościowa (za pomocą własności) implikuje tożsamość (bycie tym samym przedmiotem) – lub też, w logicznie równoważnej formie, że każde dwa numerycznie różne przedmioty czy stany rzeczy muszą się różnić przynajmniej jedną własnością. Leibniz uzasadniał zasadę tożsamości nieodróżnialnych przez odwołanie do jeszcze bardziej abstrakcyjnej metafizycznej reguły, którą określał mianem zasady racji dostatecznej. Ujmując tę zasadę w kategoriach teologicznych, możemy powiedzieć, że Bóg, podejmując jakiekolwiek działanie, ma zawsze racjonalny powód do wyboru jednej spośród wielu możliwych opcji. Według absolutystów Bóg najpierw stworzył przestrzeń i czas, a następnie umieścił w nich świat fizyczny. Problem polega na tym, że Stwórca nie mógł mieć żadnego powodu, aby wybrać jedną z nieskończenie wielu możliwych lokalizacji świata fizycznego, a zatem zasada racji dostatecznej byłaby złamana. Rozwiązaniem tego teologiczno-metafizycznego problemu było według Leibniza przyjęcie, że dopiero stworzenie świata wraz z czasoprzestrzennymi relacjami pomiędzy jego elementami pozwoliło na zaistnienie czasu i przestrzeni jako wtórnych bytów.

Zasada tożsamości przedmiotów nieodróżnialnych może być sformułowana w postaci następującej implikacji: jeżeli przedmiot *a* i przedmiot *b* posiadają dokładnie te same własności, to *a* jest tożsamy z *b* (*a* jest tym samym przedmiotem co *b*). Jednakże treść tej zasady, a także jej spełnienie zależą od przyjętej interpretacji terminu „własność". W najbardziej liberalnym ujęciu każdemu prawdziwemu zdaniu o pewnym przedmiocie odpowiada pewna własność. Przy takiej interpretacji własnością danego przedmiotu będzie np. bycie dokładnie tym przedmiotem (taką cechę określa się często scholastycznym terminem *haecceitas*). Wtedy oczywiście Leibnizjańska zasada będzie trywialnie spełniona, jako że każde dwa przedmioty różnią się swoimi *haecceitas*. Jednakże większość filozofów uważa, że w kontekście zasady tożsamości przedmiotów nieodróżnialnych powinno się używać jedynie tzw. własności czystych (jakościowych), które nie zawierają odniesienia do konkretnych przedmiotów.

Innym ważnym rozróżnieniem pojęciowym jest podział na własności wewnętrzne i zewnętrzne (relacyjne). Własność wewnętrzna danego przedmiotu to taka, której posiadanie nie zależy od obecności innych przedmiotów, natomiast cecha zewnętrzna uwzględnia także otoczenie obiektu. Na przykład cecha bycia odległym o dwa metry od

brzozy jest zewnętrzna, gdyż ścinając brzozę pozbawiamy tej cechy położony koło niej przedmiot. W wersji silniejszej zasada tożsamości przedmiotów nieodróżnialnych głosi, że przedmioty nietożsame numerycznie muszą różnić się przynajmniej jedną cechą wewnętrzną, natomiast wersja słabsza dopuszcza ich nieodróżnialność wewnętrzną, jeśli tylko przysługują im odmienne cechy zewnętrzne. Zgodnie ze słabszą zasadą, dwa przedmioty mogą być odróżnione za pomocą relacji łączących je z innymi obiektami (na przykład gdy jeden z nich znajduje się bliżej pewnej brzozy niż drugi).

Dla naturalistycznie zorientowanych filozofów zasada tożsamości przedmiotów nieodróżnialnych może mieć bardziej empiryczne uzasadnienie. Zakładając, że jedynie jakościowe fakty oparte na posiadaniu pewnych własności mogą być dla nas empirycznie dostępne, musimy stwierdzić, że różnice czysto ilościowe, oparte na tożsamości numerycznej, nie mogą być poddane weryfikacji empirycznej. W wypadku lokalizacji całego świata względem przestrzeni absolutnej jest jasne, że nie możemy rozróżnić między dostępnymi opcjami. Fakty dotyczące absolutnej lokalizacji przestrzennej świata (a także lokalizacji w absolutnym czasie) nie mają żadnych empirycznych konsekwencji, a zatem mogą być przez nas potraktowane jako zbędny bagaż. Leibniz, filozof racjonalista, nie byłby zadowolony z takiego argumentu, jednakże zasada tożsamości przedmiotów nieodróżnialnych jest zwykle uzasadniana przez współczesnego filozofa nauki wykształconego w duchu empiryzmu logicznego właśnie w taki sposób.

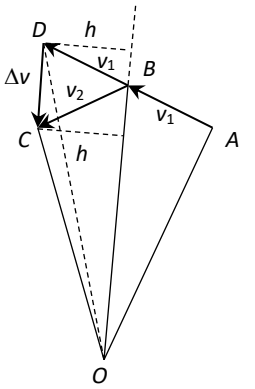
Argument Leibniza z przesunięcia uważa się powszechnie za niezwykle poważne wyzwanie dla absolutyzmu w kwestii czasu i przestrzeni, które równoważy a nawet przeważa argument Newtona z występowania sił pozornych. Ostatecznie jednak spór między absolutyzmem a relacjonizmem nie znajduje jednoznacznego rozwiązania na gruncie mechaniki klasycznej. W dalszych częściach książki zobaczymy, jak spór ten rozwinął się na gruncie nowszych teorii czasu i przestrzeni – szczególnej i ogólnej teorii względności. Pojawią się tam zupełnie nowe ujęcia i argumenty, jak na przykład współczesna wersja argumentu Leibniza z przesunięcia, znana pod nazwą argumentu dziury. Mimo znacznego zaawansowania technicznego nowych argumentów debata między obydwoma stanowiskami nadal się toczy, a jej rozstrzygnięcie jest równie odległe, jak za czasów Newtona i Leibniza.

2.6. Teoria grawitacji Newtona

Mechanika klasyczna w wersji Newtonowskiej dostarcza nam niezwykle użytecznego narzędzia do opisu i przewidywania mechanicznego zachowania ciał, w tym ruchu planet. Jednakże aby wykorzystać to narzędzie, potrzebny jest jeszcze jeden kluczowy element. Musimy w matematycznie ścisły sposób opisać siły, jakie działają na poruszające się ciała niebieskie. Ze szkolnego kursu fizyki wiemy, że są to siły przyciągania grawitacyjnego między ciałami obdarzonymi masą, a w szczególności siła przyciągania między Słońcem a pozostałymi ciałami niebieskimi. Siły te opisuje dobrze znane prawo powszechnego ciążenia. Mało kto jednak wie, w jaki sposób Newton doszedł do wniosku, że siła grawitacji między dwoma ciałami jest odwrotnie proporcjonalna do kwadratu ich odległości. Anegdota o jabłku, które spadając na głowę Newtona, miało zainspirować go do sformułowania prawa ciążenia, nie ma potwierdzenia historycznego, a do tego zaciemnia prawdziwe źródło jego odkrycia. Newton nie byłby w stanie wyprowadzić swojej formuły na podstawie badania spadku swobod-

nego ciał w polu grawitacyjnym Ziemi, gdyż w pobliżu powierzchni ziemskiej ciała spadają z jednakowym przyspieszeniem, co zauważył już Galileusz. Zależność przyspieszenia i siły grawitacji od odległości ujawnia się dopiero w skali planetarnej. Faktycznym źródłem odkrytej formuły były prawa Keplera, które syntetycznie ujmowały obserwowane prawidłowości ruchu planetarnego.

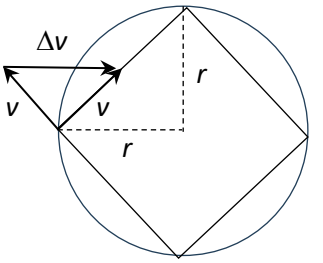
Newton postawił sobie zadanie znalezienia odpowiedniej siły, która wytłumaczyłaby to, że planety krążą po eliptycznych orbitach wokół Słońca, a także uzasadniła znane ilościowe fakty dotyczące tego ruchu. Jego rozumowanie można podzielić na kilka etapów. W pierwszym etapie pokazał, że z drugiego prawa Keplera można wyprowadzić wniosek, że siła działająca na każdą planetę jest skierowana centralnie w stronę Słońca. W drugim kroku Newton wyprowadził wzór na przyspieszenie dośrodkowe w ruchu kołowym. Końcowym etapem rozumowania Newtona było zastosowanie trzeciego prawa Keplera do obliczenia wartości siły centralnej, która dawałaby odpowiednie przyspieszenie dośrodkowe planety. Wszystkie trzy kroki oparte były na niezwykle pomysłowych argumentach geometrycznych, których walory powinniśmy docenić nawet kilkaset lat po ich sformułowaniu. Przyjrzyjmy się zatem nieco dokładniej temu rozumowaniu.



Rys. 2.10. Argument Newtona za centralnością siły grawitacji. Planeta porusza się od punktu A przez B do C (wektory prędkości są równe przesunięciom, gdyż interwały czasowe są jednostkowe). Trójkąty OAB i OBC mają równe pola z drugiego prawa Keplera. Trójkąty OAB i OBD mają również równe pola (te same podstawy AB i BD i wysokość padająca na te podstawy. Zatem trójkąty OBC i OBD muszą mieć te same pola. Skoro mają wspólną podstawę OB, wysokości padające na tę podstawę muszą być równe. Oznacza to, że wektor zmiany prędkości Δv (odcinek DC) jest równoległy do linii OB, a zatem siła działająca w momencie, gdy planeta jest w punkcie B, jest skierowana centralnie do O

Newton z powodzeniem stosował przybliżenia „gładkich" figur geometrycznych, takich jak okręgi czy elipsy, przy pomocy wielokątów. W pierwszym swoim argumencie założył, że ruch planety naokoło Słońca dokonuje się skokowo, podobnie jak w przypadku numerycznej analizy równań ruchu przedstawionej w paragrafie drugim tego rozdziału. Załóżmy więc, że w pierwszej sekundzie planeta przesuwa się o pewien odcinek, a w następnej sekundzie o inny odcinek, skierowany nieco bardziej w stronę Słońca. Z drugiej zasady dynamiki wiemy, że zmiana prędkości, która jest wektorem powstałym z odjęcia prędkości w pierwszej sekundzie od prędkości w sekundzie drugiej, odbywa się w kierunku przyłożonej siły. Rozpatrując dwa trójkąty reprezentujące pola zakreślone przez promień wodzący planety w równych, jednosekundowych odstępach, z drugiego prawa Keplera wnioskujemy, że pola tych trójkątów

muszą być równe. Ponadto pierwszy trójkąt ma takie samo pole co trójkąt powstały w wyniku przesunięcia pierwszego wektora do punktu zaczepienia drugiego wektora prędkości (por. rys. 2.10 ze szczegółowymi objaśnieniami). Stąd wynika już, że różnica dwóch kolejnych wektorów prędkości będzie równoległa do linii łączącej planetę ze Słońcem, a zatem siła zmieniająca prędkość planety musi być skierowana w stronę Słońca. Istotne jest, że argument Newtona może być powtórzony dla coraz mniejszych interwałów czasowych (jedna milisekunda, jedna nanosekunda itd.), a zatem w granicy wniosek pozostaje słuszny dla gładkiej trajektorii planety (elipsy czy okręgu).



Rys. 2.11. Obliczenie przyspieszenia dośrodkowego w ruchu po okręgu, przybliżonym za pomocą kwadratu. Zwróćcie uwagę, że w każdym z czterech wierzchołków zmiana prędkości Δv (a więc także przyspieszenie) jest skierowana w stronę środka okręgu

Podobną metodę przybliżeń zastosował Newton do wyprowadzenia wzoru na przyspieszenie dośrodkowe w ruchu po okręgu. Pierwszą aproksymacją takiego ruchu jest ruch po kwadracie wpisanym w okrąg – jest to bardzo niedokładne przybliżenie, ale jak się okazuje, może być łatwo rozszerzone na dowolny wielokąt foremny. Dla okręgu o promieniu r bok kwadratu wpisanego ma długość $\sqrt{2}r$ (rys. 2.11). Jeśli przez v oznaczymy wartość prędkości ciała (długość wektora prędkości), to w momencie przejścia przez wierzchołek kwadratu zmiana prędkości wynosi $\sqrt{2}v$. W ciągu jednego okresu obiegu T zmiana wektora prędkości odbywa się czterokrotnie, zatem średnio zmiana prędkości na jednostkę czasu (czyli inaczej średnie przyspieszenie) wynosi $\frac{4\sqrt{2}v}{T}$. Okres obiegu T powiązany jest z prędkością v prostą formułą (długość obwodu podzielona przez v): $T = \frac{4\sqrt{2}r}{v}$, z czego łatwo wyprowadzamy wzór na średnie przyspieszenie w ruchu po kwadracie:

$$a = \frac{v^2}{r}.$$

Rozumowanie powyższe można bez problemu powtórzyć dla pięciokąta, sześciokąta i ogólnie n-kąta foremnego, co pokazuje, że w granicy $n \to \infty$ przyśpieszenie dośrodkowe w ruchu po okręgu jest dane powyższym wzorem. Pozostaje jedynie obliczyć wartość siły, jaka musi działać na daną planetę, aby nadać jej przyspieszenie potrzebne do utrzymania jej w ruchu po okręgu. Wyraźmy tę siłę za pomocą okresu obiegu planety T zamiast prędkości, wstawiając w powyższej formule w miejsce v wyrażenie $\frac{2\pi r}{T}$:

$$F = ma = m\frac{4\pi^2 r}{T^2}.$$

Z trzeciego prawa Keplera wiemy, że stosunek sześcianu promienia do kwadratu okresu obiegu jest taki sam dla wszystkich planet: $\frac{r^3}{T^2} = C$, gdzie C jest pewną stałą. Mnożąc licznik i mianownik powyższej formuły przez r^2 i stosując prawo Keplera, otrzymujemy:

$$F = \frac{4m\pi^2 C}{r^2}.$$

Zatem siła utrzymująca planety w ruchu wokół Słońca musi być odwrotnie proporcjonalna do kwadratu odległości danej planety od Słońca. Newton poddał wyprowadzoną przez siebie formułę testowi przez zastosowanie jej do ruchu Księżyca wokół Ziemi. Księżyc nie podlega prawom Keplera, jako że nie obiega on Słońca, a zatem powyższe rozumowanie nie ma tu bezpośredniego zastosowania. Znając odległość Księżyca od Ziemi oraz jego okres obiegu (około 28 dni), Newton policzył jego przyspieszenie dośrodkowe a_K, a następnie porównał je ze znanym przyspieszeniem przy powierzchni Ziemi a_Z. Stosunek tych dwóch przyspieszeń okazał się równy stosunkowi kwadratów promienia Ziemi i odległości od Ziemi do Księżyca, zgodnie z prawem odwrotnej proporcjonalności do kwadratu odległości. Zatem siła grawitacji pochodząca od Ziemi wystarcza, aby utrzymać Księżyc w ruchu obiegowym wokół naszej planety.

Uogólniając powyższe rozważania, Newton doszedł do wniosku, że każde dwa ciała przyciągają się siłą odwrotnie proporcjonalną do kwadratu ich odległości. Stosując zasadę akcji i reakcji (trzecie prawo dynamiki), można argumentować, że siła ta musi być również proporcjonalna zarówno do masy jednego z ciał, jak i masy drugiego z nich. W rezultacie otrzymujemy dobrze znane prawo powszechnego ciążenia:

$$F = G\frac{m_1 m_2}{r^2},$$

gdzie G jest stałą grawitacji, której dokładną wartość wyznaczył dużo później Henry Cavendish w doświadczeniach z tzw. wagą skręceń.[9]

Zwróćmy uwagę na ważny aspekt wyprowadzenia prawa powszechnego ciążenia z praw ruchu planetarnego Keplera, który nie zawsze jest odpowiednio podkreślany. Trzecie prawo Keplera implikuje, że planeta oddalona o daną odległość r od Słońca będzie poddana takiemu samemu przyspieszeniu dośrodkowemu, niezależnie od jej masy, wielkości itp. Wynika to stąd, że ustalenie promienia r jednoznacznie określa okres obiegu T, a zatem także przyspieszenie dośrodkowe. W konsekwencji we wzorze na siłę grawitacji pojawia się masa inercyjna z drugiego prawa dynamiki, a zatem masa zaczyna pełnić podwójną rolę: jako wielkość charakteryzująca bezwładność ciała poddanego działaniu sił i jako miara oddziaływania grawitacyjnego. Tak jednak być nie musi. Na przykład w rozdziale 4. omówimy znane prawo elektrostatyki Coulomba, zgodnie z którym siła oddziaływania jest proporcjonalna do nowej wielkości charakteryzującej oddziałujące cząstki – ich ładunku elektrycznego. Gdyby ciała niebieskie oddziaływały na siebie elektrostatycznie, okres obiegu planety o danym promieniu orbity byłby dodatkowo uzależniony od jej

[9] Cały czas zakładamy oczywiście, że mamy do czynienia z ciałami punktowymi. W przypadku ciał rozciągłych sprawa się komplikuje, chyba że mają one idealnie symetryczny kształt, np. kuli. Wtedy prawo Newtona nadal obowiązuje, jeśli odległość r liczymy od środków obu kul. W ogólnym wypadku, aby policzyć całkowitą siłę, należy zsumować przyczynki pochodzące od niewielkich, prawie punktowych fragmentów ciała, stosując matematyczną metodę całkowania.

ładunku (a dokładniej, jego stosunku do masy ciała). Ta specyfika oddziaływania grawitacyjnego, której wcześniejszy przejaw poznaliśmy już przy okazji prawa spadku swobodnego Galileusza, znajdzie wyjaśnienie dopiero na gruncie ogólnej teorii względności.

Sukces Newtona spotkał się z mieszanymi reakcjami współczesnych. Z jednej strony matematyczne opisanie sił potrzebnych do utrzymania planet na swoich orbitach stanowiło ogromny przełom w nauce, z drugiej zaś natura sił grawitacyjnych pozostała dla wielu zagadkowa. W jaki sposób Słońce wpływa na ruch odległych o miliony kilometrów planet? Czy wysyła ono niewidzialne nici wiążące planety i pociągające je ku sobie? Pamiętajmy, że za czasów Newtona obowiązującym paradygmatem siły mechanicznej było oddziaływanie przez kontakt. Aby zmusić spoczywające ciało do ruchu, należy fizycznie je „popchnąć". Popularną koncepcją ruchu planet, propagowaną m.in. przez zwolenników Kartezjusza, była tzw. teoria wirów. Zakładała ona, że przestrzeń międzyplanetarna wypełniona jest niewidzialną substancją, znajdującą się w stanie ruchu wirowego wokół Słońca. Według kartezjan to właśnie ta substancja pociąga planety za sobą jak kawałki korka na powierzchni wiru wodnego. Dla kartezjan nie do pojęcia było, aby w pustej przestrzeni (której istnienia zresztą nie uznawali, idąc za swoim mistrzem) mogły pojawić się siły działające na przedmioty. Coś musi te siły przenosić – jakaś fizyczna, materialna substancja.

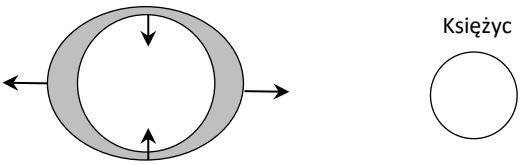
Indagowany o naturę sił grawitacyjnych, Newton odpowiedział swoim słynnym i jakże charakterystycznie aroganckim stwierdzeniem „Nie wymyślam hipotez" (*Hypotheses non fingo*). Chciał przez to zapewne powiedzieć, że podstawowym celem naukowca powinno być ujęcie faktów i danych empirycznych w matematyczne formuły, a metafizyczną interpretację tych formuł należy zostawić filozofom. Nie musimy tworzyć modelu oddziaływań grawitacyjnych opartego na znanych interakcjach mechanicznych – wystarczy, że umiemy je ściśle opisać. Takie instrumentalistyczne podejście do nauki, które poznaliśmy już przy okazji analizy przewrotu Kopernikańskiego, rozbrzmiewa w późniejszych wypowiedziach fizyków, na przykład w słynnym zawołaniu „*Shut up and calculate*" („Zamknij się i licz"), stosowanym przy okazji pytań o interpretację formalizmu mechaniki kwantowej. Choć postawa instrumentalistyczna zakłada odrzucenie wielu filozoficznych pytań jako nienaukowych, to jednak warto zauważyć, że jej przyjęcie jest samo w sobie także decyzją o charakterze filozoficznym, niewynikającą z samych danych empirycznych.

Ściśle związaną z problemem ukrytej natury sił grawitacyjnych jest kwestia tzw. działania na odległość. Oddziaływanie grawitacyjne, jak się wydaje, jest w stanie pokonać ogromne dystanse bez żadnego opóźnienia. Newton przyjął, że siła grawitacji pochodząca od odległego ciała (np. Słońca) jest odczuwalna w każdym punkcie przestrzeni natychmiastowo. Łamie to intuicyjną zasadę lokalności, o której będziemy jeszcze mówić w rozdziałach poświęconych teorii elektromagnetyzmu oraz mechanice kwantowej. Uczeni, którzy domagali się od Newtona podania mechanizmu działania sił grawitacyjnych, mieli na uwadze właśnie problem nielokalności czy też działania na odległość. Intuicyjnie wydaje się jasne, że działanie Słońca na planety musi się jakoś rozchodzić w przestrzeni w postaci zaburzenia o skończonej prędkości. Jednakże Newton, jak wiemy, odrzucił wszelkie dywagacje na ten temat. Znów przyjdzie nam czekać aż do powstania ogólnej teorii względności, aby rozwiązać ten problem interpretacyjny.

Prawo grawitacji Newtona, połączone z druga zasadą dynamiki, pozwala nam na ścisłe analityczne obliczenie trajektorii pojedynczego ciała w polu grawitacyjnym innego ciała (zagadnienie dwóch ciał). W zależności od prędkości początkowej danego ciała, jego trajektoria

będzie jedną z krzywych stożkowych: okręgiem lub elipsą dla mniejszych prędkości, a przy osiągnięciu i przekroczeniu prędkości granicznej parabolą lub hiperbolą. Te dwie ostatnie trajektorie są otwarte – ciało poruszające się po tych trajektoriach nigdy nie wróci do punktu wyjścia. Jest to tor ruchu typowy dla ciał spoza Układu Słonecznego, przechwyconych przez Słońce i „wystrzelonych" z powrotem w przestrzeń. Natomiast zagadnienie trzech ciał sprawia już duże problemy techniczne. Rozwiązuje się je zwykle przy założeniu, że jedno z ciał jest dużo mniejsze lub dużo bardziej oddalone, a zatem ruch pozostałych ciał można opisać jako złożenie zagadnienia dwóch ciał i niewielkiego zaburzenia (perturbacji) pochodzącego od trzeciego ciała. Typowym efektem niewielkiej perturbacji jest powolna rotacja orbity jednego ciała wokół ciała drugiego (rotacja ta jest zwana precesją *perihelium*, czyli punktu na orbicie położonego najbliżej centralnego ciała (Słońca)). Jednak w ogólności zagadnienie trzech ciał nie jest rozwiązywalne analitycznie, co przypomina nam o poważnych ograniczeniach epistemologicznej wersji determinizmu, o których mówiliśmy wcześniej.

Spektakularnym sukcesem teorii grawitacji Newtona było poprawne wyjaśnienie zjawiska przypływów i odpływów, które sprawiało wiele kłopotów uczonym. Galileusz wysunął śmiałą hipotezę, że przypływy i odpływy morskie są wywołane łącznym działaniem siły odśrodkowej pochodzącej od ruchu obrotowego Ziemi i siły odśrodkowej w ruchu obiegowym dookoła Słońca. Punkt na powierzchni Ziemi, który znajduje się po przeciwległej stronie planety w stosunku do Słońca „odczuwa" maksymalną siłę odśrodkową, która jest sumą siły „obrotowej" i „obiegowej", a punkt skierowany w stronę Słońca minimalną (różnicę dwóch sił), co powinno dać efekt podnoszenia i obniżania powierzchni wody. Jednakże hipoteza ta daje błędne przewidywania częstotliwości przypływów: według Galileusza powinniśmy mieć jeden przypływ i jeden odpływ w ciągu dnia (i do tego muszą one być skorelowane z położeniem Słońca na niebie). W rzeczywistości obserwujemy dwa przypływy i dwa odpływy, przy czym przypływy pojawiają się, kiedy Księżyc jest bądź w pobliżu zenitu, bądź w „nadirze" (po przeciwległej stronie sfery niebieskiej).



Rys. 2.12. Wytłumaczenie powstania przypływów i odpływów w teorii grawitacji Newtona. Szary obszar reprezentuje masę wody na powierzchni Ziemi

Prawo grawitacji Newtona natomiast implikuje, że strona Ziemi zwrócona w stronę Księżyca będzie przyciągana mocniej niż strona opozycyjna. Powoduje to powstanie efektywnej siły rozciągającej naszą planetę i znajdujące się na niej oceany wzdłuż osi łączącej Ziemię z Księżycem, co daje dwa „wybrzuszenia" oceaniczne (przypływy) i dwa miejsca o obniżonym poziomie wód (odpływy) – por. rys. 2.12. Można policzyć, że efekty pływowe pochodzące od Księżyca są dużo mocniejsze od analogicznych sił pochodzących od Słońca – nie dlatego, że grawitacja Słońca odczuwalna na Ziemi jest mniejsza od Księżycowej (w istocie jest dużo większa), ale dlatego, że średnica Ziemi jest pomijalna w stosunku do odległości

od Słońca (ale nie w stosunku do odległości od Księżyca), a zatem różnica między przyspieszeniem części bliższej i części dalszej od Słońca jest znikoma.

2.7. Mechanika klasyczna po Newtonie

Przez następne dwieście lat po śmierci Newtona mechanika klasyczna rozwijała się w niewiarygodnym tempie. Postęp dokonywał się zarówno w kwestii rozszerzenia zakresu jej stosowalności, jak i rozwijania zupełnie nowych technik matematycznych. Udział w tym postępie miała cała plejada genialnych fizyków i matematyków: Jean-Baptiste d'Alembert, Leonhard Euler, Joseph Louis Lagrange, Pierre Simone Laplace, Wiliam Hamilton i wielu innych, których nazwiska zostały uwiecznione w matematycznej i fizycznej terminologii używanej do dziś. Pierwszym naturalnym rozszerzeniem zakresu stosowalności mechaniki było uwzględnienie rozmiarów ciał poddanych działaniu sił, co doprowadziło do powstania tzw. mechaniki brył sztywnych. Bryły sztywne (czyli takie, które nie zmieniają swojego kształtu) mogą uczestniczyć zarówno w ruchu postępowym, opisanym zwykłymi prawami dynamiki zastosowanymi do ich środka masy, jak i w ruchu obrotowym. Ze szkoły wiemy, że do opisu ruchu obrotowego potrzebne są nowe pojęcia, takie jak moment siły (siła razy ramię działania), prędkość i przyspieszenie kątowe, moment pędu czy moment bezwładności. Moment bezwładności, jako uogólnienie pojęcia masy, zawiera w sobie informację zarówno o całkowitej masie obiektu, jak i jego kształcie i rozłożeniu masy, co wpływa na inercję w ruchu obrotowym. Drugie prawo dynamiki Newtona zastosowane do ruchu obrotowego względem stałej osi będzie miało postać następującej równości: moment działającej siły = moment bezwładności razy przyspieszenie kątowe.[10] Dalszym uogólnieniem ruchu brył sztywnych będzie dynamika płynów nieściśliwych, uwzględniająca możliwość względnego ruchu fragmentów ciała (np. strumienia wody), jednakże bez zmiany jego objętości. Jest to niezwykle skomplikowana matematycznie część mechaniki, która ma bardzo ważne zastosowania praktyczne np. w aerodynamice.

Kluczowym dodatkiem do praw dynamiki są zasady zachowania. Już w średniowieczu wysunięto hipotezę, że ilość ruchu danego ciała (jego tzw. *impetus*) nie powinna ulegać zmianie. W mechanice Newtonowskiej idea ta przybrała postać zasady zachowania pędu, którą dla przypadku układu dwóch ciał niepoddanych działaniu zewnętrznych sił można łatwo wyprowadzić z trzeciego prawa dynamiki. Równość siły akcji i reakcji można zapisać w następującej formie (p_1 i p_2 oznaczają chwilowe pędy ciała pierwszego i drugiego):

$$\frac{dp_1}{dt} = -\frac{dp_2}{dt}$$

skąd szybko dostajemy, iż

$$\frac{d(p_1 + p_2)}{dt} = 0,$$

[10] Jeśli dopuścimy osie obrotu zmieniające się w czasie, jak np. oś wirującego bąka, który wykonuje również tzw. ruch precesyjny („kiwający"), to równania ruchu obrotowego przyjmą postać tzw. równań Eulera, zawierających momenty bezwładności, siły i prędkości kątowe względem trzech prostopadłych kierunków w przestrzeni.

czyli pęd całkowity odosobnionego układu nie ulega zmianie. Zachowany także jest moment pędu obracającego się układu izolowanego.

Kolejną ważną regułę stanowi zasada zachowania energii. Energia kinetyczna przedmiotu poruszającego się z prędkością v dana jest w postaci znanego wzoru $\frac{mv^2}{2}$. Wielkość ta w sposób oczywisty nie jest zachowana przy istnieniu działających sił, które wpływają na zmianę prędkości. Jednakże układy izolowane zachowują swoją energię kinetyczną (w wypadku pojedynczego ciała to dość banalna obserwacja, gdyż z pierwszego prawa Newtona od razu wynika, że jego prędkość musi być niezmienna w czasie). Z kolei w wypadku działających sił szczególnego rodzaju (tzw. sił zachowawczych) można znaleźć pewną zachowaną wielkość, której częścią będzie energia kinetyczna. Ta wielkość to całkowita energia układu, której drugim składnikiem jest energia potencjalna zależna od pola sił. Dobrze znanym przykładem jest energia potencjalna w jednorodnym polu grawitacyjnym (np. przy powierzchni Ziemi), zależna tylko od wysokości ciała względem pewnego umownego poziomu.

Bardzo interesującym faktem, udowodnionym na początku dwudziestego wieku przez matematyczkę Emmę Noether, jest związek praw zachowania w mechanice z symetriami. Okazuje się, że każdemu prawu zachowania pewnej wielkości odpowiada szczególny rodzaj symetrii równań ruchu. Przez symetrię rozumiemy przekształcenie, które nie zmienia odpowiednich formuł matematycznych. (Na przykład transformacja Galileusza jest symetrią równań Newtona.) Zasada zachowania energii wynika z symetrii względem przesunięć (translacji) w czasie, a zasada zachowania pędu jest konsekwencją symetrii względem przesunięć w przestrzeni. Z kolei zachowanie momentu pędu to przejaw symetrii obrotowych (izotropii). Więcej szczegółów na temat twierdzenia Noether znajdziecie w paragrafie z gwiazdką.

Najbardziej spektakularnym aspektem rozwoju mechaniki klasycznej po Newtonie było wprowadzenie zupełnie nowych sformułowań podstawowych praw mechaniki i matematycznych metod ich stosowania. Najogólniej rzecz biorąc, w mechanice post-newtonowskiej można wyróżnić dwa zasadnicze nurty, które nazywa się niekiedy mechaniką Lagrange'a i mechaniką Hamiltona. Oba te podejścia, choć przy pewnych założeniach równoważne z tradycyjną mechaniką newtonowską, dają się zastosować w sytuacjach odległych od klasycznych problemów mechanicznych. Impulsem do rozwoju tych nowych technik, a w szczególności mechaniki Lagrange'a, był problem opisu zachowania układów poddanych ograniczeniom ruchu, czyli tzw. więzom. Przykładem układu z więzami może być np. koralik ześlizgujący się bez tarcia po odpowiednio wykrzywionym drucie w polu grawitacyjnym. Innego rodzaju więzy ograniczają z kolei toczące się bez poślizgu koło. Więzy pierwszego rodzaju (zwane również holonomicznymi) dają się matematycznie przedstawić w postaci równania łączącego współrzędne ciała, podczas gdy równania określające więzy drugiego rodzaju (nieholonomiczne) zawierają również prędkości. W przypadku toczącego się koła równanie więzów łączy prędkość obrotową z prędkością posuwistą.

Jest niezmiernie trudno zastosować równania Newtona bezpośrednio do opisu ruchu poddanego więzom. Na przykład w wypadku ruchu koralika musielibyśmy uwzględnić w każdym momencie siły reakcji pochodzące od drutu, które zależą od jego kształtu. Podobne trudności napotyka próba opisania ruchu złożonego wahadła, zbudowanego z dwóch lub więcej części połączonych w sposób umożliwiający ich względny ruch. Problemy takie jak powyższe rozważali osiemnastowieczni uczeni – Jakub Bernoulli, Jean-Baptiste d'Alembert i Jo-

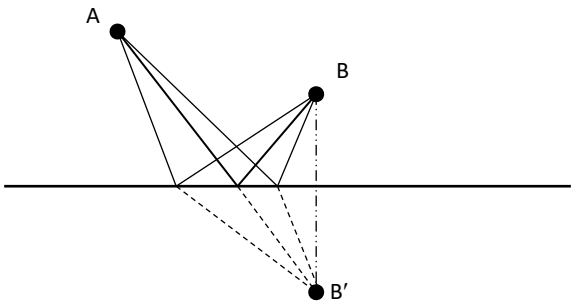
seph Louis Lagrange. Inspirację czerpali ze statyki, a konkretnie ze znanego prawa dźwigni, które głosi, że w stanie równowagi iloczyny sił działających na obiekt razy ich ramię działania powinny się sumować do zera. Doprowadziło to ostatecznie do sformułowania przez Lagrange'a tzw. zasady prac wirtualnych, znanej również pod nazwą zasady d'Alemberta. Rozważmy ogólnie układ ciał połączonych ze sobą pewnymi więzami ograniczającymi ich ruchy. Dla każdego ciała z osobna obliczmy działającą na nie siłę i odejmijmy od niej iloczyn masy i chwilowego przyspieszenia podczas niewielkiego (wirtualnego) ruchu. Do tak obliczonych sił następnie stosujemy zasadę równowagi statycznej ze względu na niewielkie przesunięcia w kierunku danej siły. Zasada ta wymaga, aby suma iloczynów wszystkich sił i ich wirtualnych przesunięć była równa zeru.

Lagrange następnie pokazał, że sformułowana wyżej zasada prac wirtualnych jest matematycznie równoważna pewnemu równaniu różniczkowemu, zwanemu równaniem Lagrange'a (inna terminologia to równanie Lagrange'a II rodzaju lub równanie Eulera-Lagrange'a). Równanie to zawiera pewną funkcję $\mathcal{L}$ określaną mianem funkcji Lagrange'a lub też w skrócie lagrangianem (czyt. lagranżjan). Lagrangian danego układu jest przedstawiany jako funkcja tzw. współrzędnych uogólnionych. Współrzędne uogólnione powstają ze „zwykłych" współrzędnych (np. kartezjańskich) przez zastosowanie do nich równań więzów. Ze względu na istnienie więzów współrzędne danego ciała nie są od siebie niezależne – np. jeśli znamy współrzędną „pionową" koralika ześlizgującego się po drucie, możemy obliczyć współrzędną poziomą. Natomiast współrzędne uogólnione są niezależne, a zatem ich liczba musi być mniejsza od liczby zwykłych współrzędnych. Do opisu ruchu koralika wystarczy jedna współrzędna uogólniona w postaci np. odległości od punktu początkowego, liczonej wzdłuż drucika.

Lagrangian ponadto zawiera odniesienie do prędkości wyrażonych we współrzędnych uogólnionych (czyli pochodnych współrzędnych uogólnionych po czasie). Równanie Lagrange'a zawiera tzw. pochodne cząstkowe lagrangianu po współrzędnych uogólnionych oraz ich prędkościach (dokładna forma równania podana jest w paragrafie „z gwiazdką"). W wypadku gdy siły działające na dany obiekt dadzą się przedstawić w szczególnej matematycznej postaci jako tzw. gradient potencjału, lagrangian układu może być przedstawiony jako różnica między energią kinetyczną a energią potencjalną układu. Zachowanie dynamiczne danego układu można więc opisać, znając zależność jego energii kinetycznej i potencjalnej od współrzędnych uogólnionych i prędkości uogólnionych. Zauważmy, że przy tym podejściu nie musimy w ogóle odwoływać się do pojęcia siły. Jest to dobra wiadomość dla tych, którzy tak jak np. d'Alembert, mają poważne wątpliwości co do ontologicznego charakteru sił (ich „tajemniczej" czy „ukrytej" natury). Z drugiej strony jednak można zauważyć, że status energii także nie jest całkowicie jasny, jako że energia układu nie jest dana nam w bezpośredniej obserwacji. W każdym razie przejście od klasycznej mechaniki newtonowskiej do mechaniki Lagrange'a wiąże się ze zmianą paradygmatu mechaniki – od koncepcji „popychającej" siły do całkowitej energii determinującej przyszłe zachowanie układu.

Jeszcze bardziej radykalne odejście od obrazu świata zawartego w klasycznej mechanice newtonowskiej można zauważyć w ujęciu mechaniki opartym na tzw. zasadzie najmniejszego działania. Jego źródła odnajdujemy w tak zwanych problemach „minimalizacyjnych", znanych od czasów starożytnych. Jednym z takich problemów jest pytanie o to, jaki kształt geometryczny umożliwia zamknięcie danej powierzchni najkrótszym obwodem (jest to oczywiście okrąg). Innym przykładem, bardziej zbliżonym do problemów fizycznych, jest następujący praktyczny problem: jaką drogę należy wybrać od punktu A (którym może być np.

dom) do pewnej prostej (rzeki), a następnie do punktu B (grządki, którą należy podlać wodą zaczerpniętą z rzeki), aby całkowita droga była jak najkrótsza (rys. 2.13). Nawiasem mówiąc, problem ten można łatwo rozwiązać, korzystając z odpowiednio zastosowanej symetrii (odbicia względem prostej). Rozwiązaniem jest droga, której części tworzą takie same kąty z wybraną prostą.

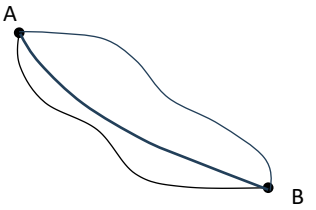


Rys. 2.13. Minimalizacja drogi między punktem A, „rzeką" i punktem B. Odbijając punkt B względem prostej, widzimy, że problem można przeformułować jako poszukiwanie najkrótszej drogi między A i B′, którą jest oczywiście linia prosta AB′

Ten ostatni przykład nasuwa skojarzenia ze znanym z optyki geometrycznej prawem odbicia. Prawo to może być wyprowadzone z tzw. zasady Fermata, która głosi, że światło porusza się w taki sposób, aby minimalizować czas dotarcia do danego punktu. W wypadku rozchodzenia się światła w jednorodnym ośrodku ze stałą prędkością, minimalizacja czasu jest oczywiście równoważna minimalizacji drogi, a zatem otrzymujemy analogiczne prawo odbicia jak w powyższym przykładzie z „rzeką". Natomiast ciekawszym przypadkiem jest zjawisko refrakcji (załamania) przy przejściu z jednego ośrodka do drugiego, kiedy w różnych ośrodkach prędkość jest różna. Okazuje się, że przy założeniu odwrotnej proporcjonalności prędkości światła w danym ośrodku do współczynnika załamania tego ośrodka, z zasady Fermata można wyprowadzić znane z optyki prawo załamania Snelliusa (stosunek sinusa kąta padania do sinusa kąta załamania jest stały dla danych dwóch ośrodków).

Ogólny sposób postępowania zastosowany w zasadzie Fermata jest następujący: wybieramy punkt początkowy A i punkt końcowy B, a następnie rozważamy wszystkie możliwe trajektorie od A do B i obliczamy dla każdej z nich pewną wielkość (np. czas dotarcia). Trajektoria, dla której wielkość ta przyjmuje wartość minimalną, jest trajektorią faktycznie wybraną w realnym zjawisku (rys. 2.14). Zastosowaniem tego schematu do ruchu przedmiotów materialnych zajęli się Pierre Louis Maupertuis i Leonhard Euler. Wielkość podlegającą minimalizacji nazwali „działaniem", a zatem podstawą proponowanego podejścia była zasada najmniejszego działania. Dla Maupertuisa działanie było określone przez iloczyn pędu oraz drogi pokonanej przez ciało. Dokładniej, aby obliczyć całkowite działanie dla danej trajektorii ciała, należy zsumować działania obliczone na niewielkich odcinkach drogi, na których zmiana pędu (prędkości) jest pomijalna. W granicy, gdy wybrane odcinki zmierzają do zera, odpowiednia suma przechodzi w matematyczną operację całkowania („odwrotność" różniczkowania). Zasada najmniejszego działania nakazuje wybrać spośród wszystkich możliwych trajektorii ciała o danej energii tę, dla którego opisana wyżej całka daje minimalną wartość.

Należy podkreślić, że w ujęciu Maupertuisa należy rozważać tylko te trajektorie, dla których całkowita energia ciała jest taka, jak w rzeczywistości.



Rys. 2.14. Ilustracja zagadnienia minimalizacji. Dla każdej możliwej trajektorii od A do B obliczamy działanie i wybieramy tę trajektorię, dla której to działanie jest minimalne (linia pogrubiona)

Genialny matematyk Euler sformalizował powyższą procedurę, nadając jej matematycznie precyzyjną postać tzw. rachunku wariacyjnego. Rachunek wariacyjny może być postrzegany jako uogólnienie newtonowskiego i leibnizjańskiego rachunku różniczkowego. W rachunku różniczkowym obliczamy pochodną danej funkcji zmiennej rzeczywistej $f(x)$ w celu ustalenia lokalnego minimum lub maksimum tej funkcji, czyli takiego argumentu x_0 funkcji, dla którego wartość funkcji jest lokalnie największa lub najmniejsza. Jak wiadomo, kryterium minimalizacji czy też maksymalizacji jest zerowanie się pochodnej w tym punkcie (należy jeszcze wykluczyć możliwość istnienia punktu przegięcia – w tym celu oblicza się drugą pochodną). Natomiast w wypadku zasady najmniejszego działania chcemy ogólnie znaleźć nie liczbę, a funkcję (trajektorię), dla której odpowiednia wartość działania jest minimalna. Odwzorowanie, przypisujące odpowiednie wartości funkcjom (a nie liczbom), nazywa się „funkcjonałem". Euler pokazał, że warunek minimalizacji (bądź też maksymalizacji) danego funkcjonału można wyrazić za pomocą tzw. wariacji funkcjonału, która w tym wypadku musi przyjąć wartość zerową.

Zasadę najmniejszego działania w najogólniejszej formie podał Lagrange.[11] Okazuje się, że działanie, którego minimalizacja (bądź też w niektórych wypadkach maksymalizacja) wybiera rzeczywistą trajektorię ze zbioru wszystkich możliwości, może przyjąć postać całki z wprowadzonej wcześniej funkcji Lagrange'a. Jest to całka, której parametrem całkowania jest czas, ale zawiera ona w sobie „ukrytą" informację o wybranej trajektorii, ponieważ dla każdej chwili wartość lagrangianu jest obliczana przy założeniu, że układ znajduje się na odpowiednio wybranej trajektorii. Przy tym ogólnym podejściu nie musimy już ograniczać się do trajektorii opisujących ruch obiektu o ustalonej energii – możemy brać pod uwagę wszystkie matematycznie dopuszczalne trajektorie. Lagrange udowodnił, że założenie o zerowaniu wariacji z tak zdefiniowanego działania (czyli warunek maksymalizacji lub minimalizacji funkcjonału) implikuje spełnienie równania Eulera-Lagrange'a, o którego wyprowadzeniu z zasady prac wirtualnych mówiliśmy wcześniej. Przekonujemy się zatem, że równanie Eulera-Lagrange'a można wyprowadzić na dwa różne sposoby.

[11] Ze względu na to, że w niektórych wypadkach właściwą trajektorię układu wybiera się przez maksymalizację działania, powinno się mówić o zasadzie ekstremalnego działania. Jednakże z powodów historycznych stosuje się nadal termin „zasada najmniejszego działania".

Zasada najmniejszego działania jest niezwykle interesująca z filozoficznego punktu widzenia. Porównajmy ją ze standardową metodą obliczania trajektorii za pomocą drugiej zasady dynamiki Newtona, omówioną w paragrafie 2.2. Aby obliczyć przyszłą trajektorię ciała lub układu ciał, należy wziąć pod uwagę obecny stan układu (położenia i prędkości) oraz działające obecnie siły. Zatem kierunek przewidywania jest zwrócony w stronę przyszłości, zgodnie z intuicyjnym przekonaniem, iż przyczyny poprzedzają czasowo swoje skutki. Inaczej natomiast sprawa wygląda w wypadku podejścia opartego na minimalizacji (lub maksymalizacji) działania. W celu obliczenia przyszłej trajektorii musimy się zwrócić w kierunku przyszłości, tj. wziąć pod uwagę całkowite działanie na każdej z możliwych dróg ewolucji układu i wybrać drogę z najmniejszą wartością. Jednakże całkowite działanie jest znane dopiero pod koniec zakończonego ruchu, a nie na początku. Wygląda zatem na to, że przynajmniej część przyczyny przyszłego ruchu ciała (czy też ciał) znajduje się w jeszcze dalszej przyszłości, na zakończenie samego ruchu.

W filozofii dobrze znana jest sytuacja, w której przyczynowe wyjaśnienie danego zdarzenia odwołuje się do późniejszego stanu rzeczy. Jest to wyjaśnianie oparte na tzw. przyczynie celowej („teleologicznej"), rozważanej już przez Arystotelesa. Wyjaśnianie takie stosujemy przede wszystkim do działań i zachowań ludzkich, a szerzej wszystkich istot obdarzonych świadomością. Jest naturalne wyjaśnić zachowanie łucznika mierzącego do tarczy jego przyszłym celem, czyli trafieniem strzały w tarczę. Trafienie w tarczę zostaje uznane za przyczynę celową wcześniejszego strzału z łuku. Jednakże pojawiają się tutaj wątpliwości co do fundamentalnego charakteru takiego wyjaśnienia. Wielu uważa, że faktyczną przyczyną strzału jest nie tyle późniejsze trafienie, co wcześniejsza myśl o trafieniu (chęć trafienia czy dążenie do niego). Dlatego w wyjaśnianiu celowym tak istotne jest ograniczenie do istot świadomych, zdolnych podejmować decyzje.

Mimo powyższych zastrzeżeń przyczyny celowe pojawiają się w naukach, głównie biologicznych. Skomplikowane struktury organów, takich jak serce czy płuca, często wyjaśnia się pełnioną przez te organy funkcją (pompowanie krwi, wymiana tlenu i dwutlenku węgla). Podobnie biologia ewolucyjna wyjaśnia wykształcenie się odpowiednich cech organizmów (fenotypu) tym, że ich obecność zwiększa szanse na przetrwanie organizmu. Jednakże zwykle podkreśla się cząstkowy i zastępczy charakter takiego wyjaśniania. „Prawdziwą" przyczyną wykształcenia się nowej cechy organizmów (np. pojawienia się przeciwstawnego kciuka u naczelnych) są złożone procesy fizyko-chemiczne na poziomie komórkowym uczestniczące w tworzeniu organizmów potomnych z komórek organizmów rodzicielskich, których opisem zajmuje się genetyka. Ze względu na niezwykły stopień komplikacji i zakres tych procesów, obejmujących niezliczone pokolenia, łatwiej zastosować upraszczające wyjaśnienie oparte na celowości danej zmiany. Jednak nauka podstawowa, jaką jest fizyka, nie powinna odwoływać się do tego typu uproszczeń.

Rola zasad ekstremalnych w mechanice podważa takie optymistyczne przekonanie. Sugerowane przez te zasady wyjaśnienie zachowania mechanicznego ciał opiera się na przyszłym „celu", jakim jest minimalizacja (maksymalizacja) pewnej wielkości, mimo że ciałom fizycznym nie możemy oczywiście przypisać świadomości, chęci i dążeń. Zauważmy, że istnieje zasadnicza różnica między stosowaniem wyjaśnień celowych w biologii a metodą opartą na zasadach ekstremalnych w fizyce. Biologiczne przyczyny celowe nie umożliwiają nam szczegółowych przewidywań. Nie jesteśmy w stanie precyzyjnie przewidzieć, jakie nowe cechy mogą pojawić się w organizmach w celu zwiększenia ich szansy na przetrwanie. Natomiast metoda ekstremalizacji działania w fizyce daje nam jednoznaczną odpowiedź na

pytanie, jaka będzie przyszła trajektoria ciała. Z tego powodu nie możemy traktować metody opartej na „celowości" jedynie jako zgrubnego przybliżenia dla dokładnej metody newtonowskiej, opartej na przyczynach sprawczych. Oczywiście przy pewnych założeniach obie metody są równoważne, można więc utrzymywać, że ontologiczny obraz świata klasycznego formalizmu Newtonowskiego jest bardziej fundamentalny. Czy jednak taka postawa nie jest rezultatem historycznego przypadku, że mechanikę klasyczną sformułowano najpierw w języku sił i przyczyn sprawczych? A co z argumentem, że zasady ekstremalne i oparte na nich równanie Eulera-Lagrange'a znajdują szerokie zastosowania poza obszarem zjawisk mechanicznych? Pozostawmy tę sprawę Czytelnikowi do samodzielnego rozważenia.

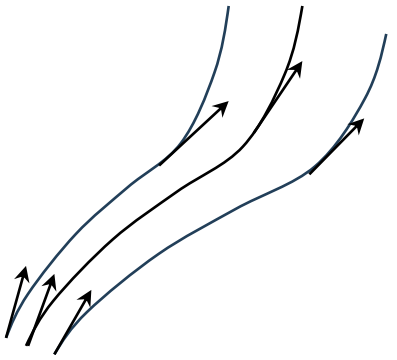
Ostatnim kamieniem milowym na drodze rozwoju mechaniki klasycznej, jaki omówimy w niniejszym rozdziale, będzie tzw. mechanika hamiltonowska. Podejście to powstało w dziewiętnastym wieku w rezultacie matematycznych badań nad istniejącym już formalizmem Lagrange'a. Zatem rozwinięcie formalizmu hamiltonowskiego nie było zainspirowane chęcią rozwiązania jakiegoś nowego problemu czy uogólnienia istniejącej teorii. Zasadniczo ujęcie mechaniki Lagrange'a, z fundamentalnym równaniem Eulera-Lagrange'a na czele, było zupełnie wystarczające. Jednakże nowatorstwo podejścia Hamiltona ujawniło się dużo później, kiedy zostało ono z powodzeniem zastosowane do nowych teorii, takich jak mechanika kwantowa czy ogólna teoria względności. Formalizm hamiltonowski odgrywa także bardzo ważną rolę w mechanice statystycznej.

Jak pamiętamy, mechanika Lagrange'a wprowadziła pojęcie uogólnionych współrzędnych, czyli w zasadzie dowolnych, niezależnych od siebie parametrów, charakteryzujących położenie obiektu w danym momencie. W mechanice Hamiltona dodatkowo definiuje się dla każdej takiej współrzędnej skorelowane z nią pojęcie uogólnionego pędu (taki pęd nazywa się sprzężonym z daną współrzędną). W wyniku analiz matematycznych okazało się, że dla typowych współrzędnych, takich jak liniowe współrzędne kartezjańskie, zróżniczkowanie lagrangianu po uogólnionej prędkości dla danej współrzędnej daje po prostu zwykły pęd (masa razy prędkość) wzdłuż tej współrzędnej. Podsunęło to myśl, aby taką „cząstkową" pochodną lagrangianu względem uogólnionej prędkości nazwać uogólnionym pędem dla wszelkich możliwych współrzędnych. Na przykład jeśli uogólnioną współrzędną jest kąt określający położenie rotującego ciała, to sprzężony z nią uogólniony pęd będzie momentem pędu tego ciała.

Podstawowe równania dynamiki w mechanice hamiltonowskiej uwzględniają nowy obiekt, zwany funkcją Hamiltona lub w skrócie hamiltonianem. Hamiltonian danego układu jest funkcją jego współrzędnych uogólnionych i sprzężonych z nimi pędów uogólnionych. Definiuje się go w dość skomplikowany sposób przy pomocy matematycznej operacji na lagrangianie[12], ale w większości typowych przypadków hamiltonian okazuje się równy całkowitej energii układu (energia kinetyczna plus potencjalna). Mechanika hamiltonowska opiera się na dwóch bardzo prostych równaniach. Pierwsze z nich zrównuje tempo zmiany danej współrzędnej uogólnionej (a zatem prędkość uogólnioną) z pochodną hamiltonianu względem sprzężonego z tą współrzędną pędu. Drugie równanie jest prawie „lustrzanym odbiciem" pierwszego: tempo zmiany uogólnionego pędu jest równe pochodnej hamiltonianu względem sprzężonej współrzędnej, ale ze znakiem minus. Dla typowych sytuacji drugie równanie Hamiltona jest w istocie równoważne z drugim prawem dynamiki Newtona (pochodna energii

[12] Operacja przekształcająca lagrangian danego układu w hamiltonian nazywa się transformacją Legendre'a.

po danej współrzędnej jest miarą siły działającej w kierunku tej współrzędnej). Pierwsze równanie z kolei to coś w rodzaju „definicji" prędkości, lub też lepiej związek prędkości z pędem.



Rys. 2.15. Przestrzeń fazowa (pędów i położeń) z zaznaczeniem trzech możliwych trajektorii. Strzałki reprezentują „uogólnione" prędkości (tempa zmian położeń i pędów), obliczone za pomocą równań Hamiltona

Mechanika hamiltonowska wykorzystuje ważne pojęcie przestrzeni fazowej. Jest to abstrakcyjna przestrzeń matematyczna, w której współrzędne punktów określone są za pomocą współrzędnych i pędów danego układu. Na przykład punktowe ciało swobodne wymaga sześciu parametrów, aby określić jego położenie i pęd – trzy parametry reprezentujące trzy współrzędne przestrzenne i trzy liczby charakteryzujące składowe pędu w kierunkach odpowiednich współrzędnych. Zatem przestrzeń fazowa dla takiego ciała musi mieć sześć wymiarów. Z kolei w wypadku opisywania układu N cząstek swobodnych wymiar przestrzeni fazowej rośnie do $6N$. Pojedynczy punkt w przestrzeni fazowej reprezentuje momentalny stan danego układu fizycznego (jak pamiętamy, położenia i pędy określają jednoznacznie stan cząstek materialnych w mechanice Newtonowskiej). Natomiast trajektorie (linie krzywe) wyznaczają możliwe ewolucje układu w czasie, czyli to, jak zmieniają się położenia i pędy ciał wchodzących w skład układu (rys. 2.15). Równania Hamiltona określają tempo zmian współrzędnych położenia i pędów, można więc powiedzieć, że charakteryzują one abstrakcyjną „prędkość" w przestrzeni fazowej. Dla danego hamiltonianu możemy w każdym punkcie przestrzeni fazowej zaczepić wektor reprezentujący prędkość zmiany każdej ze współrzędnych. Wektor ten wyznacza przebieg trajektorii dla danego punktu. Zatem ogólnie wszystkie możliwe ewolucje układu reprezentowane są w przestrzeni fazowej analogicznie do przepływu cieczy. Analogia z przepływem cieczy jest jeszcze pełniejsza ze względu na pewne fakty matematyczne, ale do tego wrócimy w paragrafie poświęconym mechanice statystycznej.

2.8.* Elementy mechaniki analitycznej

W niniejszym paragrafie, przeznaczonym dla bardziej zaawansowanych Czytelników, przedstawimy nieco ściślej matematyczne podstawy mechaniki analitycznej, a w szczególności podejścia Lagrange'a i Hamiltona. Zacznijmy od wprowadzenia pojęć energii potencjalnej i kinetycznej oraz pokazania, że druga zasada dynamiki implikuje zachowanie sumy obu energii dla pojedynczego ciała. Wzór na energię kinetyczną danego ciała jest dobrze znany ze szkolnego kursu fizyki:

$$T = \frac{1}{2}mv^2,$$

gdzie *v* jest prędkością ciała (ściślej, długością wektora prędkości). Energię potencjalną możemy wprowadzić przy dodatkowym założeniu dotyczącym sił. Zakładamy mianowicie, że siłę wypadkową działającą na ciało w danym kierunku *x* można przedstawić jako pochodną pewnej funkcji położenia *x*, którą oznaczymy jako *V*(*x*):

$$F_x = -\frac{dV(x)}{dx}.$$

Funkcję *V*(*x*) nazywamy energią potencjalną.[13] Założenie powyższe oznacza, że siła może być potraktowana jako nachylenie pewnej krzywej, w pewnej analogii do nachylenia pagórka. Energia *V* w ogólności jest funkcją trzech współrzędnych przestrzennych *x*, *y* i *z*. W takiej sytuacji formalnie pochodna względem danej współrzędnej nazywana jest pochodną cząstkową i oznaczana jest nieco inaczej niż „zwykła" pochodna. Jednakże w istocie jest to taka sama matematyczna operacja, tylko wykonuje się ją przy założeniu, że pozostałe zmienne są stałymi liczbami:[14]

$$F_x = -\frac{\partial V(x,y,z)}{\partial x}.$$[15]

Zapiszmy wzór na energię całkowitą (sumę energii kinetycznej i potencjalnej) dla przypadku jednowymiarowego:

$$E(x) = \frac{1}{2}m\left(\frac{dx}{dt}\right)^2 + V(x)$$

Funkcja *E*(*x*) nie ulega zmianie, gdy jej pochodna względem czasu jest równa zeru. Zróżniczkujmy powyższe wyrażenie po czasie, stosując zasadę łańcuchową, określającą jak różniczkować złożenie dwóch funkcji *f*[*g*(*x*)]. Zasada jest prosta: pochodna całości jest iloczynem pochodnej pierwszej funkcji *f* i drugiej funkcji *g* (w naszym wypadku *f* to funkcja kwadratowa prędkości $\frac{1}{2}mv^2$, a *g* to sama prędkość *v*):

$$\frac{dE(x)}{dt} = m\left(\frac{dx}{dt}\right)\left(\frac{d^2x}{dt^2}\right) + \frac{dV(x)}{dt}.$$

[13] Znak minus wynika z tego, że jeśli *V* jest rosnącą funkcją współrzędnej *x* (czyli energia potencjalna rośnie w kierunku dodatnich *x*-ów), to działająca siła będzie skierowana przeciwnie, w stronę malejących energii.

[14] Należy jednak pamiętać, że w ogólności pochodna cząstkowa funkcji wielu zmiennych po jednej zmiennej nie będzie równa pochodnej całkowitej po tej zmiennej. Pochodna całkowita uwzględnia również to, że inne zmienne mogą być zależne od zmiennej różniczkowania. Typowym przykładem takiej sytuacji jest funkcja, która zależy od zmiennej czasowej i zmiennych przestrzennych. Pochodna cząstkowa funkcji *f* po zmiennej czasowej *t* uwzględnia tylko jawną zależność *f* od czasu; pochodna całkowita bierze pod uwagę także zależność współrzędnych przestrzennych od czasu.

[15] Operacja brania pochodnych cząstkowych po współrzędnych *x*, *y*, *z* z pewnej funkcji skalarnej nazywa się gradientem z tej funkcji. Gradient funkcji *V* jest wektorem, którego składowe mają postać $\frac{\partial V}{\partial x}, \frac{\partial V}{\partial y}, \frac{\partial V}{\partial z}$. Oznacza się go w skrócie symbolem ∇V.

Zakładamy, że zależność energii V od czasu pochodzi jedynie od zmieniającego się położenia x ciała. W takiej sytuacji pochodna energii potencjalnej po czasie może być przedstawiona za pomocą tej samej reguły łańcuchowej jako pochodna po x razy pochodna x po czasie:

$$\frac{dE(x)}{dt} = m\left(\frac{dx}{dt}\right)\left(\frac{d^2x}{dt^2}\right) + \frac{dV(x)}{dx}\frac{dx}{dt}.$$

Widzimy zatem, że zrównanie powyższej formuły z zerem oznacza, że:

$$m\left(\frac{d^2x}{dt^2}\right) = -\frac{dV(x)}{dx}.$$

Prawa strona to nic innego jak działająca siła, a lewa to iloczyn masy i przyspieszenia. Jest to znane nam dobrze równanie ruchu Newtona (druga zasada dynamiki) $F = ma$. Zatem z drugiej zasady dynamiki wynika zasada zachowania całkowitej energii (jeśli tylko taka energia jest dobrze określona, tj. istnieje funkcja energii potencjalnej, która nie jest jawnie zależna od czasu).

Wprowadzimy teraz funkcję Lagrange'a, oznaczając ją przez $\mathcal{L}$.

$$\mathcal{L} = T - V.$$

Ponieważ energia kinetyczna jest funkcją prędkości, a potencjalna – położenia, lagrangian będzie funkcją obu tych zmiennych. W przypadku jednego wymiaru, zmiennymi lagrangianu będą: współrzędna x oraz prędkość w tym kierunku, czyli pochodna czasowa x. Dla uproszczenia zapisu będziemy stosowali przyjęte w fizyce oznaczenie pochodnej funkcji względem czasu jako funkcję „z kropką". Zatem prędkość w kierunku x to $\dot{x}$, a lagrangian przedstawimy jako $\mathcal{L}(x, \dot{x})$. Przy pomocy lagrangianu definiujemy całkowite działanie $\mathcal{A}$ na pewnej drodze pokonywanej przez ciało w interwale od chwili t_1 do t_2:

$$\mathcal{A} = \int_{t_1}^{t_2} \mathcal{L}(x, \dot{x})dt.$$

Całkę powyższą możemy traktować w przybliżeniu jako sumę wartości lagrangianu w niewielkich interwałach czasu Δt razy dany interwał: $\sum \mathcal{L}\Delta t$, gdzie $\sum \Delta t = t_2 - t_1$. Ważne jest, aby zwrócić uwagę, że wartość działania zależy od przyjętej trajektorii ciała, czyli funkcji $x(t)$. Zależność lagrangianu od czasu jest bowiem wynikiem „złożenia" dwóch zależności: położenia i prędkości ciała od czasu (czyli jego trajektorii) oraz zależności lagrangianu od tych dwóch parametrów. Na działanie można więc spojrzeć jako na funkcję, przypisującą danej trajektorii $x(t)$ pewną liczbę. Taką funkcję nazywamy funkcjonałem.

Możemy następnie zastanowić się, co będzie się działo z naszym funkcjonałem przy niewielkich zmianach trajektorii. Proces ten nazywamy obliczaniem wariacji danego funkcjonału. Jeśli dla danej trajektorii wariacja funkcjonału jest zerowa, tj. niewielkie (infinitezymalne) odejścia od tej trajektorii nie powodują zmiany wartości funkcjonału, to znaczy, że funkcjonał dla tej trajektorii przyjmuje wartość ekstremalną (minimalną bądź maksymalną). Zatem warunek ekstremalizacji przyjmuje postać równania:

$$\delta\mathcal{A} = 0.$$

gdzie δ jest wariacją. Warunek zerowania się wariacji danego funkcjonału można zamienić na warunek wyrażony przy pomocy „zwykłych" pochodnych odpowiednich funkcji wchodzących w skład tego funkcjonału. Nie będziemy tego pokazywać, ale okazuje się, że w przypadku działania zdefiniowanego w powyższy sposób za pomocą lagrangianu, zerowanie się jego wariacji jest równoważne następującemu równaniu, zwanemu równaniem Eulera-Lagrange'a (dla przypadku jednowymiarowego)[16]:

$$\frac{d}{dt}\left(\frac{\partial \mathcal{L}}{\partial \dot{x}}\right) - \frac{\partial \mathcal{L}}{\partial x} = 0 \tag{2.2}$$

Jest to równanie różniczkowe, czyli charakteryzuje ono nasz układ w każdej chwili jego ewolucji. Wynika z niego, że jeśli w danej chwili obliczymy, jak zmienia się nasz lagrangian wraz ze zmianą jednego argumentu x i drugiego argumentu $\dot{x}$, to tempo zmiany w czasie jednego parametru zmienności (względem $\dot{x}$) będzie równe drugiemu z nich (względem x). Warto zwrócić uwagę, że w ten sposób zmieniamy perspektywę z „globalnej" (minimalizacja działania na całej drodze of t_1 do t_2) na „lokalną" (współzależność pewnych parametrów chwilowych). Wracamy zatem do ujęcia zjawisk mechanicznych w kategoriach przyczyn sprawczych działających w danym momencie, zamiast przyczyn celowych obejmujących przyszłe trajektorie. Nie będziemy rozstrzygać, która z tych filozoficznych perspektyw jest bardziej fundamentalna.

Możemy się przekonać, korzystając z definicji lagrangianu, że równanie (2.2) jest w istocie drugą zasadą dynamiki Newtona. Zrobimy to, obliczając pochodne cząstkowe lagrangianu po zmiennych x i $\dot{x}$. Zauważmy, że dla pojedynczego ciała w polu sił jawna zależność jego lagrangianu od prędkości $\dot{x}$ bierze się tylko z energii kinetycznej, a zależność od x jest zawarta w energii potencjalnej. Mamy zatem następujące równania:

$$\frac{\partial \mathcal{L}}{\partial \dot{x}} = m\dot{x}$$
$$\frac{\partial \mathcal{L}}{\partial x} = \frac{\partial V}{\partial x}.$$

Wstawiając powyższe do równania Eulera-Lagrange'a, otrzymujemy

$$m\ddot{x} = -\frac{\partial V}{\partial x},$$

czyli oczywiście dobrze znane równanie $F = ma$ (dwie kropki oznaczają drugą pochodną po czasie). Przy okazji przekonaliśmy się, że pochodna lagrangianu po $\dot{x}$ to nic innego, jak pęd ciała mv. Prowadzi to wprost do definicji pędu uogólnionego sprzężonego ze współrzędną x. W ogólnym wypadku lagrangian może być traktowany jako funkcja dowolnych współrzędnych q (i ich tempa zmian), a wtedy sprzężony z daną współrzędną pęd p będzie zdefiniowany dokładnie jako $\frac{\partial \mathcal{L}}{\partial \dot{q}}$.

W ten sposób przechodzimy do formalizmu hamiltonowskiego. Rozważmy ogólny przypadek układu fizycznego, który charakteryzuje się pewną liczbą współrzędnych uogólnio-

[16] W przypadku trójwymiarowym równanie Eulera-Lagrange'a wygląda tak samo dla każdej współrzędnej z osobna.

nych q_i. Tak jak powyżej, definiujemy dla każdej współrzędnej q_i sprzężony z nią pęd uogólniony:

$$p_i = \frac{\partial \mathcal{L}(q_i, \dot{q}_i)}{\partial \dot{q}_i}.$$

Wprowadzimy teraz nową funkcję H współrzędnych i pędów, zwaną hamiltonianem:

$$H = \sum_i p_i \dot{q}_i - \mathcal{L}. \tag{2.3}$$

Definicja powyższa może wydać trochę dziwna i nieintuicyjna, ale przemawia za nią pewna ważna obserwacja. Otóż jeśli obliczy się pochodną całkowitą wyrażenia na H względem czasu $\frac{dH}{dt}$, to po dokonaniu odpowiednich transformacji okaże się, że będzie ona dokładnie równa pochodnej cząstkowej lagrangianu po czasie $\frac{\partial \mathcal{L}}{\partial t}$ (ze znakiem minus). Pochodna cząstkowa po czasie charakteryzuje jawną zależność danego parametru od czasu. Lagrangian oczywiście jest zwykle zależny od czasu ze względu na jego zależność od współrzędnych, które z kolei zmieniają się w czasie podczas ewolucji układu. Jednakże typowe lagrangiany nie zawierają jawnej zależności od czasu – są innymi słowy niezmiennicze względem przesunięć w czasie. Zatem hamiltonian danego układu będzie funkcją, która nie zmienia się w ogóle w czasie, jeśli tylko lagrangian nie zależy jawnie od czasu.

Nietrudno sprawdzić, że dla typowych sytuacji hamiltonian jest po prostu tożsamy z całkowitą energią układu, która, jak już wiemy, powinna być zachowana. Zbadajmy to znowu dla przykładu cząstki poruszającej się w jednym wymiarze x w polu sił. W tym wypadku lagrangian ma postać (zapisaną za pomocą pędu p_x):

$$\frac{{p_x}^2}{2m} - V(x).$$

Oczywiście $p_x \dot{x} = \frac{p_x^2}{m}$, a zatem wstawiając do wzoru (2.3), otrzymamy całkowitą energię:

$$H = \frac{{p_x}^2}{2m} + V(x). \tag{2.4}$$

Równania Hamiltona mają następującą ogólną postać:

$$\begin{aligned} \dot{q}_i &= \frac{\partial H}{\partial p_i} \\ \dot{p}_i &= -\frac{\partial H}{\partial q_i}. \end{aligned} \tag{2.5}$$

Znów, wstawiając za q_i współrzędną liniową x, za p_i sprzężony z nią pęd p_x, a w miejsce hamiltonianu formułę (2.4), z równań Hamiltona dostajemy

$$\begin{aligned} \dot{x} &= \frac{p_x}{m} \\ \dot{p}_x &= -\frac{\partial V(x)}{\partial x}. \end{aligned}$$

Pierwsze z powyższych równań to dobrze znana zależność pęd = masa razy prędkość. Drugie to oczywiście zasada dynamiki Newtona: tempo zmiany pędu równa się przyłożonej sile. Jeszcze raz podkreślmy, że choć równania Hamiltona w szczególnych przypadkach są równoważne zasadom dynamiki newtonowskiej, to jednak stosuje się je również tam, gdzie mechanika newtonowska już przestaje działać.

Na koniec rozdziału poświęconego mechanice klasycznej przyjrzyjmy się bliżej wspomnianej wcześniej relacji pomiędzy symetriami a prawami zachowania (tw. Noether). Symetrią danego obiektu nazwiemy przekształcenie, które nie zmienia tego obiektu (obiekt ten jest inwariantem tego przekształcenia). Pamiętamy z wcześniejszego paragrafu, że transformacje Galileusza są symetriami dla równań Newtona. Obecnie będziemy rozważać symetrie konkretnych funkcji Lagrange'a. Na przykład może się zdarzyć, że dany lagrangian nie zmienia się przy przesunięciach przestrzennych (translacjach) lub przy obrotach. Oznacza to, że oddziaływania, którym podlega dany układ, nie wyróżniają żadnego miejsca w przestrzeni albo żadnego kierunku.

Formalnie transformacje przestrzenne opisujemy w postaci transformacji współrzędnych. Na przykład translacja w danym kierunku może być przedstawiona jako transformacja $x \to x + d$, gdzie d jest pewną liczbą. Przekształcenia, takie jak translacje i obroty, należą do kategorii przekształceń ciągłych, dla których można podać tzw. infinitezymalne generatory. Na przykład infinitezymalnym przesunięciem wzdłuż osi x będzie transformacja $x \to x + \delta$, gdzie δ jest arbitralnie małą wielkością. Ogólnie rzecz biorąc, jeśli rozważamy dowolne przekształcenie geometryczne działające na współrzędnych przestrzennych q_i, to jego infinitezymalny generator będzie miał postać

$$\delta q_i = f(q)\delta,$$

gdzie f jest pewną funkcją, a q symbolicznie oznacza wszystkie zmienne q_j. Funkcja f zatem może zależeć od innych współrzędnych niż transformowana współrzędna q_i (tak jest np. w wypadku obrotów). Natomiast dla translacji funkcja $f(q)$ przyjmuje po prostu wartość 1.

Jak będzie wyglądała infinitezymalna zmiana (wariacja) lagrangianu poddanego powyższej transformacji współrzędnych? Odpowiedzi dostarcza rachunek różniczkowy, a jest ona nieco skomplikowana. W najprostszym przypadku, kiedy mamy funkcję jednej zmiennej $F(x)$, zmiana tej funkcji podczas infinitezymalnej zmiany wartości x dana jest następującym wzorem:

$$\delta F = \frac{dF}{dx}\delta x.$$

Można sobie wyobrażać, że dx w mianowniku i δx ulegają „skróceniu", a w tym wypadku d i δ oznaczają te same infinitezymalne zmiany F. Jednak w ogólnym przypadku wariacja lagrangianu pod wpływem odpowiedniej infinitezymalnej transformacji będzie bardziej złożona z tego względu, że lagrangian zależy od wielu współrzędnych q_i oraz $\dot{q}_i$. Ogólna formuła, będąca uogólnieniem powyższego wzoru, jest następująca:

$$\delta\mathcal{L} = \sum_i \left(\frac{\partial\mathcal{L}}{\partial q_i}\delta q_i + \frac{\delta\mathcal{L}}{\partial\dot{q}_i}\delta\dot{q}_i\right).$$

Zauważając, że $\frac{\delta \mathcal{L}}{\partial \dot{q}_i}$ to nic innego, jak pęd uogólniony p_i, i stosując równanie Eulera-Lagrange'a do pierwszego składnika sumy, możemy uzyskać następującą formułę na infinitezymalną wariację lagrangianu:

$$\delta \mathcal{L} = \sum_i (\dot{p}_i \delta q_i + p_i \delta \dot{q}_i).$$

Czytelnicy zaznajomieni nieco z podstawowymi regułami różniczkowania wiedzą, że pochodna iloczynu funkcji *FG* ma postać pochodnej z *F* razy *G* plus *F* razy pochodna z *G*. Wyrażenie pod sumą ma podobną strukturę (pamiętamy, że kropka oznacza różniczkowanie po czasie), a zatem możemy pochodną czasową niejako wyciągnąć przed nawias:

$$\delta \mathcal{L} = \frac{d}{dt} \sum_i (p_i \delta q_i).$$

Teraz możemy wstawić wyrażenie na infinitezymalną zmianę współrzędnej q_i, otrzymując:

$$\delta \mathcal{L} = \frac{d}{dt} \sum_i (p_i f(q)) \delta.$$

Jeśli rozważana transformacja jest symetrią lagrangianu, jego wariacja pod wpływem tej transformacji powinna być równa zeru. Oznacza to, że pochodna po czasie z odpowiedniego wyrażenia jest również zerowa, a zatem wyrażenie to nie zmienia się w czasie. Innymi słowy, jest ono zachowane. Uzyskaliśmy zatem rezultat, że istnienie symetrii przestrzennych zawsze łączy się z pewną zasadą zachowania. Jeśli rozważaną symetrią jest translacja, to ponieważ w tym wypadku $f(q) = 1$, wielkością zachowaną będzie suma $\sum_i p_i$, czyli pęd całkowity układu. Zatem zasada zachowania pędu wynika z niezmienniczości względem translacji przestrzennych. Z kolei można pokazać (nie będziemy tego robić), że dla obrotów wielkością zachowaną będzie moment pędu.

Warto w tym momencie zauważyć, że nie każdy lagrangian spełnia warunek symetryczności względem translacji przestrzennych. Na przykład lagrangian $\mathcal{L} = T - V$ dla pojedynczej cząstki w polu sił nie jest symetryczny, chyba że funkcja $V(x)$ jest stała. Jest dość oczywiste, że w takiej sytuacji pęd cząstki nie zostaje zachowany, gdyż będzie ona przyspieszać pod wpływem siły. Natomiast jeśli rozważymy dwie oddziałujące cząstki, których energia potencjalna zależy jedynie od ich relatywnej odległości $V(x_1 - x_2)$, to taki lagrangian już jest symetryczny względem przesunięć, a zatem pęd całkowity będzie zachowany.

Istnieje również zależność między symetrią czasową a zachowaniem pewnej wielkości. Jak już zauważyliśmy powyżej, funkcja Hamiltona została zdefiniowania przy pomocy lagrangianu w taki sposób, aby zachodziła następująca równość:[17]

$$\frac{dH}{dt} = -\frac{\partial \mathcal{L}}{\partial t},$$

[17] Można to pokazać, obliczając jawnie pochodną całkowitą lagrangianu po czasie. Otrzyma się wtedy następujący wzór: $\frac{d\mathcal{L}}{dt} = \frac{\partial \mathcal{L}}{\partial t} + \sum_i \left(\frac{\partial \mathcal{L}}{\partial q_i} \dot{q}_i + \frac{\partial \mathcal{L}}{\partial \dot{q}_i} \ddot{q}_i \right)$. Następnie należy z równania Eulera-Lagrange'a zamienić pochodną cząstkową lagrangianu po q_i na pochodną czasową tempa zmiany po prędkości $\dot{q}_i$ oraz skorzystać z definicji pędu uogólnionego p_i, co da nam następującą sumę $\sum_i (\dot{p}_i \dot{q}_i + p_i \ddot{q}_i) = \frac{d}{dt} \sum_i p_i \dot{q}_i$.

czyli całkowite tempo zmiany hamiltonianu względem czasu jest równe cząstkowemu tempu zmiany lagrangianu z minusem. Oznacza to, że jeśli w lagrangianie nie występuje jawnie zmienna czasowa, to hamiltonian całego układu nie zmienia się w czasie, a zatem jest zachowany. Ale hamiltonian to nic innego jak energia układu. Zatem istnienie symetrii lagrangianu względem translacji czasowych implikuje, że energia układu będzie zachowana.

Pytania i problemy

1. Czy można zinterpretować prawa dynamiki Newtona jako ukryte definicje niektórych występujących w nich terminów? Które to będą terminy? Czy przyjęcie odpowiednich postulatów znaczeniowych redukuje prawa Newtona do zdań analitycznych *a priori*?

2. Czy możliwe jest definitywne obalenie pierwszego i drugiego prawa dynamiki Newtona?

3. Pokaż, że definicja masy przedmiotu w formie postulatu „Każde dwa przedmioty oddziałują ze sobą w taki sposób, że iloczyn ich mas jest równy odwrotności iloczynu przyspieszeń" jest twórcza. Wskazówka: rozważ trzy ciała *a*, *b* i *c* i zastosuj powyższy postulat do ich oddziaływań parami: *a* z *b*, *b* z *c* i *a* z *c*. Udowodnij, że z przyjętego postulatu wynika pewna zależność, która nie zawiera w ogóle mas ciał *a*, *b* i *c*.

4. Wytłumacz, w jaki sposób drugie prawo dynamiki Newtona umożliwia obliczenie trajektorii danego ciała poddanego działaniu sił. Jakie dodatkowe informacje na temat tego ciała są tutaj potrzebne?

5. Przedstaw dwie wersje tezy determinizmu: ontologiczną i epistemologiczną. Jakie są między nimi zależności logiczne? Od czego zależy sens determinizmu epistemologicznego i jego ewentualna trafność?

6. Jak można ogólnie scharakteryzować pojęcie stanu układu fizycznego w danej chwili? Dlaczego chwilowy stan układu nie powinien zawierać żadnej informacji na temat zachowania układu po tej chwili? Czy stan układu opisany funkcją prędkości (interpretowaną jako pochodna funkcji położenia) spełnia ten warunek?

7. Omów przykłady sugerujące, że mechanika klasyczna nie jest teorią w pełni deterministyczną.

8. Czy zjawiska chaotyczne podważają ważność determinizmu ontologicznego, czy epistemologicznego?

9. Jak wygląda matematyczna forma transformacji Galileusza i jej fizyczny sens? Podaj przykład inwariantu tej transformacji.

10. Przedstaw dwa rozumienia sporu o absolutność czasu i przestrzeni: fizyczne i filozoficzne (metafizyczne).

11. Co to jest przestrzeń „migawkowa" (momentalna)? Czy jest ona inwariantem transformacji Galileusza? Rozważ to samo pytanie w odniesieniu do przestrzeni rozciągłej czasowo (trwającej w czasie).

12. Przeanalizuj argument Newtona z wiadrem za istnieniem przestrzeni absolutnej. Czy argument ten podaje nam metodę identyfikacji absolutnych punktów przestrzeni?

13. Jaką lukę w argumencie Newtona ujawnił Ernst Mach? Przedstaw słabą i silną interpretację argumentu Macha.

14. Przedstaw argument Leibniza z przesunięcia przeciwko absolutyzmowi (za relacjonizmem) w sporze o status przestrzeni. Jaką rolę w tym argumencie odgrywają założenia metafizyczne (zasada tożsamości przedmiotów nieodróżnialnych, zasada racji dostatecznej)?

15. Argument Leibniza za relacjonizmem można sformułować w wersji „dynamicznej", w której zamiast o absolutnym położeniu wszechświata mówi się o jego prędkości względem absolutnej przestrzeni. Spróbuj zrekonstruować taki argument i porównać go z wersją „statyczną". Czy argument dynamiczny musi opierać się na zasadzie tożsamości przedmiotów nieodróżnialnych, czy też wystarczy przesłanka głosząca, że nie powinniśmy przyjmować istnienia żadnych faktów, które nie mogą być empirycznie zweryfikowane?

16. Przeanalizuj, w jaki sposób Newton uzasadnił, że planety muszą być utrzymywane na orbitach siłą centralną od Słońca odwrotnie proporcjonalną do kwadratu odległości. Wskaż na rolę praw Keplera w kolejnych krokach argumentacji Newtona. Zwróć uwagę na metodę przybliżania gładkich trajektorii (po okręgu czy elipsie) wielokątami.

17. Omów problem „działania na odległość" w przypadku sił grawitacyjnych. Jakie było podejście Newtona do tego problemu?

18. Na czym polega ogólna idea wyznaczania trajektorii ruchu przy pomocy zasady najmniejszego działania? Jakie konsekwencje dla przyczynowego wyjaśniania ruchu ma to podejście?

19. Przedstaw w ogólnych zarysach istotę mechaniki w ujęciu hamiltonowskim, wykorzystując pojęcie przestrzeni fazowej. Jaki jest wymiar przestrzeni fazowej opisującej ruch N ciał punktowych?

Literatura uzupełniająca

Problem statusu praw dynamiki Newtona oraz zagadnienie czasu i przestrzeni w mechanice klasycznej omówione są w znanej pracy: E. Nagel, *Struktura nauki: zagadnienia logiki wyjaśnień naukowych*, PWN Warszawa 1970, rozdziały 7 i 8.

Kwestia determinizmu mechaniki klasycznej jest dyskutowana w monumentalnej pracy: J. Earman, *A Primer on Determinism*, D. Reidel, Dordrecht 1983.

Godne polecenia wprowadzenie do teorii chaosu dla filozofów: P. Smith, *Explaining Chaos*, Cambridge University Press, Cambridge 1998.

Szczegóły wyprowadzenia przez Newtona prawa grawitacji i wiele innych interesujących faktów dotyczących astronomii i mechaniki niebieskiej zawarte są w pracy: J. Cushing, *Philosophical Concepts in Physics*, Cambridge University Press, Cambridge 1998.

Omówienie historycznego rozwoju mechaniki klasycznej od starożytności do czasów współczesnych, ze szczególnym uwzględnieniem mechaniki analitycznej po Newtonie, znajdziecie w książce: L. Sklar, *Philosophy and the Foundations of Dynamics*, Cambridge University Press, Cambridge 2013.

Dla osób zainteresowanych bardziej gruntowną analizą matematycznego aparatu mechaniki klasycznej wartościową pozycją będzie niedawno przetłumaczona książka: L. Susskind, G. Hrabowsky, *Teoretyczne minimum. Co musisz wiedzieć, żeby zacząć zajmować się fizyką*. Prószyński i S-ka, Warszawa 2022.

ROZDZIAŁ 3. NAUKA O CIEPLE

Zjawiska mechaniczne, takie jak ruch, zderzenia ciał, przekazywanie energii i pędu itp., stanowią trzon fizyki newtonowskiej. Jednakże istnieją typy procesów występujących w świecie fizycznym, które wydają się zasadniczo odmienne od prostych interakcji mechanicznych. Do zjawisk tych zaliczają się przede wszystkim różnego rodzaju procesy cieplne – ogrzewanie, ochładzanie, spalanie, topnienie, krzepnięcie, parowanie, kondensacja. Szczególnie interesujące jest zachowanie cieczy i gazów, które nie posiadają dobrze określonego kształtu jak ciała stałe, ale przyjmują kształt pojemnika. Już w starożytności zwrócono uwagę na specyficzne zachowanie się płynów, formułując interesujące prawa hydrostatyki, np. znane każdemu prawo Archimedesa. Zjawiskami dotyczącymi gazów zaczęto się natomiast dokładniej zajmować dużo później – w siedemnastym i osiemnastym wieku, kiedy technologia umożliwiła produkcję szczelnych naczyń, dzięki którym można było badać zmiany zachodzące w określonej próbce gazu. Szybko zauważono np., że ogrzewanie próbki gazu powoduje zwiększenie ciśnienia wywieranego na ścianki naczynia. Jeśli ciśnienie utrzymamy na stałym poziomie przez zastosowanie ruchomego tłoka, to dostarczenie ciepła zwiększy objętość gazu, a przy okazji tłok wykona pewną pracę. Z kolei sprężenie próbki gazu przez dostarczenie jej pewnej energii mechanicznej skutkuje zwiększeniem temperatury gazu. Podobne obserwacje posłużyły do sformułowania ilościowych praw korelujących ze sobą makroskopowe parametry ciśnienia, temperatury i objętości, co z kolei otworzyło drogę do stworzenia nowej teorii łączącej zjawiska cieplne z makroskopowymi procesami mechanicznymi (pracą mechaniczną), zwanej termodynamiką.

Prawo wyporu Archimedesa głosi, że ciało zanurzone w wodzie traci na wadze dokładnie tyle, ile waży woda wyparta przez to ciało. Zasada ta ma ciekawe i nieoczywiste konsekwencje. Rozważmy na przykład następujący problem. Wyobraźmy sobie, że pewne naczynie zostało wypełnione po brzegi wodą, w której pływa kawałek lodu. Co się stanie, kiedy lód ten się stopi? Czy woda przeleje się przez brzegi naczynia, czy też jej poziom nieco opadnie? Wydaje się, że bez odwołania się do doświadczenia nie będziemy mogli tego rozstrzygnąć. A jednak prawo Archimedesa daje nam natychmiastową odpowiedź: poziom wody się nie zmieni (możecie to sprawdzić sami, jeśli nie wierzycie słynnemu fizykowi z greckich Syrakuz). Powód jest prosty: ponieważ kawałek lodu pływa po wodzie, wypiera dokładnie tyle wody, ile sam waży, a zatem po stopnieniu woda powstała z lodu

zajmie dokładnie objętość wypieraną uprzednio przez lód. Przy okazji możemy dostrzec powód, dla którego pływające kawałki lodu wystają nieco ponad powierzchnię – jest to wynik różnicy między gęstością wody a gęstością lodu. Woda o objętości równej objętości zanurzonej części lodu musi ważyć tyle samo, co cały kawałek lodu, a zatem lód ma mniejszą gęstość niż woda.

Dlaczego zjawiska cieplne zasługują na uwagę filozofa? Powodów jest wiele. Na plan pierwszy wysuwa się pytanie o charakterze ontologicznym: czy ciepło i procesy z nim związane stanowią osobną kategorię własności obiektów materialnych, czy też może są one tylko przejawami ukrytych procesów o *stricte* mechanicznym charakterze, takich jak ruch i zderzenia? Innymi słowy, pytamy o ontologiczną naturę ciepła i jego związki z innymi znanymi własnościami materii. Naturalną odpowiedzią na to pytanie, przyjętą na początku rozwoju nauki o cieple, jest teza o ontologicznej niezależności zjawisk cieplnych od pozostałych zjawisk fizycznych. Konkretną realizacją tej doktryny była teoria tzw. cieplika, czyli niewidzialnego fluidu, który miał być odpowiedzialny za wszelkie procesy cieplne – ogrzewanie, oziębianie itd. Na przykład proces przekazywania ciepła od ciała gorącego do zimnego przedstawiany był jako przepływ pewnej ilości cieplika zawartego w pierwszym ciele do ciała drugiego aż do osiągnięcia poziomu równowagi. Jednakże rozwój nauki zadał cios tej doktrynie. Bliska zależność między ciepłem a energią mechaniczną, o której będziemy mówić dokładniej w dalszych częściach rozdziału, nasunęła myśl o możliwej redukcji zjawisk cieplnych do procesów czysto mechanicznych, zachodzących na mikroskopowym poziomie molekularnym. Prowadzi to do pytania o charakterze filozoficzno-metodologicznym, czym jest redukcja jednego zjawiska, własności czy opisu do innych zjawisk (własności, opisów).

Z pojęciem redukcji możemy się zetknąć w wielu obszarach filozofii i metodologii. Typowe filozoficzne redukcje dotyczą pewnych kategorii ontologicznych – na przykład niektórzy usiłują zredukować kategorię przedmiotów konkretnych do własności i relacji; inni z kolei proponują redukcje odwrotne, w których przedmioty abstrakcyjne sprowadzane są do pewnych konstruktów opartych na konkretach. Na styku filozofii i nauki spotykamy niezmiernie ważne pytanie o możliwość redukcji procesów i własności mentalnych (takich jak myśli, pragnienia i emocje) do zjawisk czysto fizycznych czy też neurologicznych zachodzących w mózgu. Dotyka to fundamentalnego sporu o naturę umysłu między dualizmem a różnymi formami fizykalnego monizmu. Analiza zjawisk termodynamicznych w kategoriach mechaniki newtonowskiej dostarcza wzorcowego przykładu udanej redukcji między teoriami, dlatego też warto przyjrzeć się dokładniej strukturze i własnościom tej procedury redukcyjnej.

Kolejnym powodem zainteresowania filozofów termodynamiką jest jej zaskakujący i subtelny związek z problematyką filozofii czasu. Doświadczenie potoczne uczy nas, że czas posiada jednoznacznie wyznaczony kierunek. Jak się wydaje, istnieje obiektywna różnica między przeszłością a przyszłością – przeszłość jest ustalona, zamknięta, a przyszłość otwarta. Natomiast podstawowa teoria fizyczna, jaką jest mechanika klasyczna, nie potwierdza tej asymetrii czasu. Każdy proces dopuszczony przez prawa fizyki newtonowskiej może równie dobrze zachodzić odwrotnie w czasie. Pod tym względem termodynamika różni się zasadniczo od mechaniki klasycznej. Procesy z udziałem ciepła mają ewidentnie kierunkowy, nieodwracalny charakter – na przykład ciepło przepływa zawsze od ciała cieplejszego do

zimniejszego, nigdy na odwrót.[1] Ta nieodwracalność czasowa znajduje swój wyraz w fundamentalnej zasadzie, zwanej drugą zasadą termodynamiki. Powstaje zatem pytanie, jak pogodzić asymetrię czasową praw termodynamiki z symetrią praw mechaniki klasycznej, w świetle stwierdzonego wyżej faktu, iż zjawiska cieplne nie są niczym innym jak mechanicznym ruchem i wzajemnym oddziaływaniem ogromnej liczby molekuł. Nad rozwiązaniem tej zagadki głowiło się wielu fizyków i filozofów, a kluczem jest tutaj statystyczny (probabilistyczny) charakter opisu na poziomie molekuł. Przyjrzymy się dokładniej sposobom wyprowadzenia asymetrycznych praw termodynamiki z symetrycznych założeń mechaniki klasycznej połączonych z pewnymi założeniami statystycznymi. Jak się okaże, aby uzyskać pożądany rezultat, będziemy musieli przyjąć dodatkowe założenia wykraczające poza obie teorie, a dotyczące pewnych globalnych własności całego wszechświata.

3.1. Ciepło i temperatura

Zacznijmy jednak od najprostszego pytania: czym w ogóle jest ciepło? Jak się je mierzy? W życiu codziennym nie mamy problemu z zastosowaniem pojęcia ciepła – wiemy, że ogień jest gorący, a lód zimny. Za pomocą zmysłu dotyku możemy porównywać ciała pod względem ich ciepłoty (uważając oczywiście, aby się przy tym nie oparzyć). Ciepło w potocznym sensie bywa utożsamiane z temperaturą, jednak w fizyce te dwa pojęcia należy odróżnić. Ciepło w sensie fizycznym należy do kategorii wielkości zwanych *ekstensywnymi*, podczas gdy temperatura jest wielkością *intensywną*. Wielkości ekstensywne charakteryzujące dany obiekt zależą od jego rozmiarów, podczas gdy intensywne nie (można powiedzieć nieformalnie, że wielkości ekstensywne są globalne, a intensywne lokalne, zmieniające się od punktu do punktu). Ciepło zawarte w wannie z wodą o temperaturze 35 stopni Celsjusza jest dużo większe od ciepła szklanki z wodą o tej samej temperaturze. Łączne ciepło dwóch ciał jest sumą ciepła każdego ciała z osobna, podczas gdy temperatura nie sumuje się w ten sposób.

Różnicę między ciepłem a temperaturą można elegancko objaśnić, korzystając z zasadniczo niepoprawnego, ale użytecznego modelu cieplika. Załóżmy, że w każdym ciele zawarta jest pewna ilość niewidocznego fluidu cieplikowego, który może przyjmować różne stopnie gęstości. Im wyższa jest temperatura danego ciała, tym większa gęstość cieplika, co przy stałej objętości ciała daje nam większą ilość fluidu. Ciepło jest miarą całkowitej ilości cieplika w ciele, podczas gdy jego gęstość definiuje temperaturę. Przekazanie części cieplika z jednego ciała do drugiego powoduje jego rozrzedzenie w pierwszym ciele i zagęszczenie w drugim, czyli temperatura pierwszego ciała opada, a drugiego rośnie. Zauważmy ponadto,

[1] Uważny czytelnik może w tym momencie zaprotestować: przecież w mechanice klasycznej też mamy do czynienia z mnóstwem procesów, których nie da się odwrócić. Na przykład kulka stacza się z pagórka w polu grawitacyjnym, ale nigdy się nie zdarza, żeby się samoczynnie wtoczyła z powrotem na szczyt. Musimy jednak unikać wyciągania pochopnych wniosków. Mechanika klasyczna opisuje proces, w wyniku którego ciało o pewnej energii potencjalnej (na szczycie pagórka) zaczyna tracić tę energię na korzyść energii kinetycznej, zwiększając swoją prędkość. Pod koniec procesu staczania się kulka ma największą prędkość. Tak opisany proces jest absolutnie odwracalny, jeśli tylko zastosujemy niezbędne odwrócenie zwrotu prędkości końcowej kulki: mamy wtedy do czynienia z kulką, której nadano prędkość początkową, umożliwiającą jej wtoczenie się na szczyt pagórka. Zgoda, ale z doświadczenia wiemy, że kulka w końcu zatrzyma się u podnóża góry – to jednak wykracza już poza mechanikę newtonowską. Energia kinetyczna kulki w wyniku tarcia została zamieniona na ciepło, a to już jest proces nieodwracalny, opisywany przez termodynamikę.

że całkowita ilość cieplika w układzie izolowanym nie może ulec zmianie – ile cieplika opuszcza jedno ciało, tyle musi pojawić się w drugim ciele.

Jak wyrazić liczbowo miarę temperatury danego ciała? Jesteśmy przyzwyczajeni do używania termometrów w celu określania temperatury np. ciała ludzkiego. Jak jednak działa termometr i co naprawdę mierzy? Rozważmy prosty termometr oparty na zjawisku rozszerzalności cieplnej pewnej cieczy zawartej w cienkiej rurce. Cieczą taką może być rtęć (dzisiaj wycofana z powszechnego użytku z powodów ekologicznych) czy alkohol. Podstawą działania takiego termometru jest założenie, że rozszerzanie cieplne danego materiału jest proporcjonalne do różnicy między temperaturą końcową a początkową. Przy tym założeniu możemy wyskalować termometr, odznaczając równe odcinki i przyjmując, że odpowiadają im równe skoki temperatury. Pozostaje oczywiście wątpliwość, czy przyjęcie założenia proporcjonalnego rozszerzania się roboczej cieczy nie wikła nas w błędne koło, bo aby stwierdzić, że termometr działa poprawnie, musimy użyć termometru w celu sprawdzenia poprawności naszego założenia. Z podobnym problemem zetknęliśmy się już przy okazji analizy praw Newtona, a zatem nasze rozwiązanie będzie również podobne. Musimy odwołać się do pewnej konwencji, ale będzie ona miała niepomijalny komponent empiryczny w postaci stwierdzenia, że wskazania różnych termometrów powinny być ze sobą zgodne.

Pomiary temperatury są obarczone dodatkowym problemem, który nie dotykał pomiarów czasu dyskutowanych wcześniej. Chodzi o to, że każdy pomiar temperatury polegający na kontakcie fizycznym przedmiotu mierzonego z termometrem wpływa na mierzoną wartość. Termometr odbiera pewną ilość ciepła od mierzonego obiektu, tym samym zmniejszając jego temperaturę.[2] Oczywiście w praktyce efekt ten można uczynić pomijalnym poprzez zastosowanie niewielkich termometrów o znikomej pojemności cieplnej. Należy jednak pamiętać, że co do zasady termometr nie mierzy temperatury wyjściowej obiektu, a końcową temperaturę układu obiekt plus termometr.

Drugim ograniczeniem opisanej powyżej procedury pomiarowej jest to, że pozwala ona nam jedynie na ilościowe określenie różnicy między temperaturami, a nie jej absolutnej wartości. Rozszerzenie cieczy roboczej jest (idealnie) proporcjonalne do różnicy między temperaturą początkową termometru a temperaturą mierzonego ciała. Jednakże w praktyce używamy absolutnych wartości. Jest to zagwarantowane przyjęciem umownej wartości zera dla pewnego szczególnego rodzaju zjawisk. W wypadku skali Celsjusza, jak wiadomo, tą umowną wartością jest temperatura topnienia lodu. Związane z tym są oczywiście nowe problemy – skąd wiadomo, że lód zawsze topi się w tej samej temperaturze? W istocie wiemy, że tak nie jest – temperatura topnienia lodu zależy od wielu dodatkowych parametrów, takich jak ciśnienie powietrza czy zanieczyszczenie chemiczne lodu. Naukowcy włożyli wiele wysiłku w „oczyszczenie" definicji zera w skali Celsjusza z czynników, które mogłyby wpłynąć na poprawność takiej definicji. W rezultacie otrzymano skalę temperatury, w której można przypisać jednoznaczną wartość liczbową każdemu obiektowi materialnemu o dobrze określonych charakterystykach termodynamicznych. Jednak należy mieć na uwadze, że temperaturę można dowolnie przeskalować, stosując transformację liniową ($y = ax + b$); otrzymamy wtedy inną, równie dobrą skalę (np. Fahrenheita). Temperatura należy do rodzaju wielkości zwanych interwałowymi, w których tylko różnice mają sens fizyczny. Inaczej sprawa wygląda w wypadku takich wielkości jak masa, gdzie wartość zerowa jest „zafiksowana" i nie

[2] Lub też oddaje pewną ilość ciepła, jeśli wyjściowa temperatura obiektu była niższa.

może ulec przesunięciu. Takie wielkości nazywamy ilorazowymi, a jedyne dopuszczalne ich transformacje przeskalowania to mnożenie przez liczbę.[3]

Zajmijmy się teraz liczbową zależnością między ciepłem a temperaturą. Wiemy, że dostarczenie ciepła do danego ciała skutkuje zwiększeniem jego temperatury (o ile nie została wykonana dodatkowa praca mechaniczna). Jaka ilość ciepła ΔQ jest potrzebna do ogrzania np. pewnej ilości wody o dany interwał temperatury ΔT? Doświadczenie uczy nas, że ilość ta zależy od wielkości ciała – dokładniej od jego masy (a nie objętości). Wiemy też, że różne substancje mają różną zdolność do „akumulowania" ciepła. Ta sama ilość ciepła (np. pochodząca dokładnie z tego samego źródła) może podnieść temperaturę jednego grama danej substancji o jeden stopień Celsjusza, a innej substancji o półtora stopnia. Uwzględnienie tego faktu wymaga wprowadzenia pojęcia ciepła właściwego – odpowiedniego współczynnika przeliczeniowego różnego dla różnych substancji. Żeby uniknąć problemów z błędnym kołem, musimy przyjąć arbitralnie wartość takiego ciepła właściwego dla wybranej substancji, a następnie porównać zachowanie innych substancji podczas dostarczania tego samego ciepła i w ten sposób empirycznie wyznaczyć ich ciepła właściwe. Omówioną zależność ciepła od temperatury wyraża znany wzór:

$$\Delta Q = cm\Delta T,$$

gdzie c jest ciepłem właściwym, a m masą.

Przy pomocy tego prostego wzoru oraz korzystając z założenia, że ciepło całkowite układu zamkniętego nie może ulec zmianie, możemy rozwiązywać łatwe problemy kalorymetryczne, takie jak wyznaczenie temperatury końcowej mieszaniny dwóch substancji o danych temperaturach początkowych i ciepłach właściwych. Jednakże sytuacja staje się dużo ciekawsza, kiedy wprowadzimy do opisu pracę mechaniczną (energię).

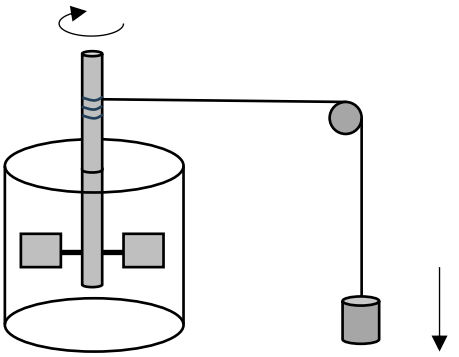
3.2. Ciepło a praca mechaniczna

Z doświadczenia wiemy, że ciepło można wytworzyć pracą mechaniczną. Potarcie rąk w zimny poranek trochę nas rozgrzeje, a długotrwałe stosowanie hamulca w samochodzie na pewno zwiększy temperaturę klocków hamulcowych, co może prowadzić do niebezpiecznego przegrzania płynu hamulcowego. Istnieje również możliwość zamiany odwrotnej, kiedy ciepło zostaje przetransformowane na pracę mechaniczną. Rozważmy porcję gazu zamkniętą ruchomym tłokiem. Podgrzewając ten gaz spowodujemy, że wskutek zwiększonego ciśnienia tłok zacznie się wysuwać, a gaz będzie zajmował coraz większą objętość. Warto zauważyć, że efekt końcowy takiego procesu jest podwójny – z jednej strony tłok wykonał pracę, która może być użytecznie wykorzystana, ale z drugiej strony gaz przyjął rozprężoną formę, która uniemożliwia dalsze wykorzystanie dostarczanego ciepła. To bardzo ważny fakt, który posłuży nam później do sformułowania drugiej zasady termodynamiki.

Problemem relacji między ciepłem a pracą mechaniczną zajął się dokładniej James Joule, który był nie tylko zdolnym eksperymentatorem, ale także cenionym browarnikiem. Już w szkole uczymy się, że Joule udowodnił równoważność ciepła i energii mechanicznej. Jest

[3] Należy tutaj dodać, że istnieje tzw. skala absolutna temperatury, w której również zero zostaje zafiksowane. Jest to związane z redukcyjną definicją temperatury jako miary średniej energii kinetycznej molekuł, o której będziemy mówić później. Zerowa temperatura odpowiada zerowej energii (brak ruchu), która jest absolutna.

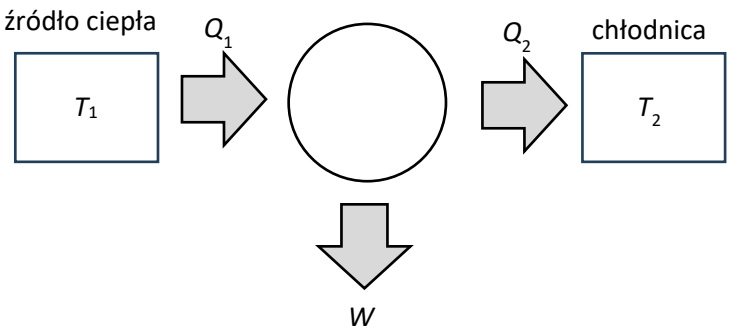
to jednak stwierdzenie trochę „na wyrost". Eksperymenty Joule'a umożliwiły porównanie ze sobą dwóch pozornie niezwiązanych wielkości: wykorzystanej energii mechanicznej i ciepła wytworzonego w rezultacie zastosowania tej energii. W urządzeniu Joule'a ciężarek zawieszony na lince połączony został z wirnikiem, którego łopatki obracały się w szczelnie zamkniętym naczyniu z cieczą (rys. 3.1). Zainstalowany termometr mierzył wzrost temperatury cieczy. Joule policzył energię uzyskaną w wyniku spadku ciężarka z pewnej wysokości, a następnie ilość ciepła wydzieloną w cieczy. Na tej podstawie wyznaczył przelicznik, umożliwiający zamianę energii mierzonej w jednostkach pracy mechanicznej (np. kg razy m) na ciepło mierzone np. w kaloriach. Tak przeprowadzony eksperyment nie dowodzi jednak tezy o równoważności energii i ciepła. Żaden wynik eksperymentu nie jest w stanie obalić naszej hipotezy (chyba żeby w cieczy nie wydzieliło się żadne ciepło). Jakakolwiek niezerowa wartość ciepła wydzielonego zostałaby zarejestrowana, ma to wpływ jedynie na liczbową wartość przelicznika.



Rys. 3.1. Schemat doświadczenia Joule'a

Czy zatem teza o równoważności pracy i ciepła nie ma empirycznej treści? Na szczęście tak nie jest. Empiryczna treść tej zasady ujawnia się, kiedy powtórzymy eksperyment Joule'a przy użyciu różnych substancji. *A priori* nie ma powodu uważać, że obliczona dla tych substancji ilość wydzielonego ciepła będzie w każdym wypadku taka sama (przy tej samej wysokości spadku ciężarka). Innymi słowy, mogłoby się zdarzyć, że współczynnik przeliczeniowy energii na ciepło byłby w wypadku każdej substancji inny. Wtedy teza o równoważności by upadła. Okazuje się jednak, że niezależnie od rodzaju substancji, materiału, z którego wykonane są łopatki wirnika, pory dnia itd., dana ilość energii mechanicznej zawsze przetwarzana jest na tę samą ilość ciepła. Daje to nam dobre indukcyjne uzasadnienie ścisłej korelacji (proporcjonalności) między tymi dwiema wielkościami, a ponieważ wybór jednostek jest sprawą umowną, możemy przyjąć, że nasze dwie wielkości – ciepło i praca (energia) mechaniczna – są równe.

Matematyczna równość nie oznacza możliwości dowolnego przekształcania jednej formy energii w drugą według życzenia. Zamiana energii mechanicznej na ciepło jest sprawą łatwą – można do tego celu wykorzystać urządzenie Joule'a. Spróbujmy jednak odzyskać wydatkowaną energię mechaniczną kosztem ciepła wydzielonego w naczyniu. Jak to zrobić? Trudno sobie wyobrazić sytuację, w której temperatura substancji w naczyniu zacznie opadać, łopatki wirnika zaczną się nagle obracać w przeciwnym kierunku, a ciężarek wzniesie się na początkową wysokość. Takich procesów nie obserwujemy w rzeczywistości.



Rys. 3.2. Schemat działania maszyny cieplnej

Pomysłowość ludzka nie zna jednak granic. Rewolucja przemysłowa osiemnastego i dziewiętnastego wieku stała się możliwa dzięki uporczywym próbom stworzenia machiny, przy pomocy której można byłoby przekształcić ciepło na użyteczną pracę w trybie ciągłym. Pierwszą udaną konstrukcją tego typu była oczywiście dobrze znana maszyna parowa, zastosowana zarówno do produkcji, jak i transportu. Równolegle z doskonaleniem technicznych szczegółów maszyn parowych trwały prace nad stworzeniem teoretycznych podstaw ich działania. W rezultacie powstał ogólny schemat działania każdego urządzenia zdolnego w ciągły, cykliczny sposób przetwarzać ciepło na pracę mechaniczną. Schemat ten zamieszczam powyżej (rys. 3.2). Podstawowymi elementami maszyny (zwanej silnikiem cieplnym) są: element roboczy (woda w wypadku maszyn parowych), źródło ciepła (kocioł) oraz chłodnica (otoczenie). W trakcie jednego cyklu element roboczy pobiera pewną ilość ciepła Q_1 ze źródła, a następnie wykonuje pracę W oraz oddaje niewielką ilość ciepła Q_2 do chłodnicy (np. w postaci gorącej pary wodnej wydostającej się do otoczenia). Następnie cykl ulega powtórzeniu. Możemy teraz napisać proste równanie, będące konsekwencją równoważności pracy i ciepła Joule'a:

$$Q_1 = W + Q_2.$$

Procent ciepła Q_1 zamieniony efektywnie na pracę dany jest równaniem:

$$\frac{W}{Q_1} = \frac{Q_1 - Q_2}{Q_1}.$$

Jest to tzw. sprawność silnika cieplnego. W wyniku wielu obserwacji i rozważań teoretycznych stwierdzono, że sprawność silnika nigdy nie może osiągnąć 100%; byłoby to możliwe tylko wtedy, gdyby żadna ilość ciepła nie została oddana do chłodnicy ($Q_2 = 0$). W rzeczywistości, maksymalna wartość sprawności możliwa do osiągnięcia jest dana wzorem:

$$\frac{T_1 - T_2}{T_1},$$

gdzie T_1 jest temperaturą źródła ciepła, a T_2 temperaturą chłodnicy (w skali absolutnej). Innymi słowy, niemożliwa jest całkowita zamiana ciepła na pracę mechaniczną w jednym cyklu roboczym, w którym element roboczy wraca do punktu wyjścia i jest gotowy do przyjęcia kolejnej porcji ciepła. Strata części ciepła na rzecz chłodnicy jest nieunikniona. Przy czym należy pamiętać, że nie jest to strata, którą można uczynić dowolnie małą. Na przykład

z doświadczenia wiemy, że w wyniku oporów czy tarcia każdy ruch na Ziemi ulega spowolnieniu. Jest to jednak efekt, który można uczynić niemal bliski zeru np. przez użycie substancji smarujących. Natomiast w wypadku ograniczenia sprawności silników cieplnych istnieje bariera nie do pokonania. Jest nią właśnie powyższy stosunek temperatur. Można powiedzieć, że praktyczne niedoskonałości maszyny obniżają jej sprawność poniżej tej wartości i z nimi możemy walczyć, natomiast nigdy nie osiągniemy wyniku lepszego niż ten stosunek. Zabrania nam tego jakaś fundamentalna prawidłowość charakteryzująca rzeczywistość fizyczną, o której będziemy mówić szerzej w następnym paragrafie.

3.3. Dwie zasady termodynamiki

Termodynamika jest częścią fizyki, która syntetycznie ujmuje wszystkie opisane powyżej zjawiska (i dużo więcej). Jak każda fundamentalna teoria, formułuje kilka podstawowych praw zwanych zasadami. Pierwsza zasada termodynamiki to nic innego jak uogólniona zasada zachowania energii, która bierze pod uwagę fakt, stwierdzony przez Joule'a, że ciepło jest też formą energii. Można tę zasadę przedstawić w formie następującego prostego równania:

$$Q + W = U,$$

gdzie Q jest ciepłem dostarczonym do danego ciała (odizolowanego od otoczenia), W – pracą wykonaną nad tym ciałem, a U – zmianą jego energii wewnętrznej. Przyjmujemy ponadto konwencję, że wartości ujemne pracy i ciepła oznaczają odpowiednio pracę wykonaną przez ciało i ciepło przez nie oddane. Równanie powyższe można zastosować do różnych możliwych sytuacji. Na przykład gdy $U = 0$, czyli energia wewnętrzna ciała nie ulega zmianie, cała praca wykonana nad ciałem zostaje oddana przez ciało w formie ciepła: $Q = -W$. Z kolei gdy U jest ujemne, zasada implikuje, że ciało może oddać pewną ilość ciepła oraz wykonać pracę kosztem energii wewnętrznej. Zauważmy, że pierwsza zasada termodynamiki dopuszcza sytuację, w której całe ciepło dostarczone ciału zostanie zamienione na pracę mechaniczną bez zmiany energii wewnętrznej: $W = -Q$. Jednakże proces taki, jak się za chwilę przekonamy, podlega ograniczeniom sformułowanym w drugiej zasadzie termodynamiki.

Druga zasada termodynamiki posiada kilka wersji, które wydają się niezależne, lecz okazują się wzajemnie równoważne. Wszystkie mają postać reguły zabraniającej zachodzenia pewnego procesu. Sformułowanie znane pod nazwą zasady Kelvina opiera się na stwierdzonym powyżej fakcie nieistnienia silników cieplnych o stuprocentowej sprawności:

> Nie istnieje proces fizyczny, którego *jedynym* rezultatem byłaby całkowita zamiana pewnej ilości ciepła na pracę mechaniczną.

Przede wszystkim musimy podkreślić wyrażenie „jedyny rezultat" występujące w powyższym sformułowaniu. Bez tego dodatku sformułowana teza byłaby jawnie fałszywa, jak pokazuje przykład z poprzedniego paragrafu z rozprężającym się gazem. Możliwe jest przekształcenie całego ciepła dostarczonego próbce gazu w pracę, pod warunkiem, że stan końcowy gazu będzie się różnił od początkowego – gaz się rozpręży. Natomiast w silnikach cieplnych chodzi o to, żeby urządzenie po wykonaniu pracy wróciło dokładnie do stanu początkowego. W takiej sytuacji druga zasada termodynamiki zabrania całkowitego wykorzystania ciepła – pewna jego część musi zostać przekazana otoczeniu.

Sformułowanie Clausiusa drugiej zasady dotyczy procesu przekazywania ciepła między ciałami o różnych temperaturach. Wiemy z doświadczenia, że ciała o wyższej temperaturze mogą przekazać ciepło ciałom o temperaturze niższej. Czy możliwa jest sytuacja odwrotna? Zasada w wersji Clausiusa potwierdza to, co znamy dobrze z codziennego doświadczenia:

> Nie istnieje proces fizyczny, którego *jedynym* rezultatem byłoby przekazanie ciepła od ciała o niższej temperaturze do ciała o wyższej temperaturze.

Znów niezmiernie istotne jest podkreślenie, że chodzi o proces, którego jedynym rezultatem jest owo przekazanie ciepła. W przeciwnym razie zasada Clausiusa wykluczałaby istnienia lodówek, które przecież odbierają ciepło od przedmiotów chłodnych i przekazują je goręt-szemu otoczeniu. Dzieje się to jednak kosztem dostarczonej energii, która musi mieć gdzieś swoje źródło. Druga zasada w wersji Clausiusa wyklucza jedynie „spontaniczny" przepływ ciepła z ciał zimnych do gorących, bez interwencji z zewnątrz.

Wersje Kelvina i Clausiusa wyglądają różnie, ale można udowodnić, że są one równo-ważne. Dla ilustracji pokażemy, jak w dość prosty sposób wyprowadzić pierwszą wersję z drugiej. Załóżmy zatem, że pierwsza wersja jest fałszywa, czyli istnieje sposób na to, aby przekształcić całe ciepło w pracę mechaniczną bez żadnej dodatkowej zmiany w układzie fizycznym. Można łatwo się przekonać, że umożliwiłoby to przekazanie ciepła z ciała zim-nego do gorącego. Weźmy bowiem ciało zimniejsze i zastosujmy do niego hipotetyczną stu-procentowo sprawną maszynę cieplną. Odbierze ona pewną ilość ciepła z tego ciała i zamieni ją bez strat na pracę mechaniczną, np. ruch posuwisty tłoka. Całkowita zamiana uzyskanej ilości pracy mechanicznej na ciepło jest już sprawą prostą – wystarczy np. wykorzystać do tego celu odpowiednio zmodyfikowane urządzenie Joule'a. W ten sposób uzyskaliśmy efekt końcowy w postaci oziębienia ciała chłodnego i ocieplenia ciała gorącego bez żadnych do-datkowych zmian w otoczeniu. Sformułowanie Clausiusa drugiej zasady termodynamiki oka-zuje się fałszywe, co pokazuje, że wersja Clausiusa implikuje wersję Kelvina. Można również pokazać, że założenie fałszywości zasady w wersji Clausiusa pociąga za sobą fałsz sformu-łowania Kelvina, ale jest to trochę bardziej skomplikowane. W każdym razie obie zasady, choć brzmiące odmiennie, są sobie równoważne.

3.4. Druga zasada termodynamiki i entropia

Powyższe sformułowania drugiej zasady termodynamiki nie są powszechnie znane, choć zapewne wielu czytelników zetknęło się z nimi wcześniej. Natomiast jestem pewien, że wszyscy słyszeli o trzecim, najsłynniejszym sformułowaniu:

> Entropia izolowanego układu nigdy nie maleje (rośnie lub pozostaje niezmienna).

Wersja ta wprowadza pojęcie entropii, któremu będziemy się musieli nieco bliżej przyj-rzeć. Często spotyka się formułkę, że entropia jest miarą nieuporządkowania, a zatem druga zasada termodynamiki głosi, że nieuporządkowanie odosobnionego układu rośnie lub co naj-wyżej pozostaje takie samo (kiedy osiągnie maksymalną wartość). Jest to oczywiście prawda, ale nieuporządkowanie jest pojęciem statystycznym, a nie termodynamicznym. Statystyczną definicją entropii zajmiemy się później, natomiast na razie wprowadzimy trochę mniej znane pojęcie entropii termodynamicznej.

Wróćmy na chwilę do naszej analizy silników cieplnych. Z przedstawionych powyżej rozważań wynika, że ilość ciepła dostarczona do danego silnika nie jest jedynym parametrem

determinującym uzyskaną pracę mechaniczną. Sprawność silnika w przekształcaniu ciepła na pracę zależy od temperatury, a dokładniej od różnicy temperatur między źródłem ciepła a chłodnicą. Jeśli różnica ta jest niewielka, to tylko drobny ułamek dostarczonego ciepła będzie mógł być efektywnie wykorzystany w formie pracy mechanicznej. Oznacza to, że ciepło może występować w różnych „formach", mniej lub bardziej użytecznych. Ilość ciepła (czy też, bardziej poprawnie, energii wewnętrznej) zawarta w garnku z wrzącą wodą jest dużo bardziej użyteczna niż ta sama ilość ciepła w wannie z letnią wodą. Entropia w sensie termodynamicznym jest właśnie wielkością, która jest powiązana z użytecznością danego układu przy przekształcaniu ciepła w pracę. Ściślej rzecz ujmując, entropia jest odwrotnością użyteczności – im większa entropia, tym mniej użyteczny jest system i zawarte w nim ciepło.

Nie powinno być zatem zaskoczeniem, że formuła wyrażająca entropię termodynamiczną powinna uwzględniać temperaturę w taki sposób, by większa temperatura implikowała mniejszą entropię. Dokładna definicja entropii jest nieco skomplikowana, ale warto włożyć nieco wysiłku w jej zrozumienie. Na wstępie podkreślmy, że nie będziemy definiowali entropii jako takiej, a tylko różnicę entropii między dwoma stanami danego układu. Zatem entropia termodynamiczna jest wielkością interwałową, podobnie jak temperatura. Ustalenie np. wartości zerowej entropii jest kwestią umowną – możemy to zrobić, ale nie musimy. Wybierzmy zatem dwa stany A i B danego ciała fizycznego (np. gazu). Stany A i B mogą się różnić wartością parametrów makroskopowych – ciśnienia, temperatury, objętości, gęstości. Istnieje nieskończenie wiele sposobów „dojścia" ze stanu A do stanu B. Spośród tej mnogości dróg, wybierzmy te, które reprezentują pewien odwracalny proces, czyli taki, który może zachodzić w „obie strony". Jak rozpoznać, że dany proces jest odwracalny? Użyteczną regułą jest tutaj warunek, że proces taki musi być powolny. Na przykład usunięcie ścianki w naczyniu zawierającym gaz pod ciśnieniem spowoduje gwałtowne rozprężenie gazu, które na pewno nie jest odwracalne – gaz spontanicznie sam nie „wepchnie się" do naczynia, w którym się uprzednio znajdował. Natomiast jeśli do naczynia podłączymy tłok i będziemy go wolno wyciągać, taki proces można łatwo odwrócić poprzez zadziałanie odwrotną siłą na tłok i sprężenie gazu.

Rozważmy zatem taki wolny, kontrolowany i odwracalny proces, który zaczyna się od stanu A i kończy w stanie B. Podzielmy ten proces na drobne fragmenty w taki sposób, aby w każdym z nich temperatura ciała była w przybliżeniu stała. Dla każdej takiej części obliczmy iloraz ciepła pobranego przez układ ΔQ oraz temperatury ciała T: $\frac{\Delta Q}{T}$. Sumując te ilorazy po wszystkich częściach drogi od A do B otrzymamy różnicę entropii między stanem końcowym B a stanem początkowym A. Oznaczymy ją przez S_{AB}. Mamy zatem następujący wzór:

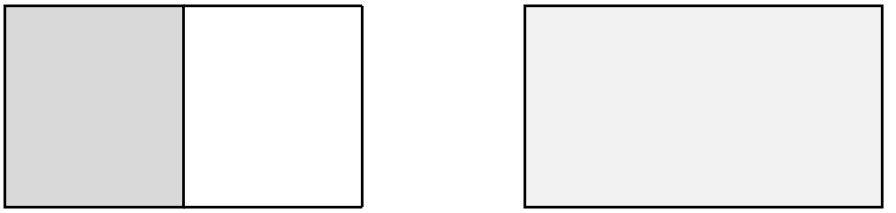
$$S_{AB} = \sum_i \frac{\Delta Q_i}{T_i},$$

gdzie indeks i reprezentuje poszczególne fragmenty, na które podzieliliśmy całą drogę od A do B. Wybierając coraz to mniejsze fragmenty, w granicy otrzymamy dokładną wartość entropii, która jest wyrażana matematyczną operacją całkowania:

$$S_{AB} = \int_A^B \frac{dQ}{T}.$$

Chociaż definicja entropii odwołuje się do konkretnej drogi (procesu) od stanu A do B. to jednak można pokazać, że rezultat jest niezależny od szczegółów tej drogi, jeśli tylko jest ona odwracalna. Innymi słowy, wielkość S_{AB} jest taka sama dla każdej odwracalnej drogi dojścia od A do B.[4] Możemy więc przyjąć, że S_{AB} charakteryzuje wyłącznie stan początkowy i końcowy.

Posłużmy się prostym przykładem dla zilustrowania wprowadzonego pojęcia entropii termodynamicznej. Rozważmy gaz zamknięty w jednej części naczynia, którego druga część jest pusta. Po wyjęciu przegrody, gaz spontanicznie się rozszerzy, zajmując całą objętość naczynia (rys. 3.3). Jaka będzie różnica entropii między stanem końcowym a początkowym? Na pozór może się wydawać, że zmiana entropii jest zerowa, gdyż żadne ciepło nie jest dostarczane ani pobierane. (Temperatura gazu również nie ulega zmianie – wynika to z prostego faktu, że usunięcie bariery nie ma żadnego wpływu mechanicznego na ruch molekuł, a zatem ich średnia energia kinetyczna pozostaje niezmieniona.) Jednak jest to błędny wniosek. Jak już zauważyliśmy, spontaniczne rozprężanie gazu nie jest procesem odwracalnym. Aby poprawnie obliczyć zmianę entropii, musimy znaleźć inną, odwracalną drogę dojścia od stanu początkowego do końcowego.



Rys. 3.3. Gaz zamknięty w połowie naczynia (po lewej) ma mniejszą entropię niż gaz rozprężony w całej objętości (po prawej)

Jedna z możliwości jest następująca. Zamiast bariery wprowadźmy ruchomy tłok i zacznijmy wolno go wyciągać, zwiększając w ten sposób objętość gazu. Z doświadczenia wiemy, że taki proces skutkuje ochłodzeniem gazu (można to uzasadnić, korzystając z molekularnej teorii gazu – cząsteczki gazu zderzają się z wysuwającym się tłokiem i tracą część swojego pędu). Aby zachować stałość temperatury w stanie końcowym, musimy zetknąć nasz gaz ze źródłem ciepła, które uzupełni braki. Oznacza to jednak, że w ciągu całego procesu ilość ciepła dostarczonego do układu ΔQ będzie niezerowa (dodatnia). Dzieląc tę wartość przez stałą wielkość temperatury T, otrzymamy dodatnią wartość różnicy entropii S_{AB}. Zatem w wyniku spontanicznego rozprężenia gazu jego entropia rośnie.[5]

[4] Musimy jeszcze oczywiście założyć, że dla każdych dwóch stanów istnieje odwracalny proces je łączący.

[5] Ściśle rzecz biorąc, udowodniliśmy, że entropia gazu zajmującego większą objętość przy stałej temperaturze jest większa od entropii tego samego gazu w stanie sprężonym, niezależnie od drogi dojścia. Pamiętajmy, że entropia charakteryzuje stany, a nie drogi dojścia (wykorzystanie w definicji odwracalnych procesów dojścia jest tylko użytecznym narzędziem, które może zostać na końcu odrzucone, jak Wittgensteinowska drabina). Zwykle jednak rozszerza się pojęcie entropii na „typowe" procesy prowadzące od jednego stanu do drugiego i mówi nieformalnie o zmianie entropii procesu.

Skąd wiadomo, że druga zasada termodynamiki w wersji z entropią jest równoważna dwóm pozostałym sformułowaniom (Kelvina i Clausiusa)? Znowu nie będziemy tego udowadniać ogólnie, ale pokażemy, że zasada wzrostu entropii implikuje wersję Clausiusa (a zatem także i Kelvina, skoro poprzednio udowodniliśmy, że „Clausius" implikuje „Kelvina"). Załóżmy więc, wbrew temu, co głosi wersja Clausiusa, że możliwe jest przekazanie ciepła z ciała zimniejszego do gorącego bez żadnych dodatkowych zmian. W istocie pociąga to tezę, że możliwy jest proces, w którym dwa układy o takiej samej temperaturze zaczną spontanicznie zmieniać swoją temperaturę – temperatura jednego z nich wzrośnie, a drugiego zmaleje.[6] Jaka jest różnica entropii między stanem układu w „równowadze termicznej" a stanem, w którym jedna część układu ma temperaturę wyższą niż druga? Nietrudno pokazać, że entropia tego pierwszego stanu (równowagi) będzie wyższa od entropii stanu drugiego (nierównowagi).

Aby to udowodnić, rozważmy dwa ciała o temperaturach odpowiednio T_1 i T_2, gdzie $T_1 < T_2$. Po ich zetknięciu temperatura końcowa będzie równa T_3, przy czym $T_1 < T_3 < T_2$. Jak można „zasymulować" ten proces wyrównywania temperatur przy pomocy odwracalnej procedury? Aby w kontrolowany sposób ochłodzić ciało o wyższej temperaturze T_2, można np. zetknąć je z próbką gazu zamkniętą w naczyniu z tłokiem i wolno wyciągać tłok. Temperatura gazu będzie opadać, zatem zacznie on pobierać ciepło od naszego ciała, aż do momentu, kiedy osiągnie ono żądaną temperaturę T_3. W podobny sposób postąpimy z ciałem chłodniejszym – zetkniemy je z gazem w naczyniu, ale tym razem zaczniemy wpychać tłok. Wydzielone w ten sposób ciepło ogrzeje nasze ciało do temperatury T_3. Zmiana entropii w przypadku ciała zimniejszego będzie dodatnia, gdyż ciepło zostaje pobrane, natomiast zmiana entropii ciała gorącego jest ujemna – ciepło jest oddawane do otoczenia. Entropia całego układu to suma entropii poszczególnych części (entropia jest wielkością addytywną, jak np. masa). Jaki jednak będzie znak sumy liczby dodatniej i liczby ujemnej? Zależy to oczywiście od ich wartości bezwzględnych. Zauważmy, że obliczając np. zmianę entropii ciała gorętszego, na każdym kroku musimy dzielić dostarczone ciepło przez liczbę z zakresu od T_3 do T_2. Natomiast w wypadku ciała zimnego mianownikiem jest liczba pomiędzy T_1 and T_3. Ponieważ w obu wypadkach całkowita ilość ciepła pobranego i oddanego jest taka sama, wartość bezwzględna zmiany entropii dla ciała gorętszego będzie niższa od wartości bezwzględnej zmiany entropii ciała chłodniejszego. Zatem ostatecznie zmiana entropii przy wyrównywaniu temperatur jest wielkością dodatnią – stan równowagi termicznej charakteryzuje się wyższą entropią termodynamiczną niż stan nierównowagi.[7]

Jednak, jak już wskazaliśmy, gdyby zasada w wersji Clausiusa była złamana, możliwy byłby spontaniczny proces prowadzący od stanu równowagi o wyższej entropii do stanu nierównowagi o entropii niższej. W takim wypadku również zasada niemalenia entropii zostałaby naruszona. Dowodzi to, że zasada entropijna implikuje zasadę w sformułowaniu Clausiusa.[8] Druga zasada termodynamiki w wersji wykorzystującej entropię jest jawnie niesyme-

[6] Spróbujcie sami uzasadnić, że jeśli zasada Clausiusa jest złamana, taki proces jest jak najbardziej możliwy.

[7] Ciekawe, że w tym wypadku do podobnego wniosku doszlibyśmy nawet rozważając zasadniczo niepoprawną, nieodwracalną drogę dojścia od stanu początkowego do końcowego. Jednakże wartość liczbowa tak obliczonej „entropii" różniłaby się od prawdziwej wartości.

[8] Zauważmy, że udowodniliśmy już dwie implikacje: Entropia → Clausius i Clausius → Kelvin. Aby udowodnić równoważność trzech sformułowań, wystarczy pokazać, że zachodzi trzecia implikacja, która „domyka" cały cykl: Kelvin → Entropia.

tryczna w czasie – procesy dopuszczalne w jedną stronę są wykluczone przy odwróceniu strzałki czasu. Dokładniejszą analizę tego fenomenu natury zostawimy na później, natomiast teraz przyjrzymy się bliżej interpretacji termodynamiki w kategoriach molekularnej teorii materii.

Przy okazji analizy treści drugiej zasady termodynamiki możemy zadać następujące pytanie o charakterze bardziej filozoficznym: czy świat, w którym obowiązywałaby zasada głosząca, że we wszystkich izolowanych układach fizycznych entropia maleje lub pozostaje niezmienna, byłby światem zasadniczo różnym od naszego? Na pozór wydaje się, że tak. Jednak pojęcia „rośnięcia" i „malenia" *implicite* zakładają, że czas ma już pewien wyróżniony kierunek od wartości mniejszych do większych. Dana funkcja czasu F (np. funkcja entropii) jest rosnąca, gdy dla dowolnych wartości czasu $t_1 < t_2$, $F(t_1) < F(t_2)$. Skąd jednak wiemy, że czas t_1 jest „wcześniejszy" od t_2 i co to w ogóle znaczy? Istnieje hipoteza, zgodnie z którą uporządkowanie czasowe momentów jest wtórne względem uporządkowania entropii: moment t_1 uznamy za wcześniejszy od t_2, gdy entropia wszechświata w t_1 jest mniejsza od entropii w t_2. W takim jednak razie opisanie świata, w którym entropia zawsze maleje, jest niepoprawne: entropia z definicji musi rosnąć. Jeśli wydaje nam się, że entropia maleje, to tylko dlatego, że odwróciliśmy uporządkowanie czasu. Pojawia się oczywiście tutaj nasz dobry znajomy – problem sensu empirycznego drugiej zasady termodynamiki. Czy zdefiniowanie porządku czasowego przy pomocy globalnego wzrostu entropii sprowadza drugą zasadę do poziomu konwencji (zdania analitycznego *a priori*)? Sprawa nie jest oczywista. Jak się wydaje, możliwy jest świat, w którym zasada ta byłaby ewidentnie złamana. Byłby to świat, w którym występowałyby zarówno izolowane procesy zwiększające, jak i zmniejszające entropię. Natomiast świat, w którym entropia zawsze maleje, to po prostu nasz świat, tylko błędnie opisany, w którym zamieniliśmy przyszłość z przeszłością.

3.5. Redukcja termodynamiki do mechaniki

Idea, że cała materia zbudowana jest z małych, niewidocznych i niepodzielnych składników, zakiełkowała w umysłach ludzkich bardzo wcześnie. Starożytni atomiści greccy i rzymscy (Demokryt, Leucyp) opierali swoją hipotezę na obserwacjach pyłków unoszących się w powietrzu, jednakże zasadnicza część ich doktryny atomistycznej była czystą spekulacją bez potwierdzenia empirycznego. Należy mimo wszystko podkreślić, że nawet w naiwnej wersji teoria atomistyczna umożliwiała wyjaśnienie szerokiego spektrum zjawisk, od zmian stanów skupienia ciał do aktów percepcji. Ta zdolność atomizmu do syntetycznego ujmowania i wyjaśniania wielości obserwowanych zjawisk stanowiła mocny impuls do rozwinięcia jego współczesnej wersji. Jak się za chwilę przekonamy, atomistyczny model budowy ciał daje nam możliwość ujrzenia procesów termodynamicznych w zupełnie nowym świetle.

Podstawowe założenie atomistycznej (czy też lepiej molekularnej) teorii budowy ciał ma oczywiście charakter ontologiczny (substancjalny). Stwierdza ono coś na temat struktury obiektów materialnych – to mianowicie, że nie można ich dzielić w nieskończoność. U podstawy każdego ciała fizycznego leży ogromna liczba identycznych elementarnych składników – molekuł lub pojedynczych atomów. Zatem mamy tu do czynienia z redukcją ontologiczną poprzez utożsamienie przedmiotów materialnych ze zbiorowiskiem cząstek. Następnym krokiem procedury redukcyjnej może być próba objaśnienia obserwowalnych czy mie-

rzalnych własności ciał w kategoriach własności grup molekuł tworzących te ciała. Zauważmy jednak, że samo założenie identyczności ciał ze zbiorowiskami molekuł nie gwarantuje powodzenia takiej redukcji. Do pomyślenia jest, że w wyniku odpowiedniego połączenia podstawowych elementów powstanie struktura, której cechy nie dadzą się zredukować do odpowiedniej kombinacji własności składników. Takie nieredukowalne własności kompleksów nazywa się własnościami emergentnymi. W historii nauki popularnością cieszyła się np. koncepcja, zgodnie z którą procesy życiowe nie mogą być całkowicie zredukowane do własności fizykochemicznych molekuł wchodzących w skład danego organizmu. Wraz ze wzrostem złożoności obiektów fizycznych, od prostych atomów, molekuł (np. DNA), komórek, do całych organizmów, pojawiają się coraz to nowe charakterystyki, których istnienia nie można wydedukować z własności elementów składowych.

Jednakże nurt redukcjonizmu w fizyce był zawsze bardzo mocny. Spektakularnym tego przykładem jest właśnie nauka o cieple. Przyjęcie założenia o molekularnej budowie materii było dla naukowców punktem wyjścia do interpretacji własności i procesów termodynamicznych w kategoriach mechanicznych oddziaływań między cząsteczkami wchodzącymi w skład danej substancji. Wraz z odrzuceniem substancjalnej koncepcji ciepła w formie cieplika, porzucona została również idea emergencji własności cieplnych ciał. Wysiłki naukowców zostały skoncentrowane na przedefiniowaniu znanych pojęć termodynamicznych przy pomocy pojęć odnoszących się do kolekcji cząsteczek i ich mechanicznych zachowań. Zacznijmy od fundamentalnych pojęć ciepła i temperatury. Jak powszechnie wiadomo, obie te wielkości uzyskują interpretację w kategoriach dobrze znanej z mechaniki klasycznej energii kinetycznej ruchu cząsteczek. Jednym z fundamentalnych założeń molekularnej teorii materii głosi, że cząsteczki tworzące daną substancję znajdują się w ciągłym ruchu, a zatem ich energia kinetyczna jest niezerowa. Całkowita energia kinetyczna wszystkich cząsteczek jest miarą ciepła – lub też, bardziej poprawnie, ciepło pobrane przez dane ciało jest równe wzrostowi całkowitej energii kinetycznej ruchu cząsteczek. Z kolei temperaturę ciała definiuje się jako uśrednioną energię kinetyczną. Ze względu na chaotyczny charakter wewnętrznego ruchu cząstek, różne cząsteczki będą posiadały różną energię kinetyczną (mówimy w takim wypadku o statystycznym rozkładzie prędkości, a zatem także energii kinetycznej). Uśredniając ten rozkład, otrzymamy absolutną wartość temperatury danego ciała. Dzięki przyjętej definicji możemy w niearbitralny sposób ustalić zerową wartość temperatury, która odpowiada zanikowi ruchu (z powodu efektów kwantowych taki zanik jest fizycznie niemożliwy, ale możemy się do niego zbliżać).

Zwróćmy uwagę na jeden ważny aspekt proponowanej definicji temperatury – mianowicie jej statystyczny charakter. Jak się okazuje, redukcja pojęć termodynamicznych do czysto mechanicznych nie jest wykonalna z dość oczywistych powodów – nie jesteśmy w stanie „śledzić" ruchu wszystkich pojedynczych cząsteczek. Dlatego też musimy odwoływać się do statystyki, która operuje pojęciami średnich, rozkładów, prawdopodobieństwa itd. Teoria, do której dokonamy redukcji nauki o cieple, nosi nazwę mechaniki statystycznej. Jest to w istocie nowa teoria w stosunku do mechaniki newtonowskiej. Zawiera ona oczywiście dobrze znane elementy tej ostatniej, natomiast wprowadza również zupełnie nowe pojęcia. Co więcej, mechanika statystyczna przyjmuje pewne dodatkowe założenia, nierzadko w ukryty sposób, co do których można mieć wątpliwości, czy są one uzasadnione. Do tej sprawy wrócimy jeszcze w dalszej części rozdziału.

Proponowane definicje temperatury i ciepła łatwo wyjaśniają relację między tymi dwoma pojęciami, opisaną w pierwszym paragrafie niniejszego rozdziału. Ponieważ masa danego

ciała jest proporcjonalna do liczby jego cząsteczek, iloczyn masy ciała i średniej energii kinetycznej jego molekuł daje nam miarę całkowitej energii, czyli ciepła (pamiętajmy, że średnia energia to suma wszystkich energii cząstek podzielona przez ich liczbę). Ten prosty przykład ilustruje kolejny ważny etap każdej udanej redukcji jednej teorii do drugiej. Po zdefiniowaniu podstawowych pojęć redukowanej teorii w kategoriach teorii podstawowej, następnym krokiem jest wyprowadzenie praw teorii redukowanej z praw teorii redukującej, przy zastosowaniu wprowadzonych definicji. Idealnie powinno być tak, że dane prawo teorii redukowanej, po zastąpieniu wszystkich występujących w nim terminów przez ich definicyjne odpowiedniki w teorii redukującej, powinno przejść w prawo tej podstawowej teorii lub przynajmniej w logiczną konsekwencję takich praw. W praktyce sprawa nieraz się komplikuje, ale zasadniczy cel jest taki, jak to opisałem.

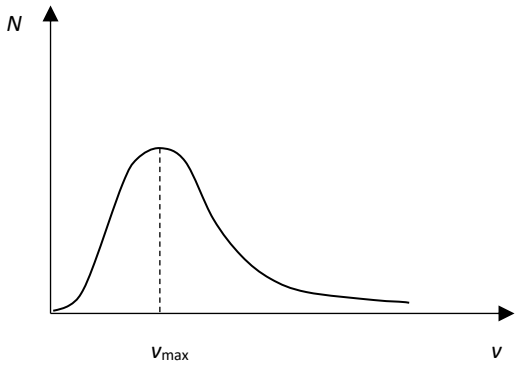
Postępując w naszkicowany powyżej sposób, możemy również zredukować pierwszą zasadę termodynamiki do prawa mechaniki klasycznej, jakim jest zasada zachowania energii. Skoro przekazane ciepło jest energią kinetyczną dostarczoną ciału, z zasady zachowania energii wynika automatycznie, że energia kinetyczna plus praca wykonana nad ciałem musi dać rezultat w postaci zwiększenia energii wewnętrznej ciała. Ta redukcja nie wymaga zastosowania żadnych pojęć *stricte* statystycznych. Natomiast aby wyprowadzić pozostałe prawa termodynamiki, jak np. prawa rządzące zachowaniem gazów doskonałych (równanie stanu gazu doskonałego, łączące temperaturę, objętość i ciśnienie), potrzebujemy dokonać redukcji innego pojęcia, jakim jest ciśnienie. Ciśnienie wywierane przez gaz na ścianki naczynia jest rezultatem bombardowania ścianek przez cząsteczki gazu. W wyniku zderzenia ze ścianką każda cząsteczka zmienia swój pęd, oddając jego część. Zmiana pędu podzielona przez czas trwania oddziaływania daje nam siłę, z jaką cząsteczka oddziałuje na ściankę (wiemy to z drugiej zasady dynamiki Newtona). Z kolei podzielenie siły na jednostkę powierzchni to właśnie ciśnienie. Oczywiście całkowite ciśnienie będzie sumą wszystkich zmian pędów cząsteczek uderzających w ścianki na jednostkę czasu i jednostkę powierzchni. Z dodatkowego założenia, że żaden kierunek ruchu cząstek nie jest wyróżniony (tj. średnia liczba cząstek i ich prędkość w każdym kierunku jest taka sama), wynika od razu prawo Pascala, głoszące, że ciśnienie wywierane na ścianki naczynia jest we wszystkich kierunkach jednakowe.

Korzystając ze statystycznych definicji temperatury i ciśnienia, możemy wyprowadzić dobrze znane z termodynamiki równanie stanu gazu doskonałego. Być może pamiętacie to wyprowadzenie ze szkolnego kursu fizyki, więc poprzestanę na ogólnym szkicu, aby uchwycić istotę podejścia statystycznego. Rozważamy gaz zamknięty w pojemniku o kształcie sześcianu o boku l (jest to założenie upraszczające rozumowanie). Zakładamy, że zderzenia cząsteczek ze ściankami naczynia są doskonale sprężyste, czyli cząsteczka o składowej prędkości v_x prostopadłej do danej ścianki zmieni tę składową na przeciwną o tej samej wartości. Daje to zmianę pędu równą $2mv_x$. Zakładając, że v_x jest uśrednioną wartością, możemy wnioskować, że średni czas między dwoma zderzeniami ze ścianką wynosi $\frac{2l}{v_x}$, czyli liczba zderzeń na jednostkę czasu to $\frac{v_x}{2l}$. Zatem całkowita zmiana pędu w kierunku x na jednostkę czasu jest równa zmianie pędu w jednym zderzeniu razy liczba zderzeń, a więc wynosi $\frac{mv_x^2}{l}$. Ale zmiana pędu na jednostkę czasu to nic innego jak siła, która podzielona przez pole powierzchni da nam ciśnienie wywierane przez jedną cząstkę na jedną ściankę naczynia. Jeżeli pomnożymy to przez liczbę cząstek N, otrzymamy wyrażenie na ciśnienie: $p = \frac{Nmv_x^2}{l^3}$. Biorąc pod uwagę

wzór na energię kinetyczną oraz przyjmując założenie, że średnia prędkość w każdym kierunku jest jednakowa, a także zastępując l^3 przez objętość naczynia V, uzyskamy:

$$pV = \frac{2}{3}N\langle E_{\mathrm{kin}}\rangle,$$

gdzie $\langle E_{\mathrm{kin}}\rangle$ jest średnią energią kinetyczną. Ponieważ średnia energia nie jest niczym innym, jak miarą temperatury, otrzymujemy ostatecznie zależność funkcyjną między ciśnieniem, temperaturą i objętością znaną z doświadczenia.

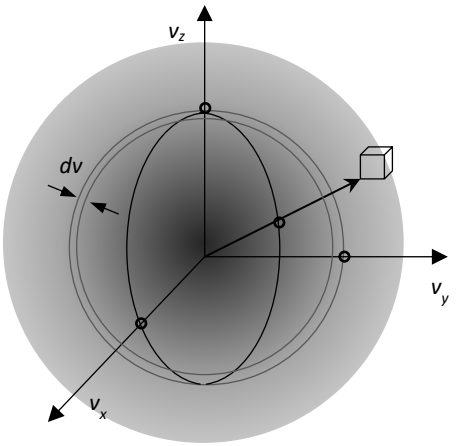


Rys. 3.4. Rozkład prędkości Maxwella. Rozkład ten określa liczbę cząstek posiadających daną prędkość *v* w dowolnym kierunku

Widać zatem, że do wyprowadzenia makroskopowego prawa potrzebowaliśmy zarówno zasad mechaniki klasycznej (druga zasada dynamiki, zależność energii od prędkości etc.), jak i założeń o charakterze statystycznym (ta sama średnia prędkość cząstek w każdym kierunku). Jak już powiedzieliśmy, nie jesteśmy w stanie określić prędkości i położeń pojedynczych cząstek z prostego powodu – nie możemy nawet zidentyfikować indywidualnych cząstek ze względu na ich zbyt małe rozmiary. Możemy jednak wprowadzić opis cząstek, który będzie zawierał więcej informacji niż same średnie wartości danych wielkości (energii, prędkości). Taki opis może obejmować statystyczny rozkład danej wielkości w całej populacji, czyli informację, jaka liczba cząstek posiada daną wartość rozważanej wielkości. Statystyczne rozkłady stosuje się powszechnie przy analizie populacji ludzkich – np. można mówić o rozkładzie indywidualnych dochodów (ile osób uzyskuje dochody w danym przedziale). W mechanice statystycznej na plan pierwszy wysuwa się kwestia rozkładu prędkości (a co za tym idzie energii). Pytanie, jakie postawili sobie badacze, było następujące. Rozważmy gaz znajdujący się w stanie równowagi termodynamicznej tj. taki, którego parametry (ciśnienie, temperatura, objętość) nie ulegają zmianie. Jaka funkcja będzie opisywała liczbę cząstek w zależności od ich prędkości?

Zagadnienie to rozwiązali niezależnie James Clerk Maxwell i Ludwig Boltzmann. Korzystając wyłącznie z ogólnych założeń, takich jak anizotropia próbki gazu (średnie prędkości w każdym kierunku są jednakowe), wyprowadzili oni formułę określającą, jaki ułamek całkowitej liczby cząstek *N* będzie charakteryzował się prędkościami z niewielkiego (infinitezymalnego) przedziału od wartości *v* do *v* + *dv*. Formuła ta nosi nazwę rozkładu Maxwella (lub też Maxwella-Boltzmanna). Funkcja *f*(*v*) rozkładu Maxwella przyjmuje wartość zero dla

prędkości $v = 0$, następnie rośnie do wartości maksymalnej dla pewnej prędkości v_{max} i zaczyna maleć do zera w nieskończoności (rys. 3.4). Można stąd wyprowadzić wniosek, że w danej próbce gazu liczba cząstek o prędkości bliskiej zerowej zbiega do zera (jednakże porównajcie wpis w ramce). Prędkości znacznie przewyższające wartość v_{max} mogą się zdarzyć, ale są statystycznie bardzo mało prawdopodobne.



Rys. 3.5. Przestrzeń prędkości z zaznaczeniem obszaru o danej prędkości wektorowej (sześcian) i o danej prędkości skalarnej w dowolnym kierunku (powłoka kuli)

Problem statystycznego rozkładu prędkości nieco się komplikuje, kiedy weźmiemy pod uwagę, że prędkość jest wektorem, a nie skalarem. Aby rozważyć tę kwestię trochę bardziej precyzyjnie, wprowadźmy abstrakcyjną trójwymiarową przestrzeń prędkości, w której każdy punkt reprezentuje wektor prędkości z początkiem umieszczonym w początku układu odniesienia (rys. 3.5). Rozkład statystyczny prędkości można zilustrować, np. zaciemniając tę przestrzeń w sposób odzwierciedlający liczbę cząstek o danej prędkości (im większe zaciemnienie, tym więcej cząstek w danym obszarze). Postawmy teraz pytanie, jaka jest gęstość obszaru w pobliżu wybranego wektora prędkości o trzech składowych v_x, v_y, v_z – obszar ten można przedstawić w formie małego sześcianu „zaczepionego" w wierzchołku wektora. Z założenia anizotropii wynika, że funkcja reprezentująca tę gęstość powinna zależeć tylko od długości wektora $v = \sqrt{v_x^2 + v_y^2 + v_z^2}$, a nie od konkretnych wartości jego składowych (czyli jego kierunku). Na podstawie ogólnych rozważań odwołujących się do założonych symetrii, można wyprowadzić wzór charakteryzujący gęstość rozkładu cząstek o prędkościach (wektorowych) w pobliżu danego wektora. Funkcja rozkładu przyjmuje postać malejącej funkcji wykładniczej od kwadratu długości wektora v (jest to tzw. rozkład Gaussa, inaczej rozkład normalny):

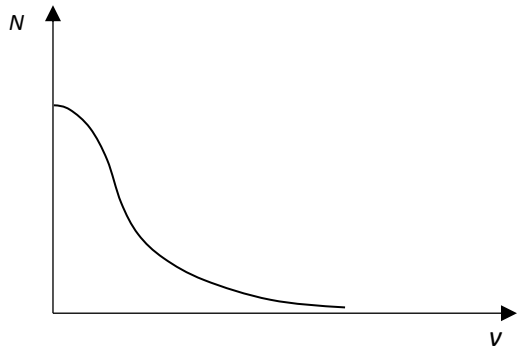
$$g(v_x, v_y, v_z) = \alpha e^{-\beta v^2}.$$

Maksimum tego rozkładu wypada w punkcie $v = 0$, a następnie jego wartości spadają (tzw. krzywa dzwonowa – por. rys. 3.6).

Aby wyznaczyć rozważany w głównym tekście rozkład cząstek o danych prędkościach traktowanych jako skalary, należy policzyć liczbę cząstek w kulistej powłoce o grubości *dv* i promieniu *v* (powłoka ta zawiera wszystkie wektory prędkości o długości z przedziału od *v* do *v* + *dv*). Dokonamy tego, mnożąc powyższą funkcję przez powierzchnię sfery $4\pi v^2$. W rezultacie otrzymamy rozkład Maxwella, którego maksimum wypada dla pewnej niezerowej wartości *v*, a którego wartość w punkcie *v* = 0 wynosi zero:

$$f(v) = 4\pi\alpha v^2 e^{-\beta v^2}.$$

Należy podkreślić, że funkcje *g* i *f*, choć obie zależą od *v*, mają inny sens fizyczny. Pierwsza z nich określa prawdopodobieństwo (dokładniej: gęstość prawdopodobieństwa) znalezienia cząstki poruszającej się z prędkością *v w danym kierunku*, a druga prawdopodobieństwo, że dana cząstka będzie miała prędkość *v* w *dowolnym* kierunku.[9] Skąd jednak bierze się paradoksalna rozbieżność między wartościami funkcji *g* i *f* w zerze? Jaki jest ostatecznie nasz wniosek dotyczący liczby cząstek poruszających się z bardzo małymi prędkościami? Niezerowa wartość funkcji *g* w punkcie *v* = 0 jest w pewnym sensie artefaktem matematycznym. Wynika ona stąd, że dla bardzo małych prędkości graniczących z 0 ich kierunek staje się nieistotny – niewielki sześcian zaczepiony infinitezymalnie blisko zera w praktyce obejmuje wszystkie kierunki ruchu. Natomiast dla dużych *v* funkcja *g* wybiera tylko jeden z nieprzeliczalnej liczby możliwych kierunków, nic więc dziwnego, że wartość funkcji spada z rosnącym *v*. W granicy $v \to 0$ kierunek wektora przestaje mieć sens fizyczny i to jest powodem anomalnego zliczania cząstek o infinitezymalnie małej prędkości przez funkcję *g*. Fizycznie pozostaje prawdą, że liczba cząstek o prędkościach coraz to bliższych zeru spada do zera, tak jak to opisuje funkcja Maxwella *f*.



Rys. 3.6. Rozkład prędkości w konkretnym wybranym kierunku. Chociaż funkcja rozkładu zależy tylko od długości wektora prędkości, to jednak jej sens jest inny niż w wypadku rozkładu Maxwella

[9] Uważny matematycznie czytelnik może zadać pytanie, dlaczego funkcje *g* i *f* mają inne wymiary, skoro widać wyraźnie, że w drugiej z nich mnożymy całe wyrażenie przez kwadrat prędkości o wymiarze $\frac{m^2}{s^2}$? Jest to związane z tym, że funkcje te określają gęstości prawdopodobieństw w innych obszarach. Aby z funkcji *g* otrzymać bezwymiarowe prawdopodobieństwo, należy ją przemnożyć przez element objętości $\Delta v_x \Delta v_y \Delta v_z$ o wymiarze $\frac{m^3}{s^3}$, natomiast funkcja *f* reprezentuje gęstość w objętości Δv, której wymiar to $\frac{m}{s}$.

Podsumujmy warunki udanej redukcji teorii T_1 do teorii T_2, jakie udało nam się określić na podstawie przypadku termodynamiki i mechaniki statystycznej. Zbierzmy je w formie punktów:

1. *Redukcja obiektów*. W pierwszym kroku identyfikujemy przedmioty opisane przez redukowaną teorię T_1 jako odpowiednie konstrukty z przedmiotów teorii T_2. W wypadku nauki o cieple utożsamiamy makroskopowe ciała fizyczne (gazy, ciecze, ciała stałe) z kompleksami składającymi się z ogromnej liczby cząsteczek.

2. *Redukcja własności*. Podstawowe własności opisywane przez teorię T_1 muszą zostać przedefiniowane przy pomocy pojęć stosowanych w teorii T_2. Pojęcia termodynamiczne, takie jak ciśnienie, temperatura czy ciepło zyskują definicje w kategoriach terminów występujących w mechanice klasycznej (oraz terminów statystycznych) – średniej energii kinetycznej, średniej zmiany pędu etc.

3. *Redukcja praw*. Prawa P_1 redukowanej teorii T_1 powinny być wyprowadzalne z praw P_2 teorii redukującej T_2, przy zastąpieniu terminów występujących w P_1 terminami teorii T_2 w sposób opisany w punkcie 2. Dopuszczalne jest przy tym stosowanie dodatkowych założeń łączących terminy teorii T_1 z terminami teorii T_2 (np. o charakterze statystycznym). Takie założenia nazywa się często „prawami pomostowymi". Przykładem redukcji praw jest wyprowadzenie pierwszej zasady termodynamiki z zasady zachowania energii, czy też prawa gazów doskonałych z zasad dynamiki Newtona i założeń statystycznych.

Pozostało nam jeszcze omówienie sposobu wyprowadzenia drugiej zasady termodynamiki z praw mechaniki statystycznej. Do tego celu będziemy jednak potrzebowali dodatkowego aparatu pojęciowego, który wprowadzimy w następnym paragrafie. Zetkniemy się w nim ponadto z fundamentalnym problemem nieodwracalności czasowej.

3.6. Druga zasada termodynamiki w ujęciu mechaniki statystycznej

Wprowadzimy teraz niezmiernie ważne pojęcia makrostanu, mikrostanu i przestrzeni fazowej. Przez makrostan danego układu fizycznego (np. próbki gazu) rozumiemy kompletny zestaw wartości makroskopowych wielkości mierzalnych charakteryzujących ten układ. Przykładami takich wielkości są dobrze znane ciśnienie, temperatura, objętość, oddane lub pobrane ciepło, gęstość, a także entropia. Natomiast mikrostan charakteryzuje cząsteczki wchodzące w skład wybranego układu. Cząsteczki te traktujemy jako nieprzenikliwe obiekty punktowe, które oddziałują ze sobą wyłącznie mechanicznie, a zatem mogą być całkowicie scharakteryzowane przez podanie ich położeń i prędkości (lub pędów). Mikrostanem układu będzie właśnie przypisanie każdej pojedynczej cząstce jej dokładnego położenia i pędu.

Zbierzmy kilka fundamentalnych faktów dotyczących mikrostanów i ich związku z makrostanami. Po pierwsze, mikrostany – w przeciwieństwie do makrostanów – nie są dla nas w praktyce dostępne. Nie mamy możliwości zaobserwowania pojedynczych cząstek ze względu na ich małe rozmiary i ogromną, przekraczającą wszelkie wyobrażenia liczbę. Zatem mikrostan jest w pewnym sensie użytecznym konstruktem teoretycznym, niemającym podłoża eksperymentalnego. Po drugie, dany mikrostan jednoznacznie wyznacza odpowiedni makrostan. Ze względu na założenie o redukowalności parametrów makroskopowych do własności cząsteczek musimy przyjąć, że dokładne wartości położeń i prędkości cząstek jednoznacznie determinują wszystko, co można zmierzyć w odniesieniu do danej próbki. Nato-

miast analogiczna zależność nie zachodzi w drugą stronę. Dla danego makrostanu istnieje ogromna liczba mikrostanów go realizujących. Wynika to stąd, że makrostany operują wielkościami uśrednionymi, a średnie wartości mogą być takie same dla różnych faktycznych rozkładów. Ciśnienie czy temperatura gazu nie ulegną zmianie, jeśli np. odpowiednio zmodyfikujemy indywidualne wartości prędkości cząsteczek.

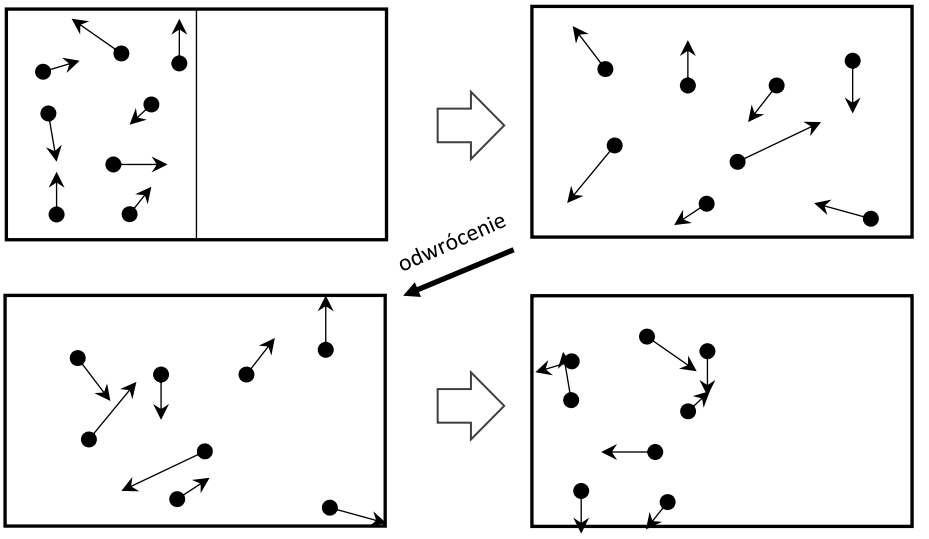
Zatem relacja między mikrostanami a makrostanami jest wielo-jednoznaczna. Mikrostany i makrostany można w wygodny sposób przedstawiać, korzystając z wprowadzonego przez nas w poprzednim rozdziale pojęcia przestrzeni fazowej. Przestrzeń fazowa dla grupy N cząstek to $6N$-wymiarowa przestrzeń, w której każdy punkt scharakteryzowany jest przez współrzędne reprezentujące liczbowo wszystkie składowe położenia pędów dla wszystkich cząstek. Ponieważ każda pojedyncza cząstka wymaga sześciu liczb do opisu jej stanu (trzy współrzędne położenia i trzy współrzędne pędu), otrzymujemy liczbę $6N$ parametrów. Zatem punkty w przestrzeni fazowej to po prostu mikrostany. A co z makrostanami? Skoro wielu mikrostanom może odpowiadać ten sam makrostan, naturalne jest przedstawiać makrostany w przestrzeni fazowej jako obszary zawierające odpowiednią liczbę sąsiadujących ze sobą punktów. Innym słowy przestrzeń fazowa rozpada się na ogromną liczbę komórek, z których każda odpowiada konkretnemu zestawowi wartości parametrów makroskopowych. Powstaje ważne pytanie, jak się mają do siebie rozmiary poszczególnych komórek – czy są one mniejwięcej tej samej wielkości, czy też różnią się od siebie? Na razie zostawimy to pytanie bez odpowiedzi, ale wrócimy do niego w następnym paragrafie.

Jak pamiętamy z poprzedniego rozdziału, mechanika w wersji hamiltonowskiej przedstawia ewolucję układu wielu cząstek w formie trajektorii (linii krzywej) w przestrzeni fazowej. W każdym punkcie przestrzeni fazowej zaczepiony jest abstrakcyjny wektor, którego zwrot i kierunek dany jest równaniami Hamiltona, i który wyznacza kierunek ewolucji układu. Ze względu na determinizm mechaniki klasycznej, z każdego punktu wychodzi dokładnie jedna trajektoria. Implikuje to od razu, że trajektorie w przestrzeni fazowej nie mogą się ze sobą przecinać. Rozważając wszystkie trajektorie wychodzące z danego makrostanu (danej komórki), możemy prześledzić (oczywiście tylko teoretycznie) ewolucję makroskopową układu fizycznego. W mechanice klasycznej obowiązuje ważne twierdzenie Liouville'a, które głosi, że objętość danego obszaru poddanego ewolucji Hamiltonowskiej nie ulega zmianie. Zatem dany makrostan musi zostać przetransformowany w wyniku ewolucji układu na obszar o tej samej wielkości. Natomiast nie musi być to obszar o takim samym „kształcie" – może on być np. bardzo rozczłonkowany.

Obecnie możemy już w nieco precyzyjniejszy sposób przedstawić problem związany z nieodwracalnością drugiej zasady termodynamiki. Jak pamiętamy, zasada ta głosi, że istnieje pewien parametr makroskopowy – entropia – który nie może maleć wraz z upływem czasu dla izolowanego układu. Oznacza to, że jeśli wykreślimy trajektorie wychodzące z komórki reprezentującej początkowy makrostan układu, to powinny one przecinać tylko komórki odpowiadające wyższej bądź równej entropii.[10] Jednakże z mechaniki klasycznej wiemy, że każda trajektoria reprezentująca ewolucję danego układu może być odwrócona w sensie zgodności z prawami dynamiki Newtona. Jeśli trajektoria w przestrzeni fazowej łączy dwa punkty A i B, implikuje to, że fizycznie dopuszczalna jest zarówno ewolucja z A

[10] Jak się przekonamy, nie jest to dokładnie spełnione w rzeczywistości. Warunek ten byłby spełniony tylko wówczas, gdyby druga zasada termodynamiki była prawem bezwyjątkowym, a tak nie jest. Na razie jednak zakładamy, że prawa termodynamiki obowiązują ściśle.

do B, jak i z B do A. To jednak prowadzi do wniosków sprzecznych z drugą zasadą termodynamiki. Rozważmy trajektorię wychodzącą z jakiegoś punktu w komórce makrostanu o niskiej entropii do punktu w makrostanie o wyższej entropii. Mechanika klasyczna dopuszcza proces odwrotny, który jednak prowadziłby ze stanu o wyższej entropii do stanu o entropii niższej, a to jest zabronione. (Fizyczna realizacja takiego procesu polegałaby na momentalnym odwróceniu zwrotu wszystkich chwilowych prędkości cząstek – w takiej sytuacji cząstki przebiegłyby swoje przeszłe drogi do punktu wyjścia.) Mamy zatem poważny problem, który stawia pod znakiem zapytania cały program redukcji termodynamiki.

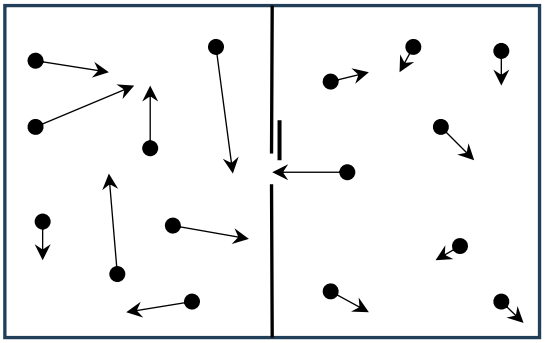


Rys. 3.7. Argument Loschmidta z odwrócenia prędkości

Problem odwracalności mikroskopowych procesów i jej niezgodności z drugą zasadą termodynamiki został zauważony na wczesnym etapie rozwoju mechaniki statystycznej. Argument za teoretyczną możliwością odwrócenia procesów termodynamicznych za pomocą odwrócenia zwrotów prędkości nosi nazwę argumentu Loschmidta od nazwiska jego twórcy. Wyobraźmy sobie sytuację, w której gaz początkowo zamknięty w pewnej części naczynia zaczyna się spontanicznie rozprężać po usunięciu przegrody. W momencie, w którym gaz zajmuje już całą objętość, każda cząsteczka ma odpowiednio skierowany wektor prędkości. Rozważmy teraz stan, w którym każda cząsteczka posiada prędkość taką samą co do wartości, lecz skierowaną przeciwnie. Co prawda nie umiemy praktycznie wytworzyć takiego stanu, ale jak się wydaje, jest on absolutnie możliwy zgodnie z fundamentalnymi prawami przyrody. W takiej sytuacji cząsteczki zaczną się „cofać" do uprzedniego stanu, a cała próbka gazu zajmie połowę naczynia bez interwencji z zewnątrz (rys. 3.7). To jednak łamie drugą zasadę termodynamiki.

Innym spektakularnym przykładem podważającym uniwersalną ważność drugiej zasady był argument, znany pod nazwą demona Maxwella (z dość oczywistym nawiązaniem do demona Laplace'a). Rozważmy pudełko składające się z dwóch części oddzielonych przegrodą (rys. 3.8). W jednej części znajduje się gaz o wyższej temperaturze, a w drugiej gaz o temperaturze niższej. W przegrodzie został wykonany otwór zasłonięty przesłoną. Mechanizm otwierający przesłonę jest zaprojektowany tak, że jeśli ze strony zimniejszego gazu zbliża się wolno poruszająca się cząstka, przesłona pozostaje zamknięta. Natomiast jeśli cząstka po-

siada odpowiednio dużą prędkość, przesłona ulega otwarciu, przepuszczając wysokoenergetyczną cząstkę[11] (ze względu na oczywiste trudności techniczne związane z wykonaniem takiego mechanizmu, jego operację powierza się nieco żartobliwie demonowi o nadludzkich zdolnościach). W wyniku działania opisanego mechanizmu cieplejszy gaz uzyska pewną liczbę wysokoenergetycznych cząstek, co zwiększy jego temperaturę, natomiast gaz zimniejszy jeszcze bardziej się ochłodzi. To jednak jest niezgodne z drugą zasadą termodynamiki w wersji Clausiusa.



Rys. 3.8. Demon Maxwella. Przesłona zostaje otwarta, kiedy z zimniejszej porcji gazu (prawa strona) zbliża się wysokoenergetyczna cząstka

Argument Maxwella opiera się na hipotezie, że opisane powyżej działanie demona nie łamie żadnej fundamentalnej zasady mechaniki klasycznej. Istnieje jeszcze bardziej przekonujący argument przeciwko drugiej zasadzie, który pokazuje, że jej złamanie jest nie tylko możliwe, ale pewne. Argument ten wykorzystuje tzw. twierdzenie Poincarégo o powrotach. Henri Poincaré udowodnił w matematycznie ścisły sposób, że każda trajektoria w przestrzeni fazowej podlegająca ewolucji newtonowskiej z konieczności musi kiedyś przejść dowolnie blisko punktu wyjściowego. Czas potrzebny na taki infinitezymalnie bliski powrót jest oczywiście niewyobrażalnie długi, ale jest matematycznie pewne, że kiedyś musi to nastąpić. Zatem aby pokazać, że entropia układu może spontanicznie się obniżyć, nie musimy niczego odwracać – wystarczy, że poczekamy odpowiednio długo, a układ ze stanu o wysokiej entropii sam wróci do makrostanu pierwotnego o niższej entropii.

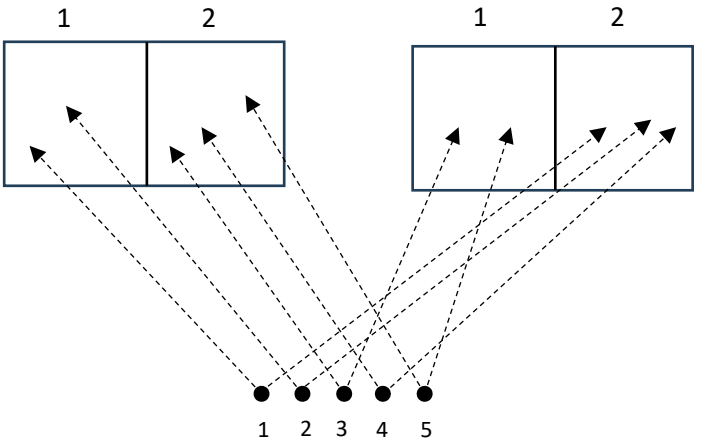
3.7. Statystyczny argument Boltzmanna i strzałka czasu

Fizycy uświadomili sobie, że stoją przed poważnym problemem przy próbie interpretacji drugiej zasady termodynamiki. Jedyną drogą, jaka wydaje się możliwa w świetle argumentów za odwracalnością, jest przyjęcie, że druga zasada nie jest uniwersalnym, bezwyjątkowym prawem przyrody, a jedynie prawidłowością statystyczną. Po prostu procesy zmniejszające entropię w skali czasu dostępnej obserwatorowi ludzkiemu są niezmiernie mało prawdopodobne. Dlaczego jednak tak miałoby być? Jak można wyprowadzić z zasad mechaniki klasycznej tezę o bardzo dużym prawdopodobieństwie (graniczącym z pewnością) wzro-

[11] Zauważmy, że zgodnie z rozkładem Maxwella (rys. 3.4) nawet w zimnym gazie znajdą się cząstki o bardzo dużej prędkości, choć ich liczba będzie niewielka w porównaniu z liczbą w gazie cieplejszym.

stu entropii? Zadania tego podjął się Boltzmann w swoim słynnym argumencie statystycznym (kombinatorycznym), którego szkic przedstawimy poniżej. Zanim jednak tego dokonamy, musimy wprowadzić jeszcze jeden rodzaj abstrakcyjnej przestrzeni (jak się już zapewne zorientowaliście, fizycy uwielbiają przeróżne przestrzenie). Jest to w zasadzie pewien wariant przestrzeni fazowej – nazywamy go przestrzenią μ. W przestrzeni fazowej jeden punkt zawiera kompletną informację na temat stanu *N* cząstek. Natomiast w wypadku przestrzeni μ potrzebujemy *N* punktów, z których każdy reprezentuje stan pojedynczej cząstki. Zatem przestrzeń μ jest „tylko" sześciowymiarowa – trzy wymiary do określenia położenia i trzy wymiary pędu. Oczywiście obie przestrzenie są ściśle ze sobą powiązane – każdemu punktowi w przestrzeni fazowej odpowiada *N* punktów w przestrzeni μ.

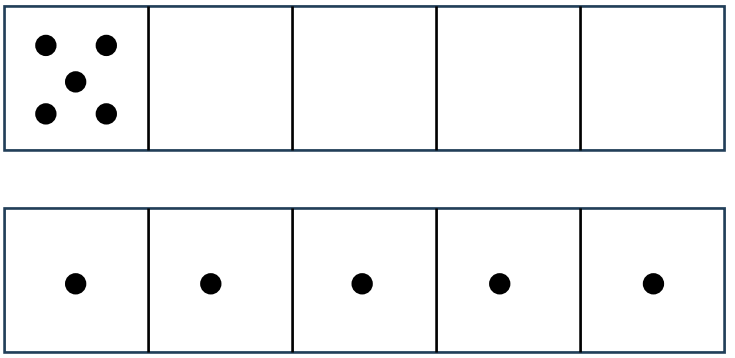
Podzielmy teraz przestrzeń μ na dużą liczbę komórek o jednakowych kształtach i rozmiarach. Każda komórka reprezentuje pewien przedział możliwych wartości pędu i położenia, w którym mogą się znaleźć pęd i położenie każdej cząstki. W takim wypadku będziemy po prostu mówić w skrócie, że dana cząstka znalazła się w określonej komórce. Przypisanie każdej cząstki do ściśle określonej komórki nazwiemy aranżacją. Natomiast określenie dla każdej komórki liczby przypadających na nią cząstek bez podania, które cząstki są w której komórce, określimy mianem dystrybucji. Na jedną dystrybucję przypada oczywiście w ogólności wiele aranżacji. Na przykład dystrybucja, w której w komórce numer jeden są dwie cząstki, a w komórce numer dwa trzy cząstki, może być zrealizowana przez aranżację: cząstki numer 1 i 2 w komórce 1, cząstki numer 3, 4 i 5 w komórce 2, ale równie dobrze przez aranżację: cząstki numer 3 i 5 w komórce 1, cząstki numer 1, 2 i 4 w komórce 2 (rys. 3.9). Różnice między aranżacjami dla danej dystrybucji nie są istotne z fizycznego punktu widzenia, gdyż zakładamy, że wszystkie cząstki są identyczne. Boltzmann przyjął, że można dobrać wielkość komórek w taki sposób, aby każda dystrybucja odpowiadała dokładnie jednemu makrostanowi. Można więc zadać pytanie, ile aranżacji realizuje dany makrostan (daną dystrybucję).



Rys. 3.9. Dwie aranżacje realizujące te same dystrybucje

Odpowiadając na to pytanie, zauważmy przede wszystkim, że liczba aranżacji zależy bardzo silnie od równomierności danej dystrybucji. Nierównomiernie rozłożone dystrybucje mają mało realizujących je aranżacji, a równomiernie rozłożone dużo. Dla ilustracji rozważmy nierównomierną dystrybucję, w której wszystkie pięć cząstek znajduje się w jednej

z pięciu komórek (rys. 3.10). W takiej sytuacji mamy tylko jedną aranżację realizującą ten makrostan. Natomiast idealnie równomierna dystrybucja przydzielająca każdej z pięciu komórek po jednej cząstce ma 5! = 120 aranżacji. Jednakże liczba aranżacji dla danej dystrybucji jest jednoznacznie skorelowana z wielkością odpowiadającego tej dystrybucji obszaru w przestrzeni fazowej! Dystrybucja, która ma 120 aranżacji, odpowiada obszarowi sto dwadzieścia razy większemu niż dystrybucja z jedną aranżacją.[12] Można z tego wyciągnąć wniosek, że różnym makrostanom będą odpowiadać obszary w przestrzeni fazowej o bardzo zróżnicowanych wielkościach i że największe będą makrostany o względnie równomiernie rozłożonych dystrybucjach. Dla realistycznych liczb cząstek i komórek różnice wielkości między komórkami makrostanów będą niewiarygodnie wielkie, przekraczając nasze codzienne wyobrażenia. W istocie przestrzeń fazowa dla typowego makroskopowego układu fizycznego będzie wyglądała tak, że niemal całą jej objętość zajmuje jedna gigantyczna komórka jednego makrostanu, a pozostałe komórki będą znikomymi pyłkami na jej obrzeżach, niewidocznymi nawet pod mikroskopem (rys. 3.11). Ten olbrzymi makrostan to stan równowagi, w którym żadne parametry układu nie ulegają zmianie.



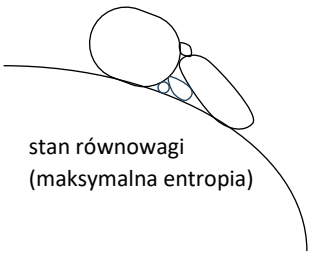
Rys. 3.10. Dwie możliwe dystrybucje pięciu cząstek w pięciu komórkach – nierównomierna (góra) i równomierna (dół)

Rozróżnienie między dystrybucjami a aranżacjami może napotkać sprzeciw ze strony zwolenników zasady tożsamości przedmiotów nieodróżnialnych, o której mówiliśmy przy okazji analizy argumentu Leibniza z przesunięcia w paragrafie 2.5. Jeśli założymy, że cząstki wchodzące w skład danego zespołu statystycznego nie różnią się od siebie fizycznie poza posiadaniem różnego położenia i ewentualnie prędkości, to jak łatwo zauważyć, wszystkie aranżacje realizujące daną dystrybucję staną się kompletnie nieodróżnialne.

[12] Wynika to stąd, że każdej aranżacji danej dystrybucji odpowiada obszar w przestrzeni fazowej o dokładnie takiej samej wielkości. Spróbujcie przekonać się sami na uproszczonym przykładzie. Załóżmy, że mamy tylko dwie komórki w przestrzeni μ i dwie cząstki, z których każda scharakteryzowana jest tylko jednym parametrem o wartości od –1 do 1. Jedna komórka obejmuje wartości od –1 do 0, a druga od 0 do 1. Przestrzeń fazowa w tym wypadku będzie dwuwymiarowa (na osi *x* odkładamy parametr cząstki pierwszej, a na *y* drugiej). Narysujcie, proszę, obszary w przestrzeni fazowej, które odpowiadają czterem możliwym aranżacjom, i przekonajcie się, że są one kwadratami o takich samych rozmiarach.

Dla zwolennika Leibniza nie istnieje różnica między aranżacją: cząstka numer 1 w pierwszej komórce, cząstka numer 2 w drugiej, a aranżacją „zamienioną": cząstka numer 2 w pierwszej komórce, cząstka 1 w drugiej. Przy takim podejściu wszystkie aranżacje wchodzące w skład danej dystrybucji należy utożsamić – każda dystrybucja ma dokładnie jedną jakościowo zdefiniowaną aranżację. Czy jednak nie znaczy to, że argument, że równomiernym dystrybucjom odpowiadają dużo większe obszary w przestrzeni fazowej, upada?

Na szczęście argument Boltzmanna jest niezależny od tego, jak będziemy liczyć jakościowo nieodróżnialne aranżacje. Jeśli pójdziemy za Leibnizem, to konsekwentnie będziemy musieli także zmodyfikować naszą interpretację przestrzeni fazowej. Musimy utożsamić ze sobą wszystkie punkty przestrzeni fazowej, które różnią się jedynie „zamianą" (permutacją) cząstek. Odpowiada to podzieleniu każdego obszaru przez czynnik $N! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot N$ (liczba permutacji N-elementowego zbioru). Można się przekonać, że „okrojony" w ten sposób obszar w przestrzeni fazowej, odpowiadający równomiernej dystrybucji, będzie nadal dużo większy od obszaru nierównomiernej dystrybucji. Zachęcam wnikliwych czytelników do zbadania tego samodzielnie na uproszczonym przykładzie opisanym w przypisie 12. W tym wypadku przestrzeń fazowa nieodróżnialnych cząstek będzie połową płaszczyzny poniżej (lub powyżej – nie ma to znaczenia) przekątnej $x = y$. Obszar przestrzeni fazowej odpowiadający nierównomiernej dystrybucjo-aranżacji, w której obie cząstki są w jednej komórce, będzie trójkątem (połową kwadratu) o krótszym boku 1, podczas gdy obszar odpowiadający równomiernej dystrybucji będzie pojedynczym kwadratem, czyli będzie dwa razy większy.



Rys. 3.11. Porównanie rozmiarów komórek makrostanów w przestrzeni fazowej

Aby przejść do uzasadnienia drugiej zasady termodynamiki, musimy jeszcze dokonać statystycznej reinterpretacji pojęcia entropii. Nie jest to trudne – w myśl statystycznej definicji entropia danego makrostanu to po prostu objętość odpowiadającego mu obszaru w przestrzeni fazowej. Dokładniej, Boltzmann scharakteryzował entropię jako logarytm objętości z odpowiednim współczynnikiem proporcjonalności:

$$S = k \log V$$

Ponieważ, jak już wspomnieliśmy, objętość obszaru odpowiadającego danemu makrostanowi jest proporcjonalna do liczby aranżacji realizujących ten makrostan, entropia może być

również zdefiniowana jako logarytm z tej liczby. Zatem widzimy, że entropia równomiernych rozkładów będzie wyższa, a nierównomiernych niższa.[13]

Możemy teraz wytłumaczyć popularne ujęcie entropii jako miary nieuporządkowania. Dany rozkład nazywamy nieuporządkowanym, gdy istnieje bardzo duża liczba realizacji tego rozkładu powstałych przez zamianę obiektów. Na przykład jeśli rozważymy pojemnik z gazem podzielony na niewielkie komórki, a w każdej komórce znajdzie się mniej więcej ta sama liczba cząstek, to taki rozkład będzie miał wysoki stopień nieuporządkowania. Z drugiej strony, rozkład cząstek, w którym prawie wszystkie cząstki znajdują się w jednej komórce (np. w lewym górnym rogu pojemnika), jest uporządkowany. Liczba realizacji takiego rozkładu, a zatem także jego entropia, jest minimalna. To samo dotyczy także rozkładów prędkości (pamiętajmy, że przestrzeń fazowa, w której rozważamy mikro- i makrostany, zawiera informację zarówno o położeniach, jak i prędkościach cząstek). Stan gazu, w którym większość cząstek porusza się w jednym kierunku z podobną prędkością, jest uporządkowany, natomiast stan, dla którego cząstki poruszają się chaotycznie w różnych kierunkach z różnymi prędkościami, charakteryzuje się wysoką entropią. Istnieje silny związek entropii i nieuporządkowania z prawdopodobieństwem. Jeśli przyjmiemy założenie, że każda realizacja (aranżacja) jest równie prawdopodobna, to stany o wysokiej entropii będą bardzo prawdopodobne, gdyż istnieje bardzo dużo sposobów ich realizacji, w przeciwieństwie do stanów o niskiej entropii, które są mało prawdopodobne.

Pojęcie entropii jest także silnie związane z teorią informacji. Wyobraźmy sobie, że mamy pewne zdarzenie losowe A z kilkoma możliwymi rezultatami a_1, …, a_n (np. rzut kostką do gry). Informacja związana z konkretnym rezultatem takiego zdarzenia jest wyznaczona przez jego prawdopodobieństwo – rezultaty bardzo prawdopodobne (bliskie pewności) niosą ze sobą mało informacji, gdyż i tak wiedzieliśmy, że raczej na pewno zajdą. Natomiast mało prawdopodobne rezultaty dają odpowiednio więcej informacji. Informacja „zawarta" w jednym wyniku a_i zdarzenia losowego o prawdopodobieństwie p_i może być zdefiniowana następująco (jest to tzw. definicja informacji Shannona):

$$H(a_i) = -\log_2 p_i$$

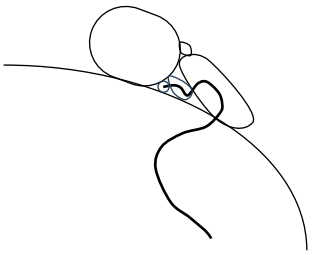
W formule powyższej stosujemy logarytm o podstawie 2, co jest standardem w informatyce (znaczy to, że mierzymy informację w bitach). Znak minus bierze się stąd, że logarytm liczby mniejszej od 1 (prawdopodobieństwa) jest ujemny. Następnie możemy obliczyć informację zawartą w całym zdarzeniu losowym A. Będzie to średnia informacji poszczególnych rezultatów, ważona oczywiście ich prawdopodobieństwami:

$$H(A) = -\sum_i p_i \log_2 p_i$$

[13] Użycie funkcji logarytmicznej w definicji entropii jest pewną konwencją, która jednak ma dobre uzasadnienie. Po pierwsze, logarytmy z dużych liczb są stosunkowo małymi wielkościami, co pozwala na łatwiejsze operowanie entropią. Po drugie i ważniejsze, funkcja logarytmiczna zapewnia nam istotną własność entropii, a mianowicie jej addytywność. Jeśli rozważymy dwa niezależne rozkłady, dla których liczby realizujących je aranżacji wynoszą odpowiednio N i M, to suma tych rozkładów będzie miała $N{\cdot}M$ aranżacji. Ponieważ logarytm iloczynu liczb jest sumą ich logarytmów, entropia całego układu będzie sumą entropii jego składników.

Formuła powyższa zachowuje się analogicznie do funkcji entropii i jest w istocie tożsama z entropią. Przyjmuje ona największą wartość dla rozkładów symetrycznych, dla których prawdopodobieństwa p_i są mniej więcej równe, natomiast minimalną wartość dla rozkładów niesymetrycznych, gdzie jedno prawdopodobieństwo dominuje pozostałe. Na przykład dla dwóch rezultatów (rzut monetą) informacja zawarta w rozkładzie symetrycznym $p_1 = p_2 = ½$ wynosi 1 bit, natomiast informacja rozkładu asymetrycznego $p_1 = 1$, $p_2 = 0$ jest zerowa.

Jesteśmy już gotowi do przedstawienia argumentu Boltzmanna formułującego statystyczne uzasadnienie drugiej zasady termodynamiki. Rozważmy układ fizyczny znajdujący się w makrostanie o niskiej entropii – na przykład gaz zawarty w jednej części pojemnika zaraz po usunięciu przegrody. Z obszaru reprezentującego w przestrzeni fazowej ten makrostan możemy wyprowadzić wiele trajektorii. Ponieważ, jak już zauważyliśmy, przytłaczająca część przestrzeni fazowej będzie zajęta przez komórkę odpowiadającą makrostanowi o wysokiej entropii, zdecydowana większość trajektorii będzie musiała „wpaść" do tej megakomórki i tam pozostać przez bardzo długi czas (rys. 3.12).[14] Prawdopodobieństwo, że wybrana na chybił trafił trajektoria poprowadzi nas do jeszcze mniejszych obszarów niż początkowy makrostan jest zupełnie pomijalne. Takie trajektorie istnieją, ale są one w zdecydowanej mniejszości.



Rys. 3.12. Typowa trajektoria w przestrzeni fazowej prowadząca od makrostanu o małej entropii do makrostanu o dużej entropii (stan równowagi)

Ten poglądowy, jakościowy argument może być uzupełniony o dokładne wyliczenia, które pokazują, jak bardzo mało prawdopodobny jest scenariusz, w którym np. gaz zamknięty w naczyniu spontanicznie skupi się w jednej jego części, albo też jedna z jego części spontanicznie się ogrzeje, a druga oziębi. Wynika z tego, że średni czas oczekiwania na takie wydarzenie przekracza znany wiek całego wszechświata, nic więc dziwnego, że w praktyce tego nie obserwujemy. Zatem wydaje się, że nasz cel został osiągnięty. Druga zasada termodynamiki nie jest bezwyjątkowym prawem przyrody, takim jak np. zasady dynamiki Newtona. Natomiast jest niezwykle mało prawdopodobne, abyśmy mogli kiedykolwiek napotkać proces łamiący tę zasadę.

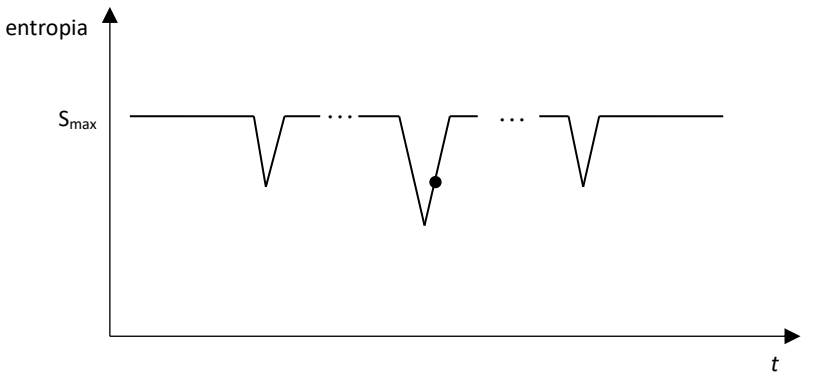
[14] Do uzasadnienia tego wniosku przydatne jest wspomniane wcześniej twierdzenie Liouville'a, które zapewnia, że obszar zajmowany przez docelowe punkty trajektorii będzie miał tę samą objętość, co wyjściowy makrostan. Gdyby obszar ten znacznie zwiększył swoją objętość, nie byłoby oczywiste, że zdecydowana większość punktów końcowych znajdzie się w makrostanie równowagi.

Niestety argument Boltzmanna obciążony jest poważnymi wadami, co stawia pod znakiem zapytania jego trafność uzasadnienia nieodwracalności zjawisk termodynamicznych. Po pierwsze, można mieć zastrzeżenia natury filozoficznej do sposobu, w jaki Boltzmann wykorzystuje pojęcie prawdopodobieństwa. Argument powyższy opiera się na założeniu o równym prawdopodobieństwie wszystkich mikrostanów wchodzących w skład danego makrostanu wyjściowego. Na jakiej podstawie możemy przyjąć to założenie? Typową odpowiedzią jest tu przyjęcie interpretacji prawdopodobieństwa opartej na niewiedzy (ignorancji). Skoro nie wiemy, jaki konkretnie jest mikrostan naszego układu, to bezpiecznie jest przyjąć, że wszystkie możliwości są równouprawnione. Jednakże takie założenie nie mówi nic na temat świata, a tylko charakteryzuje stan naszej wiedzy. W sytuacji całkowitej niewiedzy racjonalne będzie przyjęcie równych prawdopodobieństw, ale dlaczego przyroda miałaby dostosowywać się do naszej racjonalności? Jest do pomyślenia, że istnieje jakiś ukryty, nieznany nam mechanizm, który preferuje pewne mikrorealizacje danego makrostanu. Dopuszczenie takiej możliwości rujnuje cały argument, gdyż nie możemy już opierać się na prostym liczeniu trajektorii w celu określenia prawdopodobnego przebiegu ewolucji systemu. Niektóre z trajektorii mogą okazać się same z siebie zupełnie nieprawdopodobne lub wręcz wykluczone (zobaczymy zresztą za chwilę, że taką hipotezę uznaje wielu naukowców).

Drugi zarzut pod adresem argumentu Boltzmanna jest bardziej fundamentalny. Otóż łatwo pokazać, że argument ten w żaden sposób nie uzasadnia czasowej asymetryczności procesów termodynamicznych. W istocie jest on całkowicie symetryczny! Złudzenie asymetrii bierze się stąd, że rozważaliśmy ewolucję układu w kierunku przyszłości. Spróbujmy jednak zastanowić się, jak powinna była wyglądać ewolucja układu w przeszłości. Rozważmy zatem wszystkie możliwe trajektorie, które prowadzą do aktualnego makrostanu o niewielkiej entropii. Jest oczywiste, że będą to w istocie te same trajektorie, które braliśmy pod uwagę w powyższym argumencie. (Trajektorie w przestrzeni fazowej nie mają żadnych „strzałek", które wyznaczałyby ich kierunek – mogą one reprezentować ewolucję zarówno w jedną stronę, jak i w przeciwną.[15]) Zatem dokładnie ten sam statystyczny argument przekonuje nas, że przytłaczająca większość trajektorii będzie pochodziła od stanów o większej entropii. Oznacza to, że jest niezwykle prawdopodobne, iż obecny stan układu wyewoluował ze stanu o wyższej entropii, a więc druga zasada termodynamiki nie sprawdza się dla przeszłych procesów. Co więcej, wniosek ten nie zgadza się zupełnie z doświadczeniem. Weźmy na przykład momentalny stan gazu, w którym zajmuje on jedną trzecią objętości całego naczynia. Co jest bardziej prawdopodobne: to, że minutę wcześniej zajmował on jedną czwartą objętości, czy że zajmował całą objętość? Jest oczywiste, że pierwsza opcja jest dużo bardziej prawdopodobna – nie widzieliśmy nigdy, żeby gaz zajmujący całą objętość naczynia „skurczył" się do jednej trzeciej objętości. Jednakże argument Boltzmanna zastosowany do przeszłości zdecydowanie preferuje drugą opcję. Pokazuje, że najbardziej prawdopodobny scenariusz

[15] A co ze strzałkami reprezentującymi uogólnione prędkości, obliczone z równań Hamiltona (rys. 2.15 z poprzedniego rozdziału)? Ich wybór uzależniony jest od milcząco przyjętego kierunku upływu czasu, wyznaczonego zastosowanym parametrem t. Jeśli zamienimy parametr czasu t na jego „lustrzane" odbicie $-t$ i rozwiążemy odpowiednie równania Hamiltona, kierunki strzałek zmienią się na przeciwne. Zatem sama trajektoria w żaden sposób nie określa, w którą stronę będziemy się po niej posuwali.

ewolucji danego układu jest taki, że entropia stanu teraźniejszego jest najniższa, a stanów przeszłych i przyszłych wyższa.[16]



Rys. 3.13. Hipoteza fluktuacji Boltzmanna. Obecny stan wszechświata jest zaznaczony punktem na krzywej wznoszącej

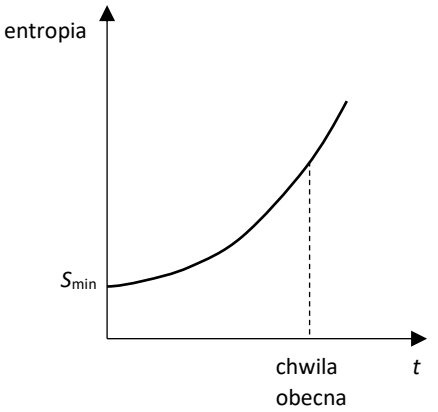
Sam Boltzmann był świadomy wagi naszkicowanego powyżej problemu. Aby się z nim uporać, sięgnął do rozważań o charakterze kosmologicznym, czyli dotyczących całego wszechświata. Przyjął, że w historii wszechświata dominujący powinien być stan o maksymalnej entropii, czyli tzw. śmierci cieplnej. Jest to stan, w którym całkowita energia wszechświata jest maksymalnie rozproszona – nie istnieją żadne skupiska materii i energii takie jak gwiazdy czy galaktyki. Wszechświat jest po prostu jednolitym, ciemnym pyłem o gęstości energii ledwo przekraczającej absolutne zero. Jednakże co jakiś czas pojawiają się w nim drobne fluktuacje, które skutkują obniżeniem globalnej entropii – materia zaczyna się skupiać w formie gwiazd i planet. Wszechświat osiąga pewną minimalną wartość entropii, po czym zaczyna ona rosnąć, zgodnie z drugą zasadą termodynamiki, aż ponownie zostanie osiągnięty stan śmierci cieplnej. Cykl ten powtarza się przy następnej fluktuacji, całe eony później (rys. 3.13). Aktualny stan wszechświata znajduje się na krzywej wznoszącej, dlatego obserwujemy wokoło procesy, które zwiększają entropię. Na pytanie, dlaczego mamy szczęście znajdować się w takim uprzywilejowanym stanie, którego trwanie jest niewiarygodnie krótsze od czasu trwania okresu maksymalnej entropii, jest prosta odpowiedź – powstanie złożonych organizmów jest możliwe tylko w świecie, dysponującym bogatymi źródłami energii o niskiej entropii (na przykład we wnętrzu gorących gwiazd, których wybuchy w postaci supernowych tworzą pierwiastki niezbędne do budowy życia). W stanie śmierci cieplnej nie może powstać żadna skomplikowana struktura fizyko-chemiczna.

Hipoteza fluktuacji Boltzmanna podaje globalne wyjaśnienie drugiej zasady termodynamiki i kierunkowości czasowej procesów. Warto przy tym zauważyć, że nie jest istotne, czy aktualny stan wszechświata znajduje się na wznoszącej czy opadającej krzywej entropii (por. wpis w ramce na s. 96). Jeśli powiążemy strzałkę czasu z globalnym wzrostem entropii,

[16] Można zauważyć, że konkluzja ta jest w istocie absurdalna. Z jednej strony, argument pokazuje, że entropia stanu układu w chwili t_1 pięć minut po chwili obecnej t_0 powinna być wyższa od aktualnej entropii. Z drugiej strony, stosując ten sam argument do momentu t_1, wnioskujemy, że entropia pięć minut wcześniej w chwili t_0 powinna być wyższa. Dostaliśmy zatem sprzeczność.

krzywa nazwana opadającą będzie w istocie dla nas wznosząca. Istotne jest tylko to, że globalny gradient entropii ma być niezerowy.

Mimo swojej atrakcyjności hipoteza Boltzmanna narażona jest na poważne zarzuty. Przede wszystkim należy zauważyć, że opiera się ona na założeniu o nieskończoności czasowej wszechświata, które w czasach Boltzmanna było powszechnie przyjmowane. Założenie to jest potrzebne do wprowadzenia fluktuacji, których zajście jest niesłychanie mało prawdopodobne. Jednakże według obecnego stanu wiedzy, wszechświat liczy sobie kilkanaście miliardów lat, a to jest zdecydowanie zbyt krótki czas, żeby połączyć okresy śmierci cieplnej z fluktuacjami entropii. Co więcej, wydaje się, że kosmologiczny argument Boltzmanna nie rozwiązuje podstawowego problemu, jakim jest czasowa symetria argumentu statystycznego. Zgodnie z rozważaniami statystyczno-kombinatorycznymi najbardziej prawdopodobnym stanem wszechświata podczas okresu fluktuacji jest stan o najniższej entropii, dla którego zarówno przeszłość, jak i przyszłość charakteryzuje się pozytywnym gradientem entropii. Jednakże to nie zgadza się z naszymi obserwacjami. Nadal więc potrzebujemy wyjaśnienia, dlaczego znajdujemy się na krzywej wznoszącej, a nie w samym minimum entropii.



Rys. 3.14. Hipoteza małej entropii początkowej wszechświata. Przedstawiona krzywa reprezentuje najbardziej prawdopodobną ewolucję od momentu początkowego do chwili obecnej, zgodnie z argumentem statystycznym Boltzmanna

Wyjaśnienia takiego dostarcza inna hipoteza, znana pod nazwą hipotezy małej entropii początkowej. Zauważmy, że odwrócony czasowo argument Boltzmanna opiera się na założeniu, że wszystkie trajektorie w przestrzeni fazowej prowadzące z przeszłości do chwili teraźniejszej są fizycznie możliwe. Można jednak wyeliminować większość z nich przez przyjęcie, że rozważany układ w dalekiej przeszłości zajmował makrostan o jeszcze mniejszej niż obecnie entropii. Na przykład rozważając sytuację, w której gaz zajmuje jedną trzecią całej objętości naczynia, możemy wprowadzić informację, że godzinę temu gaz był jeszcze bardziej sprężony, powiedzmy, do jednej dziesiątej objętości. Wtedy z argumentu statystycznego wynika, że najbardziej prawdopodobna ewolucja gazu w ciągu ostatniej godziny to rozprężanie się połączone ze stałym wzrostem entropii. Większość anomalnych trajektorii rozpoczynających się od dużej entropii została po prostu wyeliminowana przez dodatkowe założenie. W przypadku wszechświata potrzebujemy założenia o globalnym charakterze. Jest to właśnie hipoteza, że stan wszechświata w chwili początkowej (Wielkiego Wybuchu) cha-

rakteryzował się niezmiernie małą entropią. Przyjmując tę tezę łatwo możemy wyjaśnić, dlaczego entropia stanów przeszłych jest mniejsza od entropii w chwili obecnej – ponieważ dawniej była jeszcze mniejsza (rys. 3.14).

Dlaczego entropia stanu początkowego wszechświata była tak niewielka? Na to pytanie nauka nie znajduje odpowiedzi – jest to po prostu fakt, który nie może być wyjaśniony. Niektórzy filozofowie uważają, że teza o małej początkowej entropii ma status prawa przyrody, podobnie jak prawo powszechnego ciążenia czy prawa dynamiki Newtona. Nikt nie pyta, dlaczego planety przyciągają się z siłą odwrotnie proporcjonalną do kwadratu odległości – po prostu tak jest. Z kolei inni uważają, że jednostkowe tezy, takie jak „Entropia wszechświata w chwili zero wynosi tyle a tyle", nie powinny być nazywane prawami. Tylko ogólne twierdzenia mogą uzyskać status praw. Niezależnie od tego, kto ma rację w tym sporze, pozostaje faktem, że asymetrię czasową procesów fizycznych umiemy wyjaśnić tylko poprzez założenie asymetrii jednostkowych faktów dotyczących entropii – entropia w chwili początkowej była niższa od obecnej entropii.

Pytania i problemy

1. Objaśnij różnicę między ciepłem a temperaturą, wykorzystując do tego celu heurystyczne pojęcie cieplika.

2. Czy eksperymenty Joule'a istotnie pokazały, że ciepło i praca mechaniczna są sobie równoważne (są tymi samymi wielkościami)? Jaki rezultat jego doświadczeń uznalibyśmy za obalający taką hipotezę?

3. W sformułowaniu pierwszej zasady termodynamiki pojawia się pojęcie energii wewnętrznej. Czy nie można użyć tego pojęcia do efektywnego zabezpieczenia owej zasady przed możliwością eksperymentalnego obalenia? W każdej sytuacji, w której wykonana praca nie równoważy się z dostarczonym ciepłem, różnicę między tymi dwoma wielkościami interpretujemy jako wzrost lub zmniejszenie odpowiedniej porcji energii wewnętrznej. Czy pierwsza zasada nie staje się wtedy zdaniem analitycznym?

4. Podaj trzy wersje drugiej zasady termodynamiki. Pokaż, że każda z tych wersji wprowadza pewien rodzaj nieodwracalności zjawisk w czasie. Przedyskutuj kwestię równoważności tych sformułowań.

5. Wykaż, że entropia termodynamiczna gazu zajmującego całą objętość pojemnika jest większa od entropii tego gazu zajmującego połowę pojemnika, rozważając proces powolnego sprężania gazu za pomocą tłoka.

6. Czy twierdzenie „Dla każdego izolowanego układu jego entropia nie maleje w czasie" jest zdaniem analitycznym *a priori* czy syntetycznym *a posteriori*? Rozważ ten problem zakładając, że porządek czasowy jest dany niezależnie od wzrostu entropii, jak również przy założeniu, że porządek czasowy jest wyznaczony wzrostem entropii.

7. Wymień podstawowe warunki, jakie muszą spełniać teorie T_1 i T_2, aby można było stwierdzić, że teoria T_2 została zredukowana do T_1. Pokaż, że warunki te są spełnione w wypadku redukcji termodynamiki do mechaniki statystycznej.

8. W jaki sposób można zdefiniować w mechanice statystycznej ciepło przekazane danemu ciału?

9. Jak definiujemy makrostany i mikrostany danego układu fizycznego? Jaka jest między nimi zależność? Przedstaw reprezentację makrostanów i mikrostanów w przestrzeni fazowej.

10. Omów argument z odwracalności Loschmidta i przypadek demona Maxwella. Co dokładnie pokazują te argumenty?

11. Jaka jest różnica między przestrzenią fazową a przestrzenią μ? Wyjaśnij pojęcia dystrybucji i aranżacji w odniesieniu do komórek w przestrzeni μ. Jakiemu elementowi przestrzeni fazowej odpowiada dana dystrybucja w przestrzeni μ?

12. Przedstaw statystyczny argument za twierdzeniem, że makrostany w przestrzeni fazowej będą się od siebie istotnie różnić rozmiarami. Jaka jest charakterystyczna cecha wyróżnionego makrostanu, który zajmuje największą objętość w przestrzeni fazowej?

13. Dokonaj krytycznej analizy argumentu Boltzmanna za tezą, że istnieje ogromne prawdopodobieństwo, że entropia danego układu będzie rosła w czasie. Jakie są podstawowe zarzuty wobec tego argumentu?

14. Porównaj hipotezę fluktuacji Boltzmanna z hipotezą o małej entropii początkowej wszechświata. Która z tych hipotez lepiej wyjaśnia fakt występowania dodatniego gradientu entropii w obecnym stanie wszechświata?

Literatura uzupełniająca

Fascynująca historia rozwoju pojęcia temperatury i technik jej mierzenia, połączona z głęboką filozoficzną i metodologiczną analizą, przedstawiona jest w książce: H. Chang, *Inventing Temperature: Measurement and Scientific Progress,* Oxford University Press, Oxford 2004.

Znaczna część niniejszego rozdziału oparta została na przystępnym, lecz dogłębnym omówieniu relacji między termodynamiką a mechaniką statystyczną w książce: D. Albert, *Time and Chance*, Harvard University Press, Cambridge, Mass. 2000.

Popularne omówienie pojęć termodynamicznych, w tym entropii, można znaleźć w przetłumaczonej na język polski książce: R. Penrose, *Nowy umysł cesarza*, PWN, Warszawa 1995, rozdział 7.

Następująca praca australijskiego filozofa zawiera analizę problemu strzałki czasu w termodynamice i innych teoriach fizycznych: H. Price, *Strzałka czasu i punkt Archimedesa*, Amber, Warszawa 1998.

Bardzo wnikliwą i obszerną analizę filozoficznych problemów mechaniki statystycznej (uwzględniającą zagadnienie redukcji teorii naukowych) przedstawia praca: L. Sklar, *Physics and Chance*, Cambridge University Press, Cambridge 1993.

ROZDZIAŁ 4. ELEKTRYCZNOŚĆ I MAGNETYZM

W poprzednim rozdziale rozważaliśmy zjawiska cieplne, które choć z pozoru odmienne od procesów mechanicznych (ruchu, zderzeń itp.), po bliższym zbadaniu okazały się redukowalne do tych ostatnich. Obecnie zajmiemy się inną kategorią zjawisk fizycznych, dla których taka redukcja do procesów o charakterze czysto mechanistycznym jest niemożliwa. Mowa tutaj o oddziaływaniach elektrycznych i magnetycznych. Co więcej, nauka dokonała w tym przypadku zwrotu o sto osiemdziesiąt stopni, w istocie rzeczy redukując zjawiska czysto mechaniczne właśnie do elektromagnetyzmu. Na przykład fundamentalna m.in. dla kartezjan forma oddziaływania mechanicznego przez kontakt podczas zderzenia dwóch ciał materialnych faktycznie okazuje się oddziaływaniem elektrostatycznym na odległość między powłokami elektronowymi atomów wchodzących w skład tych ciał. Nieprzenikliwość ciał materialnych, stanowiąca podstawę obrazu świata mechaniki klasycznej, jest konsekwencją istnienia ogromnych sił odpychających między naładowanymi cząstkami. Gdyby elektrony na powłokach atomowych nie oddziaływały z siłą rosnącą do nieskończoności wraz ze zbliżaniem się do siebie, ciała materialne mogłyby przenikać się jak duchy, skoro większość przestrzeni pomiędzy elektronami, protonami i neutronami zawartymi w atomach jest praktycznie pusta.

Teoria elektromagnetyzmu jest istotna dla filozofa z wielu powodów. Po pierwsze, wprowadza do opisu świata nowe ontologiczne pojęcie – pojęcie pola fizycznego. Fundamentalną ontologią mechaniki klasycznej jest ontologia ciał materialnych – nieprzenikliwych obiektów obdarzonych masą o dobrze określonych lokalizacjach czasowych i przestrzennych. Co prawda oddziaływania grawitacyjne między ciałami materialnymi mogą być opisane za pomocą pojęcia pola grawitacyjnego, jednakże opis ten ma charakter skrótu, ułatwiającego formułowanie odpowiednich praw. W istocie nie musimy zakładać ontologicznej realności pola grawitacyjnego – wystarczy powiedzieć, że istnieją jedynie ciała, które przyciągają się wzajemnie z odpowiednią siłą. W wypadku zjawisk elektromagnetycznych sytuacja jest inna, gdyż pewne prawa teorii elektromagnetyzmu jawnie wymagają zastosowania pojęć pola elektrostatycznego i magnetycznego oraz ich wzajemnych powiązań. To prowadzi do dalszych pytań o naturę pól – rozciągłych i wzajemnie przenikających się obiektów wypełniających przestrzeń, o nieostrych granicach czasoprzestrzennych.

Teoria elektromagnetyzmu jest także ważna dla historyka i metodologa nauki. Okazuje się, że zjawiska elektryczne i magnetyczne są ze sobą powiązane w taki sposób, iż można

dokonać unifikacji tych pozornie odmiennych aspektów w jeden obraz rzeczywistości, posługując się pojęciem pola elektromagnetycznego. Stawia to nas przed pytaniem, co to znaczy zunifikować dwie teorie, dwa opisy albo dwa typy zjawisk. Czy wystarczy do tego podanie zestawu praw łączących te zjawiska, jak to ma miejsce w wypadku tzw. praw elektromagnetyzmu Maxwella, czy też potrzeba czegoś więcej? Innym metodologicznym aspektem teorii elektryczności i magnetyzmu w wersji podanej przez Maxwella jest jej fenomenalny sukces w postaci sformułowania hipotezy istnienia fal elektromagnetycznych, która została następnie potwierdzona doświadczalnie oraz wykorzystana w praktyce do celów komunikacji. Tego typu epizody są rzadkością w historii nauki, warto zatem przyjrzeć się dokładniej, w jaki sposób Maxwell wyprowadził ze swoich równań tak daleko idącą konsekwencję.

Na tym nie kończy się lista filozoficznie ważkich aspektów zjawisk elektromagnetycznych. Powszechnie wiadomo, że pewne teoretyczne problemy w obrębie teorii Maxwella dały impuls do sformułowania radykalnie nowatorskiej koncepcji czasu, przestrzeni i ogólnie procesów mechanicznych w formie Einsteinowskiej szczególnej teorii względności. Teoria Maxwella nie zachowuje się prawidłowo przy przejściu z jednego inercjalnego układu odniesienia do drugiego, jak tego wymaga zasada względności Galileusza, omówiona w rozdziale 1. Rozwiązaniem tego problemu może być albo przyjęcie, że równania elektromagnetyzmu obwiązują jedynie w pewnym wyróżnionym układzie odniesienia, definiującym absolutny spoczynek, albo też modyfikacja reguł transformacji współrzędnych z jednego układu odniesienia do drugiego. Tę drugą opcję wybrał Einstein, akceptując zasady transformacji Lorentza, zgodnie z którymi czas i przestrzeń nie są już absolutne jak w fizyce klasycznej, lecz zależą od przyjętego układu odniesienia. Zatem gdyby nie analiza zjawisk elektromagnetycznych, prawdopodobnie nigdy nie dowiedzielibyśmy się, że czas i przestrzeń mają inną strukturę niż mówi nam doświadczenie potoczne. Wreszcie nie do pominięcia jest fakt, że teoria elektromagnetyzmu wykorzystuje piękną matematykę, której walory zarówno poznawcze, jak i estetyczne są godne uwagi filozofa.

4.1. Elektrostatyka i pola

Już starożytni zauważyli, że pocierając pewne substancje, np. bursztyn o sukno, można spowodować, że potarte przedmioty zaczną na siebie oddziaływać siłami przyciągania bądź też odpychania. Dzisiaj wiemy, że owe zjawiska można wytłumaczyć istnieniem dwóch rodzajów ładunków, przyjmując, że ładunki tego samego rodzaju się odpychają, a ładunki przeciwne – przyciągają. Wydaje się naturalne, że istnienie dwóch rodzajów ładunków elektrycznych wynika bezpośrednio z faktu występowania dwóch typów sił elektrostatycznych. Dla porównania: w wypadku grawitacji istnieje dokładnie jeden typ „ładunku" grawitacyjnego, czyli masa, gdyż siła grawitacji jedynie przyciąga. Jednakże istnienie przyciągających i odpychających sił elektrostatycznych jeszcze nie gwarantuje istnienia dokładnie dwóch rodzajów ładunku – do wyprowadzenia tego wniosku potrzebne są dodatkowe dane doświadczalne. Teoretycznie mogłoby się okazać, że w przyrodzie występuje pięć albo nawet dwadzieścia rodzajów ładunków, z których każdy rodzaj oddziaływałby tylko na siebie. Jeśli interesuje Was, jakie empiryczne fakty pozwalają na uzasadnienie istnienia dokładnie dwóch rodzajów ładunków, zajrzyjcie do tekstu w ramce.

Doświadczenie poucza nas, że oddziaływanie między naładowanymi ciałami ma następującą własność: jeśli jakieś ciało odpycha inne, a to inne z kolei odpycha jeszcze inne,

to zbliżenie pierwszego z trzecim spowoduje również powstanie siły odpychania. Własność tę nazywamy w języku logiki „przechodniością". Oznacza to, że ciała uczestniczące w zjawiskach odpychania możemy poklasyfikować w grupy, z których każda będzie obejmować wszystkie i tylko te obiekty, które wzajemnie się odpychają. (Ściśle rzecz biorąc, potrzebujemy do takiej klasyfikacji jeszcze założenia, że relacja odpychania jest symetryczna – to jest dość oczywiste – oraz zwrotna. Z tym ostatnim jest pewien problem, gdyż zwrotność oznacza, że przedmiot odpycha się sam ze sobą, co jest ewidentnie fałszem. Jest to jednak trudność czysto formalna, gdyż możemy po prostu „dodać" do relacji odpychania relację tożsamości, a tak uzyskana relacja będzie już pełnoprawną relacją równoważnościową, która umożliwia dzielenie obiektów na podgrupy.)

Zasadniczo takich grup może być dowolnie dużo, a zatem także liczba różnych ładunków mogłaby być dowolna. Jednakże musimy uwzględnić jeszcze fakty dotyczące relacji przyciągania. Przede wszystkim wiemy, że jeśli ciało *a* przyciąga się z ciałem *b*, a *b* przyciąga z ciałem *c*, to ciała *a* i *c* będą się odpychać (to drugi empiryczny fakt po przechodniości relacji odpychania). Wreszcie podajmy trzeci istotny fakt: każde dwa dowolne ciała z grupy „odpychaczy" (tj. takie, które odpychają się z jakimiś innymi ciałami – czyli po prostu ciała naładowane elektrycznie) albo będą się wzajemnie przyciągać, albo odpychać, przy czym niektóre z nich się przyciągają.

Z trzech powyżej wymienionych faktów możemy już wyprowadzić wniosek, że istnieją dokładnie dwie grupy „odpychaczy" takie, że w obrębie każdej z nich każdy obiekt odpycha się z każdym, natomiast obiekty z różnych grup się przyciągają. Wiemy na pewno, że istnieją co najmniej dwie klasy zamknięte ze względu na relację odpychania, gdyż niektóre z „odpychaczy" wzajemnie się przyciągają, nie mogą więc należeć do jednej kategorii. Załóżmy zatem, że moglibyśmy wyróżnić więcej niż dwie takie grupy – nazwijmy trzy z nich A, B i C. Z założenia elementy grupy A przyciągają się z elementami grupy B, a elementy grupy B przyciągają z elementami grupy C. Ale z drugiego z wcześniej wymienionych faktów wynika, że w takiej sytuacji elementy grupy A będą się odpychać z elementami grupy C, a to jest w sprzeczności z założeniem, że grupy A, B i C są zamknięte ze względu na relację odpychania. Zatem mogą być tylko dwie takie grupy. Jedną z nich umownie nazywamy obiektami naładowanymi dodatnio, a drugą naładowanymi ujemnie.

Ilościowy opis przyciągania i odpychania elektrostatycznego zawarty jest w tzw. prawie Coulomba, które jest analogiczne do prawa powszechnego ciążenia Newtona. Prawo Coulomba głosi, że siła oddziaływania elektrostatycznego między dwoma ładunkami punktowymi jest proporcjonalna do ich ładunków q_1 i q_2 i odwrotnie proporcjonalna do kwadratu odległości r między nimi:

$$F = k\frac{q_1 q_2}{r^2}. \tag{4.1}$$

Stała k jest dużo większa od stałej grawitacji G, a zatem także siły oddziaływania elektrostatycznego są nieporównanie większe od sił grawitacji. Oddziaływanie grawitacyjne między dwoma naładowanymi elektrycznie ciałami o średniej wielkości można praktycznie pominąć. Jeśli zaś chodzi o same ładunki elektryczne, to doświadczenie zdaje się sugerować, że jedno i to samo ciało fizyczne może posiadać różną wartość ładunku w zależności od stanu naładowania elektrycznego – od wartości równej zeru (ciało neutralne elektrycznie) do arbitralnie dużej wartości. Byłaby to zatem kolejna różnica między elektrostatyką a grawitacją, gdyż masa ciała jest jego niezmienniczą własnością. (Oczywiście dane ciało może utracić

część swojej masy, ale dzieje się to kosztem utraty części jego materii, a zatem można argumentować, że mamy wtedy do czynienia już z innym przedmiotem fizycznym.) Rozwój nauki pokazał, że nasze intuicyjne podejście do ładunku elektrycznego danego ciała jako własności mogącej ulegać zmianie jest błędne. Dzisiaj wiemy, że naelektryzowanie ciała może się odbyć jedynie poprzez dostarczenie lub odjęcie mu pewnej liczby elementarnych składników materii obdarzonych stałym ładunkiem. W normalnych okolicznościach liczba dodatnio naładowanych cząstek (protonów) równoważy się z liczbą cząstek ujemnych (elektronów). Jeśli ciało ma nadmiarową liczbę elektronów, stanie się naładowane ujemnie, a jeśli zostanie mu ujęta ich pewna liczba, wypadkowy ładunek będzie dodatni. Jednakże ładunek pojedynczego elektronu czy protonu nie może ulec zmianie – nie da się „naelektryzować" cząstek elementarnych. Pod tym względem ładunek nie różni się zasadniczo od masy, która również nie może ulec zmianie dla danej cząstki.[1]

Dokonajmy teraz niewielkiej transformacji wzoru (4.1) na siłę Coulomba:

$$F = q_2 \left(k \frac{q_1}{r^2} \right).$$

Wyrażenie $k \frac{q_1}{r^2}$ w nawiasie nazwiemy wartością (natężeniem) pola elektrycznego E pochodzącego od ładunku q_1 w punkcie gdzie znajduje się ładunek q_2. (Oczywiście należy pamiętać, że pole elektryczne w danym punkcie jest wyrażane przez wektor, mający kierunek i zwrot zgodny z kierunkiem i zwrotem siły, jaka działałaby na dodatni ładunek umieszczony w tym miejscu.) Gdybyśmy teraz usunęli ładunek q_2 (tj. gdybyśmy położyli $q_2 = 0$), to oczywiście siła F by zniknęła. Natomiast wartość $E = k \frac{q_1}{r^2}$ pozostałaby nadal niezerowa. To sugeruje, że nawet w wypadku „samotnego" ładunku elektrycznego, w punkcie, w którym mógłby się znajdować drugi ładunek, istnieje pewnego rodzaju realność fizyczna. Ponieważ każdy punkt w przestrzeni mógłby być zajęty przez drugi ładunek q_2, mówimy o polu elektrycznym od ładunku q_1 w całej otaczającej go przestrzeni.

Powstaje jednak pytanie, czy sam fakt występowania pewnego niezerowego czynnika we wzorze na siłę Coulomba wystarcza, aby uznać istnienie nowego bytu, jakim jest pole elektrostatyczne. Zasadniczym argumentem przeciwko postulowaniu istnienia pól może być argument o charakterze empirycznym: samego pola elektrycznego nie jesteśmy w stanie zaobserwować, dopóki nie umieścimy w nim pewnego ładunku (taki ładunek nazywamy próbnym) i nie stwierdzimy, że zacznie on przyspieszać pod wpływem działającej siły. To jednak nie dowodzi, że w danym punkcie była pewna realność fizyczna, zanim umieściliśmy tam ładunek próbny. Doświadczenie wskazuje na realność siły działającej między dwoma ładunkami, natomiast status pola jest dużo bardziej problematyczny.

W filozofii napotykamy podobne sytuacje przy okazji analizy własności zwanych dyspozycyjnymi. Klasycznym przykładem własności dyspozycyjnej jest kruchość przedmiotów (np. szklanych), która nie jest obserwowalna, dopóki nie zadziałamy odpowiednią siłą na przedmiot, na przykład uderzając go młotkiem. Przejawem kruchości jest rozbicie się przedmiotu pod wpływem uderzenia. Czy natomiast kruchość istnieje także przed uderzeniem? Zdania są tutaj podzielone. Niektórzy uważają, że twierdzenie na temat kruchości materiału nie ma sensu empirycznego, dopóki materiał nie zostanie poddany odpowiedniej próbie. Większość jednak jest zdania, że kruchość istnieje jako pewnego rodzaju potencjalność, na-

[1] Kwestią zasady zachowania ładunku (i masy) zajmiemy się nieco dokładniej w rozdziale poświęconym szczególnej teorii względności.

wet jeśli nigdy nie zostanie ujawniona. Podobnie rzecz się ma z istnieniem pola – można uważać, że jego obecność jest tożsama z możliwością uzyskania odpowiedniego obserwowalnego efektu podczas umieszczenia ładunku próbnego.

Radykalnie nastawieni empiryści skłaniają się ku temu, aby traktować pola fizyczne czysto instrumentalnie jako wygodne narzędzie, któremu nie odpowiada żadna realność. Z pojęciem instrumentalizmu zetknęliśmy się już przy okazji omawiania interpretacji teorii astronomicznych. Instrumentalistyczne podejście do modelu Ptolemeuszowskiego czy Kopernikańskiego polega na tym, że traktujemy teoretyczne elementy tych modeli (np. epicykle i deferensy u Ptolemeusza czy orbity kołowe u Kopernika) jedynie jako użyteczne narzędzia do obliczania położeń planet na niebie. Podobnie w wypadku pól, ich użyteczność polega na przewidywaniu zachowania obiektów umieszczonych w pobliżu innych oddziałujących z nimi ciał. Ostatecznie jednak realność przysługuje jedynie ciałom fizycznym i ich oddziaływaniom.

Jeden z możliwych argumentów w debacie na temat realności pól fizycznych odwołuje się do pojęcia nielokalności czy też działania na odległość. Jeśli przyjmiemy stanowisko instrumentalistyczne w kwestii istnienia pól, to w konsekwencji musimy przyjąć, że oddziaływanie między naładowanymi ciałami zachodzi na odległość bez żadnego fizycznego pośrednictwa, ze złamaniem intuicyjnej zasady lokalności. (W ramce poniżej możecie przeczytać więcej na temat możliwych interpretacji pojęcia „nielokalności".) Ładunek próbny umieszczony w pewnej odległości od innego ładunku „czuje" jego obecność mimo dzielącego ich dystansu. Założenie istnienia fizycznego pośrednika w postaci pola pozwala na zachowanie intuicji lokalności. W takim ujęciu bezpośrednią przyczyną powstania siły działającej na ładunek próbny (i w konsekwencji przyspieszenia ładunku) jest fizyczne pole obecne w miejscu lokalizacji tego ładunku. Z kolei wartość natężenia pola w danym miejscu jest powiązana z wartością natężenia w obszarach bliższych ładunkowi źródła pola. Istnieje substancjalna ciągłość pomiędzy źródłem pola a miejscem lokalizacji ładunku próbnego, która to ciągłość jest ontologiczną podstawą łańcucha przyczynowo-skutkowego prowadzącego od źródła pola do ładunku odczuwającego działanie siły. Relacja przyczynowa nie zachodzi „skokowo": między każdymi dwoma elementami łańcucha przyczynowo-skutkowego istnieje przyczyna pośrednia.

Pojęcie lokalności i jego negacja – nielokalność – mogą być interpretowane wielorako. W sensie, który może być nazwany czasowym, związek przyczynowy między odległymi przestrzennie zdarzeniami *A* i *B* jest nielokalny, gdy *A* i *B* zachodzą w tym samym czasie. Innymi słowy, nielokalne oddziaływanie przyczynowe między *A* i *B* rozchodzi się z nieskończoną prędkością, bez żadnego opóźnienia. Z kolei interpretacja oparta na pojęciu „nieciągłości przestrzennej" charakteryzuje związek przyczynowy między *A* i *B* jako nielokalny, gdy w obszarze przestrzennym między *A* i *B* nie zachodzą żadne przyczyny pośrednie – tj. nie ma takiego zdarzenia *C* między *A* i *B*, które byłoby skutkiem *A* i zarazem przyczyną *B*. Dodatkowo, jeśli *B* zachodzi później niż *A*, można mówić także o nielokalności opartej na nieciągłości czasowej: w interwale pomiędzy zachodzeniem *A* i *B* nie zachodzi żadne *C*, takie że *A* jest przyczyną *C* i *C* jest przyczyną *B*. Związki przyczynowe, które są nielokalne w tym ostatnim sensie, jak się wydaje, łamią zasadę determinizmu, gdyż stan całego wszechświata w interwale między *A* i *B* nie wyznacza tego, że *B* zajdzie (można wyobrazić sobie dwa światy, które są identyczne w dowolnym punkcie czasowym między *A* i *B* i takie, że w jednym z nich zdarzenia *A* i *B* zachodzą, a w drugim nie).

Nielokalność oparta na braku ciągłości przestrzennej jest konsekwencją odrzucenia realności pól fizycznych. Jeśli do tego przyjmiemy, że oddziaływania elektrostatyczne nie są nielokalne w sensie czasowym (tj. istnieje opóźnienie czasowe między przyczyną – źródłem pola – a skutkiem w postaci siły działającej na ładunek próbny), to w rezultacie musimy przyjąć, że zachodzi także nielokalność w sensie nieciągłości czasowej, a zatem również pewnego rodzaju indeterminizm. Indeterminizmu tego można uniknąć, jeśli założy się, że oddziaływania elektrostatyczne rozchodzą się z nieskończoną prędkością. Problem nielokalności nabiera szczególnej wagi w kontekście zjawiska splątania kwantowego, o którym będziemy mówić w dalszych częściach książki.

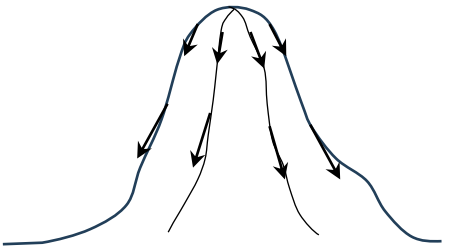
4.2. Potencjał i linie sił

Rozważając ontologiczny problem realności pól fizycznych, warto przywołać przykłady sytuacji, w których sami fizycy kwestionują obiektywne istnienie korelatów dla pewnych pojęć, możliwych do wprowadzenia w formalizmie danej teorii, mimo tego, że pojęcia te oddają spore usługi. Przykładu takiego pojęcia w elektrostatyce dostarcza potencjał elektryczny. Potencjał jest wielkością skalarną (reprezentowaną przez liczbę), przy pomocy której można obliczyć wektor natężenia pola w każdym punkcie. Relacja między potencjałem a natężeniem pola wygląda następująco. Natężenie jest wektorem skierowanym w stronę największej zmiany potencjału, a wartość natężenia jest dana szybkością tej zmiany potencjału przy niewielkiej zmianie położenia x. Intuicję tę można zobrazować, korzystając z przykładu pola grawitacyjnego. Wyobraźmy sobie pagórek o różnym nachyleniu zbocza, na którym możemy umieścić obiekt, np. kulkę (rys. 4.1). W zależności od tego, jak duże jest nachylenie zbocza w danym punkcie, kulka będzie „odczuwała" różnej wielkości siłę grawitacji – od wartości zerowej na płaskim podnóżu lub na samym szczycie góry, do wartości maksymalnej w najbardziej stromej części zbocza. Efektywne natężenie pola grawitacyjnego zależy zatem od nachylenia zbocza w danym punkcie. Wysokość zbocza w punkcie jest właśnie odpowiednikiem potencjału.[2]

Matematyczna formuła obrazująca relację między potencjałem a natężeniem pola wykorzystuje pojęcie pochodnej, które, jak już wiemy, charakteryzuje prędkość zmiany danej wielkości względem odpowiedniej zmiennej. Można też posłużyć się znaną geometryczną interpretacją pochodnej jako odzwierciedlającej kąt nachylenia stycznej do danej funkcji w punkcie (dokładniej, tangens tego kąta). W przypadku, gdy potencjał jest funkcją tylko jednej zmiennej położenia (przypadek jednowymiarowy), odpowiednia relacja między potencjałem V a natężeniem pola E wygląda następująco:

$$E = -\frac{dV}{dx}.$$

[2] Ilustracja pojęcia potencjału przy pomocy przykładu góry w stałym polu grawitacyjnym pojawia się w wielu podręcznikach i książkach popularyzatorskich. Należy jednak pamiętać, że przykład ten nie jest do końca poprawny. Składowa siły grawitacji styczna do danej powierzchni nie może przekroczyć wartości mg, natomiast pochodna funkcji wysokości zbocza w danym punkcie względem współrzędnej reprezentującej odległość poziomą (czyli gradient) może przyjmować dowolnie dużą wartość. Na przykład kiedy nachylenie zbocza zbliża się do pionu, gradient rośnie do nieskończoności, zatem siła także powinna być nieskończona. Zatem wysokość zbocza w danym punkcie nie jest, ściśle rzecz biorąc, tożsama z potencjałem, a tylko ilustruje poglądowo to pojęcie.



Rys. 4.1. Ilustracja potencjału za pomocą przykładu pagórka w polu grawitacyjnym

Znak minus związany jest z tym, że siła działająca na obiekt jest skierowana w przeciwnym kierunku w stosunku do kierunku wzrostu funkcji V. W ogólnym wypadku potencjał V to funkcja trzech współrzędnych przestrzennych x, y i z. W takiej sytuacji należy obliczyć pochodne względem każdej ze zmiennych z osobna, traktując dwie pozostałe jako stałe (są to pochodne cząstkowe, wprowadzone w rozdziale 2). W ten sposób otrzymamy trzy liczby, które reprezentują trzy współrzędne wektora natężenia E_x, E_y i E_z (szybkość zmian potencjału w kierunkach x, y, i z):

$$E_x = -\frac{\partial V}{\partial x},$$

$$E_y = -\frac{\partial V}{\partial y},$$

$$E_z = -\frac{\partial V}{\partial z}.$$

Równania te można uprościć, wprowadzając operację tzw. gradientu wielkości skalarnej. który jest oparty na następującym operatorze wektorowym, zwanym nabla (trzy wielkości w nawiasie reprezentują jego trzy składowe):

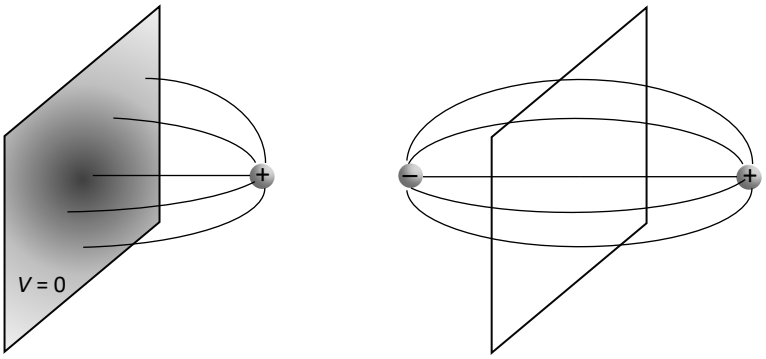
$$\boldsymbol{\nabla} = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right),$$

$$\boldsymbol{E} = -\boldsymbol{\nabla} V. \tag{4.2}$$

Wytłuszczony druk oznacza, że natężenie pola $\boldsymbol{E}$ jest wektorem (ma trzy składowe, określające jego kierunek i zwrot w przestrzeni). Operator nabla $\boldsymbol{\nabla}$, zastosowany do dowolnej wielkości skalarnej, tworzy wektor.

Pojęcie potencjału elektrycznego odgrywa bardzo ważną rolę przy rozważaniu problemów z elektrostatyki. Podstawowym zagadnieniem elektrostatyki jest pytanie, jak obliczyć wartość natężenia pola w danym punkcie, znając rozkład ładunków elektrycznych w całej przestrzeni. Teoretycznie problem taki można rozwiązać, stosując różniczkową postać prawa Coulomba, sumując wkłady pochodzące of niewielkich (infinitezymalnych) obszarów naładowanych elektrycznie. W praktyce jednak takie zadanie jest częstokroć trudne do wykonania matematycznie. Okazuje się, że wprowadzenie pojęcia potencjału pozwala na znaczne uproszczenie rozważanych sytuacji, jeśli tylko mamy do czynienia z pewnymi symetriami.

Przykład jednego z takich problemów wraz z rozwiązaniem opartym na pojęciu potencjału jest podany w ramce (zastosowana metoda nosi nazwę metody obrazów).



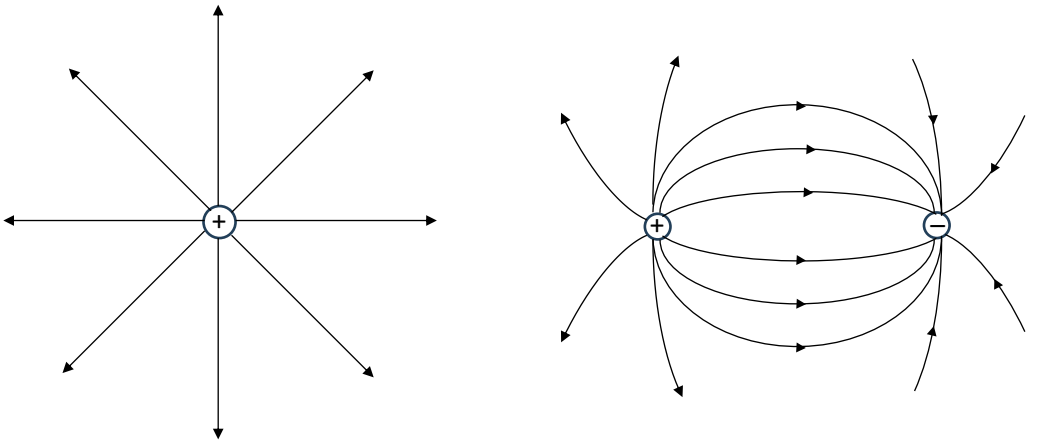
Rys. 4.2. Zastosowanie metody obrazów do rozwiązania problemu z elektrostatyki

Wyobraźmy sobie następujący układ: dodatni ładunek punktowy umieszczony w pewnej odległości od nieskończonej, uziemionej płyty metalowej (rys. 4.2). Jak będzie wyglądał rozkład pola elektrycznego w tak scharakteryzowanym układzie? Jak wiadomo, Ziemia może być uważana za nieograniczone źródło wolnych elektronów. Zatem uziemienie płyty spowoduje, że napłyną do niej ładunki ujemne, przyciągane pojedynczym ładunkiem dodatnim. Płyta się naelektryzuje, lecz rozkład ładunków na płycie nie będzie jednorodny – punkt najbliżej ładunku dodatniego będzie miał największą gęstość ładunku, a dalsze obszary odpowiednio niższą gęstość. W takiej sytuacji obliczenie wypadkowego natężenia pola elektrycznego pochodzącego od wszystkich ładunków wydaje się niemożliwością. Zauważmy jednak, że potencjał elektryczny płyty jest zerowy ze względu na jego uziemienie. Oznacza to, że na ładunki na powierzchni płyty nie działają żadne siły w płaszczyźnie płyty – znajdują się one w stanie równowagi.

Teraz skorzystamy z ważnego twierdzenia, które można udowodnić w elektrostatyce. Głosi ono, że rozkład ładunków w dowolnym obszarze oraz potencjał na brzegu tego obszaru jednoznacznie wyznaczają pole elektryczne w całej objętości. To, co dzieje się poza wybranym obszarem, nie ma żadnego wpływu na wartość pola wewnątrz. Zatem możemy zastąpić rozważany problem płyty i pojedynczego ładunku innym problemem, jeśli tylko będzie się zgadzał potencjał brzegowy oraz rozkład ładunków. Takim innym problemem jest prosty dipol elektryczny, składający się z naszego ładunku dodatniego i równego, lecz przeciwnego ładunku ujemnego, umieszczonego po drugiej stronie płyty w tej samej odległości. Zauważmy, że w każdym punkcie płaszczyzny, gdzie umieszczona była płyta, potencjał elektryczny pochodzący od obu ładunków jest zerowy, gdyż odległości od ładunków są takie same, a same ładunki równe, lecz o przeciwnym znaku (potencjał od ładunku punktowego q wynosi $-k\frac{q}{r}$). Cała półprzestrzeń „po prawej stronie” jest więc identyczna z poprzednim przypadkiem (ten sam rozkład ładunków i ten sam potencjał brzegowy). Oczywiście obliczenie pola elektrycznego od dwóch ładunków punktowych jest już sprawą banalną, więc problem został rozwiązany.

Mimo swojej niewątpliwej użyteczności potencjał elektryczny nie jest traktowany jako obiektywnie istniejąca realność fizyczna. Głównym powodem jest to, że wartość potencjału w danym punkcie nie jest jednoznacznie wyznaczona przez wartość natężenia pola. Dla danej funkcji V spełniającej równanie (4.2) przy zadanym natężeniu pola $\boldsymbol{E}$ istnieje nieskończenie wiele alternatywnych funkcji różniących się tylko dodaniem stałej: $V' = V + c$, które również spełniają (4.2) (wynika to stąd, że pochodna stałej jest równa zeru). Używając naszego obrazowego przykładu z grawitacją, możemy powiedzieć, że wysokość pagórka może być obliczana od dowolnego poziomu – poziomu morza, podnóża góry albo każdego innego miejsca. Zmiana poziomu zerowego potencjału nie ma żadnego wpływu na obliczanie natężenia pola, a w konsekwencji na działające siły, które są jedynym namacalnym przejawem pola elektrycznego. Transformacja potencjału elektrycznego polegająca na dodaniu do niego stałej wartości, należy do klasy tzw. transformacji cechowania (ang. *gauge transformations*). Transformacje cechowania i symetrie cechowania stanowią ważny element opisu wielu teorii fizycznych.

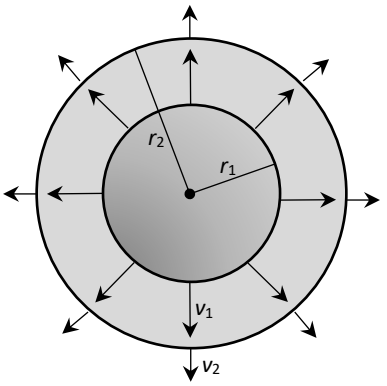
Fakt „niefizyczności" absolutnej wartości potencjału ma konkretne konsekwencje empiryczne. Na przykład naładowanie elektrostatyczne zamkniętej metalowej puszki (tzw. klatki Faradaya) nie wywołuje żadnego fizycznego skutku wewnątrz puszki. Potencjał całego układu rośnie w stosunku do potencjału Ziemi, lecz dopóki nie zetkniemy dwóch ciał o różnych potencjałach, nie zaobserwujemy żadnego efektu zwiększonego potencjału puszki. Równie dobrze można przyjąć, że potencjał puszki jest cały czas zerowy, a potencjał Ziemi maleje do odpowiednich wartości ujemnych. Sens fizyczny ma tylko różnica potencjału, nie zaś jego absolutna wartość.



Rys. 4.3. Linie sił pola elektrostatycznego

Kwestia istnienia pól elektrostatycznych reprezentowanych przez wektor natężenia nadal pozostaje otwarta. Powrócimy do niej nieco później, przy okazji omawiania praw łączących zjawiska elektryczne i magnetyczne. Na razie wspomnijmy o jeszcze jednym pomocnym pojęciu w elektrostatyce, które jednak posiada tylko status użytecznej fikcji. Chodzi tutaj o linie sił pola. Podręczniki teorii elektromagnetyzmu często przedstawiają pola elektryczne czy magnetyczne w postaci linii rozprzestrzeniających się w danym obszarze. Na przykład pole pochodzące od ładunku punktowego ma postać prostych linii wychodzących od tego ładunku

i rozciągających się do nieskończoności. Podobnie pole pochodzące od dipola (dwóch różnoimiennych ładunków elektrycznych) obrazowane jest w postaci zakrzywionych linii łączących oba ładunki (rys. 4.3). Linie sił reprezentują tory cząstek próbnych o pomijalnej masie, umieszczonych w danym polu. Można także przyjąć zasadę interpretacyjną, zgodnie z którą gęstość linii w danym obszarze reprezentuje wartość natężenia pola w tym obszarze. Na przykład w wypadku pola pochodzącego od ładunku punktowego widać, że linie sił „rozbiegają się" wraz ze wzrostem odległości, co odzwierciedla fakt, że natężenie pola słabnie, im dalej znajdujemy się od jego źródła.



Rys. 4.4. Model wypływu cieczy przez koncentryczne sfery

Ta ostatnia jakościowa obserwacja może posłużyć do wyprowadzenia ścisłej matematycznej formuły, opisującej jak natężenie pola maleje ze wzrostem odległości, czyli prawa Coulomba. Wyobraźmy sobie, że linie sił pola reprezentują wypływ pewnej nieściśliwej cieczy, a prędkość przepływu tej cieczy w danym punkcie odpowiada natężeniu pola. Rozpatrzmy dwie koncentryczne sfery o promieniach r_1 i r_2 z ładunkiem umieszczonym w centrum (rys. 4.4). Ze względu na symetrię układu prędkość wypływu cieczy powinna być taka sama w każdym punkcie danej sfery. Ilość cieczy przepływającej przez sferę w jednostce czasu jest równa powierzchni sfery pomnożonej przez prędkość. Można to uzasadnić tym, że objętość warstwy cieczy przepływającej przez sferę w krótkim czasie Δt jest równa grubości tej warstwy $v\Delta t$ razy powierzchnia S, a zatem na jednostkę czasu przepływa Sv cieczy. Jednakże ze względu na nieściśliwość cieczy, która nie jest uzupełniana ani usuwana, jej ilość, przepływająca przez każdą powierzchnię zamkniętą, musi być stała, a zatem mamy $S_1v_1 = S_2v_2$. Ponieważ powierzchnia sfery jest proporcjonalna do kwadratu jej promienia, mamy zależność $\frac{v_1}{v_2} = \frac{r_2^2}{r_1^2}$, czyli prędkość maleje wraz z kwadratem odległości, dokładnie tak, jak to wynika z prawa Coulomba.

Jednakże fakt, że prawo Coulomba można wyprowadzić z założenia o istnieniu „fluidu elektrostatycznego" nie dowodzi, że taki fluid istnieje naprawdę. Do uzasadnienia tezy o istnieniu obiektu fizycznego potrzebne są niezależne, potwierdzające ją świadectwa. Domniemany fluid nie wywołuje żadnych obserwowalnych efektów, które mogłyby niezależnie potwierdzić jego istnienie. Podobnie rzecz się ma z pojęciem linii sił – trudno wyobrazić sobie, że w przestrzeni istnieją nieskończenie cienkie „włókienka", wybiegające z jednych ładunków i zbiegające do innych. Mimo to wielu fizyków dziewiętnastowiecznych (w tym fizyk

angielski Michael Faraday) wierzyło w obiektywne istnienie linii sił. Faraday wykorzystał to pojęcie do wyjaśnienia zjawiska indukcji elektromagnetycznej oraz zjawiska oddziaływania poruszających się elektrycznych ładunków z polem magnetycznym (będziemy mówić o tych zjawiskach w kolejnych paragrafach). Wprowadził on ideę, że przecinanie linii sił pola magnetycznego przez przewodnik wywołuje fizycznie efekt przepływu prądu w tym przewodniku, co zakładało obiektywną realność linii. Jednakże okazało się, że teoria Faradaya linii sił była niespójna. Wymagała założenia, że linie sił pola wytworzonego przez magnes sztabkowy poruszają się wraz z nim, gdy przesuwa się on w danym kierunku, natomiast kiedy magnes się obraca, linie sił pozostają stacjonarne.[3] Ponieważ trudno pogodzić ze sobą te dwa założenia, idea Faradaya została porzucona.

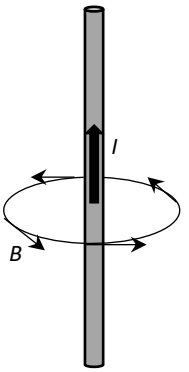
4.3. Magnetyzm

Zjawiska magnetyczne są znane ludzkości co najmniej od tak dawna jak zjawiska elektrostatyczne. Wiemy, że pewne rodzaje minerałów mają własność przyciągania metalowych przedmiotów. Wiemy również, że między magnesami mogą występować zarówno siły przyciągania, jak i odpychania. Wydawać by się mogło, że jest tutaj pełna analogia z oddziaływaniami przedmiotów naładowanych elektrycznie, a zatem powinniśmy również wprowadzić dwa rodzaje ładunków magnetycznych. Jednakże sprawa nie jest taka prosta. Naturalnie występujące magnesy mają zawsze formę dwubiegunową: przystawienie dwóch magnesów z jednej strony powoduje ich przyciąganie, a obrócenie jednego z nich w stosunku do drugiego skutkuje siłą odpychającą. Podzielenie pojedynczego magnesu na dwie części tworzy dwa nowe magnesy, które znów posiadają dwa bieguny. Nie jesteśmy w stanie „wyekstrahować" pojedynczych ładunków magnetycznych, które zawsze podlegałyby bądź działaniu sił przyciągających, bądź odpychających, niezależnie od ich ustawienia w przestrzeni. Z punktu widzenia dzisiejszej wiedzy takie zachowanie magnesów nie jest niczym niezwykłym – pole pochodzące od naturalnych magnesów wytwarzane jest nie przez pojedyncze ładunki, a przez niewielkie przepływy prądowe w całej objętości ferromagnetyka. Jednakże trwało dosyć długo, zanim nauka doszła do właściwej interpretacji zjawisk magnetycznych. W niniejszym zwięzłym opisie pominiemy aspekty historyczne i przejdziemy od razu do przedstawienia podstawowych praw magnetyzmu w formie, w jakiej są one akceptowane do dnia dzisiejszego.

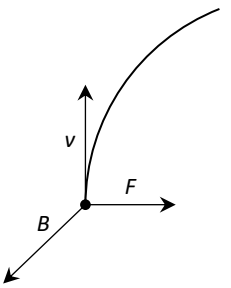
Zacznijmy od przedstawienia związków między polami magnetycznymi a poruszającymi się ładunkami elektrycznymi. Doświadczenie poucza nas, że pole magnetyczne można wywołać przez przepływ ładunków, czyli prąd elektryczny. Powszechnie znany jest fakt, że zbliżając igłę magnetyczną do przewodnika z prądem, zaobserwujemy jej obrót. Magnesy w pobliżu przewodnika z prądem układają się koncentrycznie. Sugeruje to, że wektor pola magnetycznego w danym punkcie jest styczny do okręgu przechodzącego przez ten punkt ze środkiem w miejscu przewodnika (rys. 4.5). Ilościowy opis tego zjawiska zawarty jest w prawie Biota-Savarta. Podaje ono formułę określającą kierunek i zwrot wektora natężenia pola magnetycznego $\boldsymbol{B}$, pochodzącego od niewielkiego odcinka przewodnika z prądem, przez który przepływa prąd o danym natężeniu. Wektor pola magnetycznego jest prostopadły do

[3] Szczegóły koncepcji Faradaya wraz z jej krytyką można znaleźć w książce M. Langego, *An Introduction to the Philosophy of Physics. Locality, Fields, Energy and Mass*, Blackwell, Oxford 2002, s. 51-59.

płaszczyzny zawierającej przewodnik oraz promień łączący przewodnik z danym punktem, natomiast jego wartość jest proporcjonalna do natężenia prądu, a odwrotnie proporcjonalna do kwadratu odległości (jeżeli rozważany przewodnik jest nieskończony, wartość pola jest odwrotnie proporcjonalna do samej odległości).



Rys. 4.5. Pole magnetyczne wokół przewodnika z prądem



Rys. 4.6. Siła Lorentza działająca na ładunek w polu magnetycznym

Z doświadczenia wiemy również, że pole magnetyczne wpływa na ładunki elektryczne. Oddziaływanie to jednak wygląda inaczej niż w wypadku pola elektrycznego. Po pierwsze, warunkiem koniecznym oddziałania ładunku z polem magnetycznym jest pozostawanie tego ładunku w ruchu – stacjonarne ładunki elektryczne nie „odczuwają" pola magnetycznego. Siła pochodząca od pola magnetycznego zawsze działa w kierunku prostopadłym do kierunku ruchu cząstki, zatem pole magnetyczne nie może zmienić wartości prędkości ciała, a tylko kierunek jego ruchu (ruch taki odbywać się będzie po okręgu – rys. 4.6). Po drugie, kierunek działania siły magnetycznej jest również prostopadły do wektora pola magnetycznego $\boldsymbol{B}$. Wreszcie, wartość siły magnetycznej zależy od ułożenia wektora prędkości naładowanego ciała w stosunku do kierunku pola magnetycznego – jest ona największa, kiedy prędkość jest prostopadła do linii sił pola, a spada do zera w przypadku położenia równoległego. Wszystkie te fakty można ująć syntetycznie w postaci wzoru na magnetyczną siłę Lorentza:

$$\boldsymbol{F} = \frac{q}{c}(\boldsymbol{v} \times \boldsymbol{B}).$$

Symbol × oznacza iloczyn wektorowy, który produkuje wektor prostopadły do obu mnożonych wektorów, a jego długość jest dana formułą $vB \sin\alpha$, gdzie α jest kątem między wektorami $\boldsymbol{v}$ i $\boldsymbol{B}$. Litera c oznacza pewną stałą, o której jeszcze będziemy mówić.

Jak widać z powyższego, pole magnetyczne jest ściśle związane ze zjawiskiem ruchu – zarówno tworzenie pola magnetycznego, jak i jego oddziałanie na ładunki wymaga poruszania się ciał. Jest to dla nas sygnał, że zjawiska magnetyczne (szerzej elektromagnetyczne) mogą mieć wpływ na interpretację zasady względności Galileusza, która w największym skrócie mówi, iż ruch jest względny. Wydaje się, że mamy tutaj do czynienia z pewnym problemem. Jeśli każdy inercjalny układ odniesienia jest równie dobry do opisu zjawisk fizycznych, to dziwne, że przechodząc z układu odniesienia, w którym ładunek elektryczny się porusza, do układu, w którym ten ładunek spoczywa, możemy spowodować „wyłączenie" działania siły magnetycznej. Taka konsekwencja jest niepokojąca, gdyż albo implikuje, że występowanie siły nie jest obiektywnym, realnym faktem, albo też sugeruje, że istnieją pewne wyróżnione układy odniesienia, w jakich zjawiska fizyczne zachodzą inaczej niż w pozostałych. Do tej sprawy będziemy musieli powrócić po omówieniu nomologicznych związków między polami magnetycznymi i elektrycznymi.

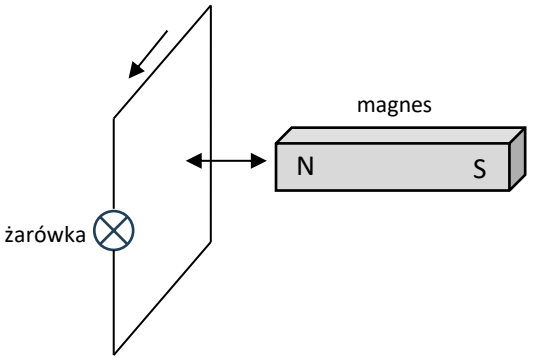
Wspomnijmy jeszcze na koniec kwestię potencjału magnetycznego. Jak pamiętamy z poprzedniego paragrafu, stacjonarne pole elektryczne $\boldsymbol{E}$ może być alternatywnie przedstawione za pomocą skalarnej wielkości V, której szybkość zmiany w przestrzeni i kierunek maksymalnej zmiany wyznaczają pole $\boldsymbol{E}$. Okazuje się, że dla pola magnetycznego $\boldsymbol{B}$ nie istnieje analogiczne skalarne pole potencjału. Ten fakt ma głębsze uzasadnienie fizyczne. Obecność skalarnego potencjału oznacza, że pole scharakteryzowane w ten sposób ma możliwość wykonania pracy. Kulka staczająca się z wierzchołka góry nabywa coraz to większej prędkości, a zatem także energii kinetycznej, kosztem energii potencjalnej w polu grawitacyjnym. Natomiast widzieliśmy już, że pole magnetyczne jest „niezdolne" do wykonania pracy nad ładunkiem elektrycznym. Ładunek poruszający się w polu magnetycznym zmienia swój kierunek ruchu, ale jego wartość prędkości – czyli także energia kinetyczna – pozostaje niezmieniona.

Istnieje jednak pewien odpowiednik potencjału elektrostatycznego, czyli tak zwany wektorowy potencjał magnetyczny $\boldsymbol{A}$. Jak sama nazwa wskazuje, jest to wektor, nie skalar. Jego relacja do wektora pola magnetycznego jest trochę bardziej skomplikowana niż w wypadku pola elektrostatycznego. Aby ją przedstawić dokładnie, trzeba mieć do dyspozycji pewne pojęcie matematyczne – pojęcie rotacji wektora – które omówimy później. Warto dodać, że wektorowy potencjał magnetyczny obciążony jest podobną wieloznacznością, co potencjał elektrostatyczny. Do danego potencjału $\boldsymbol{A}$ można dodać jakiekolwiek pole wektorowe w postaci gradientu pewnego skalara, a rezultat będzie „równie dobrym" potencjałem. Co prawda trudno ten przypadek zilustrować intuicyjnym przykładem podobnym do tego z umownym wyborem poziomu, od którego zaczynamy liczyć wysokość góry, ale w istocie jest to dokładnie taki sam rodzaj dowolności (jest to kolejny przykład tzw. swobody cechowania).[4]

[4] Należy jednak pamiętać, że w mechanice kwantowej istnieje efekt, który zdaje się sugerować, że sam potencjał magnetyczny może mieć obserwowalne skutki. Jest to tzw. efekt Aharonova-Bohma, polegający na tym, że cząstka przebiegająca w pobliży solenoidu (cewki) z prądem zmienia fazę swojej funkcji falowej, mimo że na zewnątrz solenoidu nie ma żadnego pola magnetycznego.

4.4. Strumień i krążenie pola

Stwierdziliśmy już, że zjawiska magnetyczne są powiązane z elektrycznością relacjami wzajemnego wpływu. Okazuje się, że powiązania te są jeszcze mocniejsze, a ponadto nie wymagają bezpośredniego udziału ładunków elektrycznych. Same pola elektryczne i magnetyczne pozostają ze sobą we wzajemnych relacjach przyczynowych, których szczególną naturę i elegancki opis matematyczny poznamy w niniejszym podrozdziale. Istnienie takich bliskich relacji, po pierwsze, dostarcza nam nowego, mocnego argumentu za realnością pól fizycznych, a po drugie, sugeruje, że pola elektryczne i magnetyczne nie są odrębnymi bytami, ale raczej pewnymi aspektami czy przejawami jednego bytu: pola elektromagnetycznego.



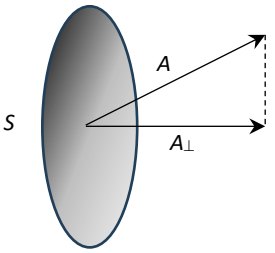
Rys. 4.7. Zjawisko indukcji elektromagnetycznej

Zacznijmy od opisu zjawiska, które jest niezwykle ważne nie tylko z teoretycznego, ale także z praktycznego punktu widzenia. Chodzi tutaj o indukcję elektromagnetyczną. Jej wystąpienie może zaobserwować każdy, kto dysponuje kawałkiem przewodu, żaróweczką i magnesem. Utwórzmy zamknięty obwód z podłączoną do niego żaróweczką. Zbliżając do tego obwodu magnes, zaobserwujemy chwilowy błysk żarówki (rys. 4.7). Kontynuując ruch magnesu względem obwodu, możemy utrzymać świecenie się żarówki, którego intensywność zależy zarówno od szybkości ruchu magnesu, jak i mocy samego magnesu. Ewidentnie ruch magnesu wzbudził powstanie prądu elektrycznego w obwodzie. To proste doświadczenie jest podstawą działania wszelkich urządzeń zamieniających energię mechaniczną na prąd elektryczny, takich jak dynama czy prądnice. Prąd dostarczany do naszych domów z elektrowni jest tworzony właśnie w ten sposób, obojętnie czy jest to elektrownia węglowa, wiatrowa, hydroelektryczna czy atomowa. Niezależnie od aspektu praktycznego, zjawisko indukcji elektromagnetycznej dało asumpt do sformułowania ważnego prawa elektromagnetyzmu, znanego jako prawo Faradaya.

Aby podać prawo Faradaya w jego pełnej formie, musimy wprowadzić pewne nowe, eleganckie pojęcia matematyczne. Są to pojęcia strumienia pola przez daną powierzchnię oraz krążenia pola wzdłuż pewnej krzywej zamkniętej. Zacznijmy od tego pierwszego. Rozważmy dowolną powierzchnię płaską zamkniętą pewną krzywą – np. powierzchnię okręgu, kwadratu albo dowolnej innej figury. Załóżmy, że w każdym punkcie tej powierzchni określony jest wektor natężenia pewnego pola $\boldsymbol{A}$ (może to być pole elektryczne, magnetyczne, ale także np. pole przepływu cieczy dane wektorem prędkości). Dla uproszczenia przyjmijmy,

że wektor $\boldsymbol{A}$ jest stały w każdym punkcie wybranej powierzchni (jego wartość, kierunek i zwrot nie ulega zmianie przy przejściu od jednego punktu do drugiego). Strumieniem pola $\boldsymbol{A}$ przez daną powierzchnię S nazwiemy iloczyn składowej wektora $\boldsymbol{A}$ prostopadłej do S (tzw. składowa normalna) i pola powierzchni S (rys. 4.8):

$$\Phi_A(S) = A_\perp S.$$



Rys. 4.8. Strumień pola przez powierzchnię

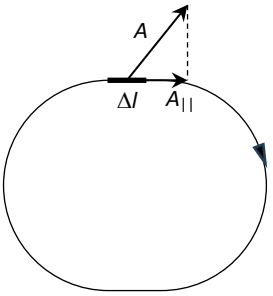
Fizyczna interpretacja strumienia jest prosta w wypadku pola prędkości przepływu. W takiej sytuacji strumień reprezentuje ilość cieczy przepływającej przez daną powierzchnię na jednostkę czasu. W wypadku innych pól, takich jak elektryczne, magnetyczne czy grawitacyjne, ta interpretacja nie może być traktowana zbyt dosłownie, gdyż, jak już wspominaliśmy, nie mamy dodatkowych argumentów za tym, że pola te oparte są naprawdę na przepływie jakiejś niewidzialnej substancji (fluidu). Jednakże intuicja przepływu jest użyteczna i można z niej korzystać w celu wizualizacji odpowiednich procesów fizycznych.

Ogólnie rzecz biorąc, nie możemy zakładać, że natężenie pola pozostaje takie samo na całej powierzchni. W takiej sytuacji musimy zastosować matematyczną metodę całkowania. Idea tej metody jest prosta, choć jej zastosowanie w konkretnych sytuacjach może być skomplikowane. Daną powierzchnię S dzielimy na drobne fragmenty ΔS, takie że w obrębie każdego fragmentu pole jest praktycznie stałe. Następnie obliczamy strumień przepływający przez każdy taki niewielki fragment i sumujemy wszystkie uzyskane w ten sposób liczby. W granicy, gdy ΔS zbiega do zera, suma poszczególnych składników „zmienia się” w całkę po powierzchni, którą zapisujemy ogólnie w następujący sposób:

$$\int_S \boldsymbol{A} \cdot d\boldsymbol{S}.$$

W powyższym zapisie zastosowaliśmy upraszczającą notację wykorzystującą iloczyn skalarny wektorów („kropka”). Symbol $d\boldsymbol{S}$ oznacza wektor prostopadły do powierzchni, którego długość jest równa polu tej powierzchni. Ważne jest to, aby wybrać jeden kierunek dla wszystkich fragmentów całkowitej powierzchni S („na zewnątrz” lub „do wewnątrz”) i konsekwentnie się jego trzymać – jest to tak zwane „zorientowanie” powierzchni. Iloczyn skalarny $\boldsymbol{A} \cdot d\boldsymbol{S}$ jest liczbą, która jest równa iloczynowi długości obu wektorów przemnożonych

przez cosinus kąta między nimi – czyli iloczyn ten odtworzy nam składową prostopadłą pola do powierzchni razy pole wybranej powierzchni.[5]



Rys. 4.9. Krążenie pola wzdłuż zorientowanej krzywej zamkniętej

Pojęcie krążenia pola jest w pewnym sensie formalnie analogiczne do pojęcia strumienia, tylko że dotyczy sytuacji jednowymiarowej, a nie dwuwymiarowej. Rozważmy dowolną krzywą zamkniętą O (może to być okrąg, kwadrat, lub jakakolwiek inna krzywa – rys. 4.9). Ustalmy najpierw pewien kierunek „obiegu" krzywej – zgodnie ze wskazówkami zegara lub przeciwnie. Następnie podzielmy krzywą na małe fragmenty o długości Δl i dla każdego takiego fragmentu obliczmy iloczyn jego długości i składowej pola $\boldsymbol{A}$ równoległej do Δl i skierowanej w stronę przyjętej orientacji krzywej (zatem rezultat będzie dodatni, jeśli składowa ma ten sam kierunek, co kierunek orientacji, a ujemny, gdy będzie on przeciwny). Na koniec zsumujmy tak uzyskane wyniki. Znów, w granicy gdy Δl dąży do zera, suma wszystkich składników przyjmie formę całki po krzywej zamkniętej:

$$\oint_O \boldsymbol{A} \cdot d\boldsymbol{l}.$$

Tak jak poprzednio, kropka oznacza iloczyn skalarny, a $d\boldsymbol{l}$ jest wektorem skierowanym w stronę orientacji krzywej. Krążenie pola to wielkość, która określa tendencję pola do wirowania. Na przykład w wypadku przepływu płynu niezerowe krążenie oznacza, że w pewnym obszarze utworzył się wir. Z drugiej strony, jeśli pole jest stacjonarne, jego krążenie wokół dowolnej krzywej zamkniętej jest równe zeru.

Możemy teraz sformułować prawo Faradaya, które jest teoretyczną podstawą zjawiska indukcji elektromagnetycznej. Zacznijmy od sformułowania słownego:

> Zmiana strumienia pola magnetycznego przez daną powierzchnię zamkniętą S skutkuje powstaniem pola elektrycznego, którego krążenie wokół krzywej zamykającej powierzchnię S jest proporcjonalne do szybkości zmiany strumienia magnetycznego przez powierzchnię S.

W powyższej wersji prawo to mówi nam, że zmieniając całkowity strumień przechodzący przez daną powierzchnię (na przykład zbliżając lub oddalając magnes), wytworzymy nową jakość: pole elektryczne, które będzie „obiegać" ową powierzchnię. Co więcej, wartość krą-

[5] Mamy tutaj do czynienia z niezręcznością języka polskiego – termin „pole" może oznaczać zarówno pole elektryczne czy magnetyczne, jak i pole powierzchni. W języku angielskim stosuje się dwa terminy – *field* oraz *area*.

żenia pola elektrycznego (oczywiście powiązana z wartością samego pola) jest jednoznacznie skorelowana z szybkością zmiany strumienia – im szybciej zmieniamy przepływ pola magnetycznego, tym większe będzie krążenie, a w konsekwencji także większe pole elektryczne. Prawo Faradaya tłumaczy powstanie prądu w obwodzie w wyniku zjawiska indukcji – swobodne elektrony zawarte w przewodniku odczuwają istnienie wytworzonego pola elektrycznego, które „popycha" je w jednym kierunku, tworząc prąd elektryczny. Jednakże pole to powstaje nawet wtedy, gdy nie ma żadnego przewodnika. Pole elektryczne jest traktowane jako realność istniejąca niezależnie od tego, czy wywołuje ona obserwowalne dla nas skutki, czy też nie.

Matematyczny zapis prawa Faradaya jest następujący:

$$\frac{1}{c}\frac{d}{dt}\int_S \boldsymbol{B}\cdot d\boldsymbol{S} = -\oint_O \boldsymbol{E}\cdot d\boldsymbol{l}. \tag{4.3}$$

Wyrażenie po lewej stronie oznacza szybkość zmiany strumienia magnetycznego (pochodna po czasie jest zawsze chwilową szybkością zmiany różniczkowanej funkcji). Prawą stronę równania rozpoznajemy jako krążenie pola elektrycznego po krzywej O, która zamyka powierzchnię S (znak minus jest pewnego rodzaju konwencją). Zauważmy, że w matematycznej formule brak określenia, który proces jest przyczyną którego. Jest to typowe dla równań fizyki matematycznej, które wymagają dopowiedzenia w języku przyczyn i skutku. Sama relacja matematyczna między wielkościami nie mówi nam jeszcze, w którym kierunku biegnie proces przyczynowy, ani nawet czy mamy w ogóle do czynienia z procesem kauzalnym czy też z innego rodzaju korelacją (np. współwystępowaniem). Powrócimy jeszcze do tej kwestii w dalszej części rozdziału.

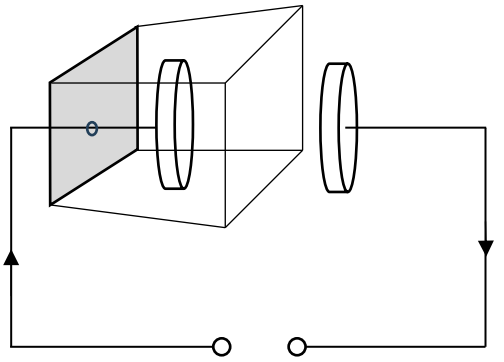
4.5. Prawo Ampère'a-Maxwella i równania Maxwella

Prawo Faradaya nie jest jedynym prawem określającym wzajemne przyczynowe korelacje między polami elektrycznymi i magnetycznymi. Istnieje zależność odwrotna, która pokazuje, jak wywołać powstanie krążącego pola magnetycznego. Z wcześniejszego podrozdziału pamiętamy, że pole magnetyczne wytworzymy, przepuszczając prąd elektryczny przez przewodnik. Obserwacja ta może być uogólniona w postaci tzw. prawa Ampère'a, które ma postać zbliżoną do prawa Faradaya. Głosi ono mianowicie, że przepływ ładunków elektrycznych przez dowolną powierzchnię wywołuje powstanie pola magnetycznego, którego krążenie wzdłuż krzywej zamykającej ową powierzchnię jest proporcjonalne do natężenia przepływającego prądu elektrycznego. Matematyczna formuła wyrażająca tę zależność jest następująca:

$$\frac{4\pi}{c}I = \oint_O \boldsymbol{B}\cdot d\boldsymbol{l}.$$

Okazuje się jednak, że prawo Ampère'a ma zasadniczą wadę. Mianowicie w pewnych sytuacjach prowadzi ono do niespójnych przewidywań, w zależności od arbitralnego wyboru powierzchni zamkniętej. Rozważmy sytuację, w której źródło prądu zostało podłączone do okładek kondensatora, jak na rys. 4.10. Przez pewien czas w obwodzie będzie płynął prąd – tzw. prąd ładowania – dopóki kondensator nie osiągnie stanu całkowitego naładowania. Weźmy teraz pod uwagę pewną powierzchnię przecinającą przewodnik (zaznaczoną szarym kolorem na rysunku). Ponieważ prąd ładowania przepływa przez tę powierzchnię, zgodnie z

prawem Ampère'a wzdłuż krzywej ją ograniczającej (prostokąt na rysunku) zacznie krążyć niezerowe pole magnetyczne. Jednakże możemy wybrać inną powierzchnię ograniczoną tą samą krzywą (na rysunku powierzchnia ta składa się ze ścianek wielościanu, z wyjątkiem ścianki zaznaczonej na szaro), która przechodzi między okładkami kondensatora, a zatem nie przecina przewodnika z prądem. W takiej sytuacji prawo Ampère'a przewiduje, że krążenie pola magnetycznego wzdłuż tej samej krzywej będzie równe zeru! Mamy zatem sprzeczność.



Rys. 4.10. Przykład ilustrujący sprzeczność w prawie Ampère'a

James Clerk Maxwell zaproponował rozwiązanie tego problemu przez wprowadzenie dodatkowego członu w równaniu Ampère'a, zwanego prądem przesunięcia. Jest to w istocie dokładny odpowiednik członu w prawie Faradaya, określający szybkość zmiany strumienia pola elektrycznego. Zatem pełna postać prawa Ampère'a-Maxwella będzie następująca:

$$\frac{4\pi}{c} I + \frac{1}{c}\frac{d}{dt}\int_S \boldsymbol{E}\cdot d\boldsymbol{S} = = \oint_O \boldsymbol{B}\cdot d\boldsymbol{l}. \tag{4.4}$$

Lewa strona równania określa dwa sposoby wytworzenia krążącego pola magnetycznego: albo przez przepuszczenie prądu przez daną powierzchnię, albo przez zmianę strumienia pola elektrycznego. Problem z prądem ładowania został rozwiązany dzięki temu, że pole elektryczne ładowanego kondensatora wzrasta, a zatem strumień przecinający tę drugą powierzchnie będzie rósł, co daje nam dokładnie takie samo krążenie wektora $\boldsymbol{B}$, jak to wyznaczone przez przepływ prądu przez pierwszą powierzchnię.

Słownie możemy wyrazić prawo Ampère'a-Maxwella w następujący sposób:

> Zmiana strumienia elektrycznego przez powierzchnię zamkniętą S lub przepływ prądu przez S wywołują powstanie pola magnetycznego, którego krążenie wokół powierzchni S jest zależne od szybkości zmiany strumienia i wielkości przepływającego prądu.[6]

[6] Na marginesie możemy zauważyć, że użycie terminologii przyczynowej („wywołuje") w sformułowaniu prawa Ampère'a-Maxwella prowadzi do pewnych trudności. Rozważając przykład z rys. 4.10, zwróćmy uwagę, że mamy tutaj dwie wzajemnie niezgodne hipotezy dotyczące tego, co jest przyczyną powstania krążącego pola magnetycznego wzdłuż zaznaczonej krzywej. W zależności od tego, którą z powierzchni uwzględnimy, odpowiedź będzie: albo przepływ prądu, albo zmiana strumienia pola elektrycznego. Aby uniknąć tego problemu, możemy porzucić sformułowanie przyczy-

Porównując prawa Faradaya (4.3) i Ampère'a-Maxwella (4.4.) zauważamy, że nie są one dokładnie symetryczne. Wynika to stąd, że w przyrodzie nie występują ładunki magnetyczne, a zatem także nie ma prądów magnetycznych. Jedyny sposób wywołania zjawisk elektrycznych przez magnetyczne jest taki, jak opisuje to prawo Faradaya: przez zmianę strumienia pola magnetycznego. Z drugiej strony, zjawiska elektryczne mogą wywołać krążenie pola magnetycznego dwojako: za pomocą zmian strumienia pola elektrycznego („polowo") albo przez przepływ ładunków („ładunkowo").

Możemy teraz dokonać podsumowania wszystkich podstawowych praw elektromagnetyzmu, które ujmują całokształt naszej wiedzy na temat tej dziedziny zjawisk. Do praw Faradaya i Ampère'a-Maxwella, które charakteryzują wzajemne relacje między elektrycznością i magnetyzmem, dodamy jeszcze dwa prawa opisujące elektryczność i magnetyzm oddzielnie. Pierwsze z nich jest uogólnieniem prawa Coulomba (czasem określa się je mianem prawa Gaussa). Głosi ono, że całkowity strumień pola elektrycznego przez powierzchnię zamkniętą (np. sferę) jest proporcjonalny do sumarycznego ładunku elektrycznego Q zawartego wewnątrz tej powierzchni. Matematycznie wyrazimy tę zależność następująco:

$$\int_S \boldsymbol{E} \cdot d\boldsymbol{S} = 4\pi Q. \tag{4.5}$$

Choć wyrażenie po lewej stronie równości jest identyczne z wyrażeniem na strumień w prawie Ampère'a-Maxwella, to jednak pamiętajmy, że w prawie Gaussa powierzchnia S musi być zamknięta, a zatem nie jest ona ograniczona żadną krzywą. Prawo Gaussa implikuje, że na całkowitą wartość strumienia przez powierzchnię zamkniętą nie mają żadnego wpływu ładunki znajdujące się na zewnątrz tej powierzchni.

Analogiczne prawo w wypadku pola magnetycznego ma następującą postać:

$$\int_S \boldsymbol{B} \cdot d\boldsymbol{S} = 0. \tag{4.6}$$

Wynika to wprost z założenia, że nie istnieją ładunki magnetyczne. Pole magnetyczne jest polem *bezźródłowym*. Dla dowolnej powierzchni zamkniętej, dokładnie tyle samo strumienia magnetycznego przez nią wpływa, co wypływa.

Dla pełnego obrazu warto przedstawić wszystkie cztery prawa elektromagnetyzmu (4.3) – (4.6) (znane również jako prawa lub równania Maxwella) w tzw. postaci różniczkowej. Różniczkowa wersja praw elektromagnetyzmu formułuje wszystkie zależności między polem elektrycznym, magnetycznym, ładunkiem i prądem elektrycznym „punktowo". Mówią one precyzyjnie, jakie matematyczne relacje łączą natężenie pola elektrycznego w danym punkcie z natężeniem pola magnetycznego w tym samym punkcie oraz z gęstością ładunku i prądu elektrycznego również w tym samym punkcie. Napiszmy może odpowiednie równania, które następnie opiszemy poglądowo. Dokładniejszą matematyczną analizę tych formuł podamy w paragrafie z gwiazdką.

$$\begin{aligned} \boldsymbol{\nabla} \cdot \boldsymbol{E} &= 4\pi\rho \\ \boldsymbol{\nabla} \cdot \boldsymbol{B} &= 0 \end{aligned} \tag{4.7}$$

nowe na korzyść korelacji: krążenie pola magnetycznego jest liczbowo skorelowane zarówno z przepływem prądu, jak i zmianą strumienia. Taka akauzalna interpretacja równań Maxwella będzie przyjęta, kiedy zapiszemy je w postaci różniczkowej.

$$\nabla \times \boldsymbol{E} = -\frac{1}{c}\frac{\partial \boldsymbol{B}}{\partial t}$$

$$\nabla \times \boldsymbol{B} = \frac{4\pi}{c}\boldsymbol{j} + \frac{1}{c}\frac{\partial \boldsymbol{E}}{\partial t}.$$

Symbole $\nabla \cdot \boldsymbol{E}$ oraz $\nabla \cdot \boldsymbol{B}$ oznaczają tzw. dywergencje pól $\boldsymbol{E}$ i $\boldsymbol{B}$. Dywergencja danego pola w punkcie jest to liczba, która określa „wypływ" danego pola z tego punktu (można to pojęcie interpretować przez analogię z przepływem cieczy – niezerowa dywergencja w danym miejscu oznacza, że w tym miejscu znajduje się źródło wypływu cieczy). Zatem pierwsze równanie Maxwella mówi tyle, że „punktowym" źródłem wypływu pola elektrycznego jest gęstość ładunku elektrycznego w tym punkcie. Analogicznie, drugie prawo głosi, że wypływ pola magnetycznego jest zawsze zerowy (gdyż nie istnieją ładunki magnetyczne). Pozostałe dwa prawa to oczywiście prawo Faradaya i Ampère'a-Maxwella. Po lewej stronie obu równań występuje operacja tzw. rotacji pola oznaczona, odpowiednio, przez $\nabla \times \boldsymbol{E}$ i $\nabla \times \boldsymbol{B}$. Operator $\nabla \times \boldsymbol{E}$ reprezentuje „skłonność" wektora $\boldsymbol{E}$ do obracania się w niewielkim obszarze otaczającym dany punkt. (Jak każdy operator różniczkowy, także rotacja mówi nam coś na temat zachowania danego obiektu nie tylko w samym punkcie, ale przede wszystkim w „infinitezymalnym" rejonie wokół punktu. Oczywiście w samym punkcie wektor $\boldsymbol{E}$ się nie obraca – ma jedną ustaloną wartość, kierunek i zwrot.)

Prawa strona trzeciego równania (Faradaya) reprezentuje szybkość, z jaką zmienia się wartość pola $\boldsymbol{B}$ w punkcie. Zatem całe równanie łączy skłonność do obracania się wektora pola elektrycznego w punkcie z tempem zmiany pola magnetycznego w tym samym punkcie. Podobnie czwarte równanie Ampère'a-Maxwella mówi nam, że punktowa rotacja pola magnetycznego jest równa szybkości zmiany pola elektrycznego plus gęstość prądu elektrycznego $\boldsymbol{j}$. Różniczkowe równania Maxwella powstają z wersji całkowych przez przejście z rozciągłych obszarów całkowania do punktu. Weźmy jako przykład prawo Faradaya w wersji całkowej. Wybierając coraz to mniejszą powierzchnię, dojdziemy do sytuacji, w której krążenie pola elektrycznego wokół infinitezymalnie małej krzywej zamkniętej (na jednostkę powierzchni) przejdzie w rotację punktową. Z kolei strumień pola magnetycznego przez powierzchnię przy przechodzeniu do punktu o zerowych rozmiarach zmienia się po prostu w iloczyn wartości samego pola razy ta powierzchnia. Bardziej ścisłe matematyczne wyjaśnienie relacji między wersją różniczkową a całkową znajdziecie w paragrafie z gwiazdką.[7]

Zestaw równań Maxwella ilustruje również pewnego rodzaju przejście na wyższy poziom abstrakcji w porównaniu z oryginalnymi prawami Faradaya i Ampère'a-Maxwella. Jak już wcześniej wspominaliśmy, pierwotnie prawa te opisywały pewne asymetryczne związki przyczynowe między zmianami odpowiednich pól z jednej strony, a wytworzeniem innych pól z drugiej. Obecnie jednak traktujemy prawa elektromagnetyzmu jako wzajemne matematyczne współzależności między wektorami $\boldsymbol{E}$ i $\boldsymbol{B}$, bez dodatkowych założeń o charakterze przyczynowo-skutkowym. Stwierdzamy po prostu, iż jeśli w danym obszarze występują pola elektryczne i magnetyczne, muszą one spełniać wszystkie cztery równania. To, czy jedne

[7] Ontologiczną zaletą różniczkowych wersji praw Maxwella w stosunku do praw w wersji całkowej jest to, że te pierwsze są jawnie „lokalne". Na przykład prawo Gaussa w wersji całkowej wydaje się sugerować, że strumień przez pewną powierzchnię zamkniętą zależy bezpośrednio od ładunku, który może być zlokalizowany w dowolnej odległości od tej powierzchni. Natomiast wersja różniczkowa uzależnia pewną punktową własność pola elektrycznego (jego dywergencję w punkcie) jedynie od gęstości ładunku w tym samym punkcie.

pola wywołują drugie, nie jest już przedmiotem naszej analizy. Oczywiście w konkretnych zastosowaniach możemy powrócić do myślenia w kategoriach przyczyny i skutku (na przykład podając intuicyjny opis rozchodzenia się fal elektromagnetycznych, o czym będziemy jeszcze pisać), ale założenia o charakterze przyczynowym nie są częścią praw Maxwella rozumianych w sposób abstrakcyjny.

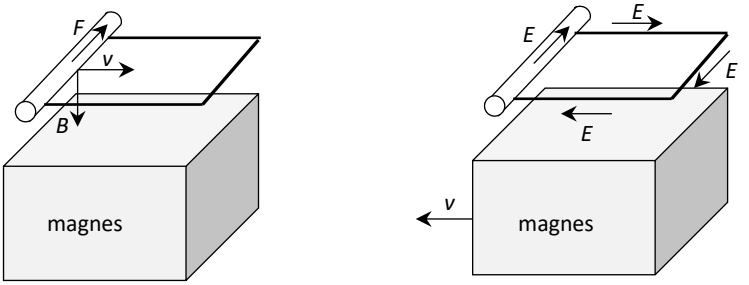
4.6. Unifikacja elektromagnetyzmu

Obecnie podejmiemy ważny temat unifikacji. Powszechnie uważa się, że przy pomocy swoich równań Maxwell zunifikował niezależne zjawiska elektryczne i zjawiska magnetyczne w jeden rodzaj zjawisk elektromagnetycznych. Ujmując sprawę bardziej ontologicznie: zamiast wprowadzać dwa typy pól fizycznych, Maxwell przyjął istnienie jednego bytu w postaci pola elektromagnetycznego. Jednakże powstaje pytanie, czy sama postać równań Maxwella uprawnia nas do tak daleko idącego wniosku ontologicznego. Niewątpliwie faktem jest, że prawa Maxwella unifikują naszą wiedzę na temat zjawisk elektrycznych i magnetycznych w sensie przeprowadzenia syntezy, tj. podciągnięcia wielu różnorodnych zjawisk pod niewielką liczbę fundamentalnych praw. Czy jednak taka synteza wystarczy, aby twierdzić, że występujące w tych prawach różne wielkości reprezentują jeden i ten sam byt? Byłaby to chyba zbyt daleko idąca teza. W nauce istnieje wiele przykładów syntetycznych praw łączących przeróżne wielkości, ale nikt nie wyprowadza stąd wniosku o utożsamieniu obiektów opisywanych przez te wielkości. Chociaż prawo gazu doskonałego stwierdza istnienie ścisłych korelacji między ciśnieniem, temperaturą i objętością, to nie prowadzi to do wniosku, że wielkości te reprezentują jeden i ten sam aspekt rzeczywistości: ciśnienio-objętościo-temperaturę. Jakie zatem dodatkowe argumenty można przedstawić na rzecz tezy, iż w wypadku zjawisk elektrycznych i magnetycznych powinniśmy dokonać ontologicznej unifikacji?

Zauważmy na wstępie, że zawsze możliwa jest „trywialna" unifikacja zjawisk za pomocą operacji logicznej sumy (alternatywy). Możemy np. uznać, że zjawisko elektromagnetyczne to zjawisko elektryczne lub magnetyczne. Lub też możemy nazwać polem elektromagnetycznym każde pole elektryczne i każde pole magnetyczne. Byłaby to jednak zupełnie nieciekawa forma unifikacji, niemająca żadnych głębszych podstaw teoretycznych. Równie dobrze można by w ten sposób dokonać unifikacji grawitacji i elektryczności, tworząc sztuczny twór pola „grawitacyjno-elektrycznego", które jest albo polem grawitacyjnym, albo elektrycznym. Źródeł unifikacji elektromagnetyzmu należy szukać w szczegółach teorii, a nie w arbitralnych ustaleniach terminologicznych. Pewną wskazówką mogą być dwa prawa łączące elektryczność i magnetyzm: prawo Faradaya i Ampère'a-Maxwella, interpretowane przyczynowo. Relacja „wywoływania", która zachodzi między zmianą jednego pola a powstaniem drugiego, może nasuwać sugestię, że te dwa rodzaje pól łączy ścisła zależność, być może nawet o charakterze ontologicznym. Jednakże nadal sprawa nie jest zupełnie jasna. Głośny wybuch może wywołać panikę u znajdujących się w pobliżu ludzi, ale jednak nie sądzi się, że wybuchy i panika są „dwiema stronami tego samego medalu".

Uzasadnienia dla elektromagnetycznej unifikacji należy szukać gdzie indziej – w szczególnym sposobie opisu zjawisk elektrycznych i magnetycznych w różnych układach odniesienia. Rozważmy następujący przykład: przewodnik przesuwamy nad stałym magnesem (rys. 4.11). Opiszmy, co dzieje się w przewodniku przy założeniu, że nasz układ odniesienia związany jest z magnesem (diagram po lewej stronie). W takiej sytuacji przewodnik wraz z zawartymi w nim swobodnymi elektronami porusza się w stałym polu magnetycznym,

a zatem elektrony zaczną odczuwać działającą na nie magnetyczną siłę Lorentza, która spowoduje ich ruch i w konsekwencji dodatnie naładowanie jednego końca i ujemne naładowanie końca przeciwnego (pojawi się zatem różnica potencjału między końcami przewodnika). Spróbujmy natomiast spojrzeć na sytuację z perspektywy przewodnika (diagram po prawej). W takim ujęciu elektrony w przewodniku są stacjonarne, natomiast przesuwa się pole magnetyczne. Jeśli narysujemy powierzchnię, której bokiem jest nasz przewodnik, to możemy stwierdzić, że strumień pola magnetycznego przechodzącego przez tę powierzchnię ulega zmianie, a zatem zgodnie z prawem Faradaya indukuje się tutaj krążące pole elektryczne. Właśnie to indukowane pole elektryczne powoduje uporządkowany ruch elektronów w przewodniku.

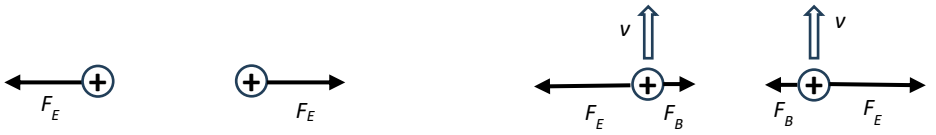


Rys. 4.11. Dwa opisy indukowania prądu w przypadku względnego ruchu przewodnika i magnesu

Porównajmy oba obrazy pod kątem istnienia odpowiednich pól. W pierwszym układzie istnieje tylko pole magnetyczne i to ono wywołuje obserwowany skutek. W drugim układzie natomiast istnieje i pole magnetyczne, i elektryczne. Zatem zmiana układu odniesienia („punktu widzenia") może spowodować „zniknięcie" jednego z dwóch pól: elektrycznego. Korzystając z zasady, którą sformułowaliśmy przy okazji analizy mechaniki klasycznej, powiemy, że prawdziwa realność przysługuje tylko obiektom czy zjawiskom, które są niezmiennicze względem transformacji z jednego do drugiego układu odniesienia. W rezultacie powinniśmy przyjąć, że samo pole elektryczne nie jest obiektywnym składnikiem rzeczywistości, skoro jego rolę może przejąć pole magnetyczne.

W istocie rzeczy zachodzi tutaj pełna symetria – pole magnetyczne również może być „zamienione" na pole elektryczne przez wybór odpowiedniego układu odniesienia. Rozważmy np. dwa jednoimienne ładunki elektryczne pozostające względem siebie w spoczynku (rys. 4.12). W układzie odniesienia związanym z tymi ładunkami występuje tylko pole elektryczne, które wzajemnie odpycha owe ładunki. Jednakże przechodząc do układu poruszającego się względem ładunków, musimy uwzględnić pole magnetyczne, które powstaje zgodnie z prawem Ampère'a-Maxwella z powodu ruchu każdego ładunku, i które powoduje powstanie odpowiedniej siły działającej na drugi z ładunków. Zatem tym razem istnienie pola magnetycznego zależy od przyjętego układu odniesienia. To, co pozostaje niezmiennicze względem wyboru układu odniesienia, to po pierwsze, obecność efektywnej siły działającej

na obiekty (niezależnie od tego, czy jej źródłem jest elektryczność czy magnetyzm)[8], a po drugie, obecność co najmniej jednego z dwóch pól: elektrycznego bądź magnetycznego. Nie jesteśmy w stanie całkowicie zlikwidować i pola elektrycznego, i magnetycznego, przez wybór odpowiedniego układu odniesienia.



Rys. 4.12. Oddziaływanie dwóch ładunków w różnych układach odniesienia

Powyższe fakty mocno sugerują, że pola elektryczne i magnetyczne traktowane osobno nie posiadają pełnej realności fizycznej. Natomiast musi istnieć pewien byt fizyczny odpowiedzialny za wszystkie obserwowalne zjawiska elektromagnetyczne. Tym bytem jest pewnego rodzaju kombinacja elektryczności i magnetyzmu, którą nazwiemy polem elektromagnetycznym. Pole elektromagnetyczne przejawia się w różnych układach odniesienia w odmienny sposób: czasem jak samo pole elektryczne, czasem jak samo pole magnetyczne, a czasem w postaci obu pól. Pojawia się jednak pytanie, w jaki sposób matematycznie reprezentować to nowe pole elektromagnetyczne. Czy w tym celu nie powinniśmy wprowadzić nowego wektora, oznaczając go np. jako $\boldsymbol{EB}$? Wektor taki powinien być wyrażalny przy pomocy oddzielnych wektorów $\boldsymbol{E}$ i $\boldsymbol{B}$, a także mieć cechę niezmienniczości przy przechodzeniu z jednego układu odniesienia do drugiego.

Okazuje się, że sprawa nie jest taka prosta. Poprawny opis pola elektromagnetycznego wymaga wprowadzenia nowego pojęcia matematycznego – tensora. Tensory odgrywają niezmiernie ważną rolę w wielu działach fizyki. Są one bardzo użyteczne przy formułowaniu relatywistycznych praw mechaniki, wynikających ze szczególnej teorii względności, a ogólna teoria względności nie mogłaby się bez nich w żaden sposób obejść. Czym dokładnie jest tensor? Zacznijmy od rozbudzenia pewnych intuicji. Pojęcie tensora może być rozumiane jako pewnego rodzaju uogólnienie pojęcia wektora. Wektor w danej przestrzeni charakteryzowany jest przez swoje składowe – trzy w wypadku przestrzeni trójwymiarowej, cztery dla czasoprzestrzeni, która ma czwarty wymiar czasowy. Można zatem ująć abstrakcyjnie wektory jako n-tki liczb, gdzie n jest wymiarem danej przestrzeni. (Oczywiście musimy pamiętać, że liczby te zależą od wyboru układu współrzędnych – jeden i ten sam wektor będzie miał inne składowe w różnych układach współrzędnych – np. takich, których osie są obrócone względem siebie). Skrótowo, dany wektor możemy symbolizować w postaci litery opatrzonej jednym indeksem, np. V_p, gdzie p jest liczbą od 1 do n.

Tensory są obiektami, które również posiadają składowe w różnych układach współrzędnych, ale mają ich więcej niż wektory. Dokładniej, liczba ich składowych jest odpowiednią

[8] Pomijamy tutaj pewien ważny problem: w różnych układach odniesienia wartość siły działającej na ładunki obliczona na podstawie praw elektromagnetyzmu wydaje się różna. Jego rozwiązanie możliwe jest dopiero na gruncie szczególnej teorii względności, w której należy zmodyfikować zarówno wzór na siłę Lorentza, jak i prawo Newtona, łączące działającą siłę z przyspieszeniem ciała (zob. paragrafy 5.7 i 5.8).

potęgą wymiaru przestrzeni n. Najprostszy przykład to tensor „dwuwymiarowy", który posiada dwa indeksy: T_{pq}. Ponieważ każdy indeks p i q może przyjąć dowolną wartość od 1 do n, ogólnie taki tensor będzie miał $n \times n$ składowych. Wygodnie jest przedstawiać tensory dwuwymiarowe w postaci macierzy, czyli tablicy liczb z n kolumnami i n wierszami. Jednak trzeba pamiętać, że – podobnie jak wektory – tensory w różnych układach współrzędnych wyglądają różnie. Nie powinniśmy zatem utożsamiać tensora z konkretną macierzą liczb, ale raczej z pewnym abstrakcyjnym obiektem, którego reprezentacje w różnych układach są odpowiednimi macierzami. Przechodząc z jednego układu współrzędnych do drugiego, musimy dokonać odpowiedniej transformacji składowych tensora.[9]

Tensor opisujący pole elektromagnetyczne jest tensorem o „rozmiarze" 4×4.[10] Wynika to stąd, że poprawne sformułowanie teorii elektromagnetyzmu wymaga zastosowania teorii względności, która oparta jest na pojęciu czterowymiarowej czasoprzestrzeni. Tensor ten symbolizuje się jako $F_{\mu\nu}$. (W fizyce obowiązuje konwencja, zgodnie z którą indeksy w postaci greckich liter przebiegają cztery wartości 0, 1, 2, 3, gdzie indeks 0 oznacza współrzędną czasową, a 1, 2 i 3 – trzy współrzędne przestrzenne. Z kolei litery alfabetu łacińskiego oznaczają zawsze współrzędne przestrzenne – 1, 2 i 3.) Składowymi tensora pola elektromagnetycznego są składowe pól elektrycznych i magnetycznych E_x, E_y, E_z, B_x, B_y, B_z z odpowiednimi znakami + lub –. Okazuje się, że jeśli dokonamy odpowiedniej transformacji tensora $F_{\mu\nu}$, przechodząc z jednego układu odniesienia do drugiego, w danym miejscu macierzy, gdzie uprzednio była tylko jedna składowa pola elektrycznego lub pola magnetycznego, może pojawić się kombinacja składowych pól elektrycznych i magnetycznych. Odpowiada to wspomnianemu wcześniej faktowi, że zmieniając układ odniesienia możemy „włączyć" lub „wyłączyć" odpowiednie pola – elektryczne bądź magnetyczne. Jednakże tensor $F_{\mu\nu}$ pozostaje tym samym obiektem, reprezentującym jeden i ten sam byt fizyczny, chociaż w różnych układach odniesienia jego składowe są różne. Warto podkreślić, że tensor $F_{\mu\nu}$ transformuje się zgodnie z zasadami szczególnej teorii względności, o których jeszcze będziemy mówić w rozdziale 5.

Aby podsumować temat unifikacji w teorii elektromagnetyzmu dodajmy, że wszystkie cztery prawa elektromagnetyzmu można przedstawić w syntetycznej formie, wykorzystującej tylko tensor $F_{\mu\nu}$ (oraz 4-składnikową gęstość prądu elektrycznego, która łącznie reprezentuje ładunek elektryczny i prąd). Nie wchodząc w matematyczne szczegóły zapisu, podam formę równań Maxwella w wersji tensorowej, aby Czytelnik mógł docenić ich zwięzłość i formalną elegancję:

---

[9] Dodajmy, że istnieją dwa ogólne sposoby transformowania tensorów przy zadanej zmianie współrzędnych, które prowadzą do wyróżnienia dwóch rodzajów tensorów (a także wektorów, będących szczególnym przypadkiem tensora). Rodzaje te zwane są tensorami kowariantnymi i kontrawariantnymi – odróżnia się je za pomocą indeksów górnych (tensory kontrawariantne) i dolnych (tensory kowariantne). Zatem T_{ab} symbolizuje tensor kowariantny, a T^{ab} – kontrawariantny. Wspomnimy więcej o tym rozróżnieniu w rozdziałach poświęconych szczególnej i ogólnej teorii względności.

[10] Ściśle rzecz biorąc, możliwe jest kompletne przedstawienie pola elektromagnetycznego przy pomocy wektora – dodajmy, wektora w czterowymiarowej czasoprzestrzeni, a zatem posiadającego cztery składowe. Jest to tzw. czterowektor potencjału elektromagnetycznego, który jest uogólnieniem i syntezą potencjału elektrostatycznego V i 3-wymiarowego wektora potencjału magnetycznego $\boldsymbol{A}$. Jednakże podobnie jak potencjały elektrostatyczny i magnetyczny, potencjał elektromagnetyczny jest dany tylko z dokładnością do pewnej transformacji (transformacji cechowania), a zatem zgodnie z wcześniej przedstawionymi argumentami, nie przysługuje mu pełna realność fizyczna.

$$\partial_\sigma F_{\nu\tau} + \partial_\nu F_{\tau\sigma} + \partial_\tau F_{\sigma\nu} = 0$$

$$\partial_\nu F^{\mu\nu} = J^\mu.$$

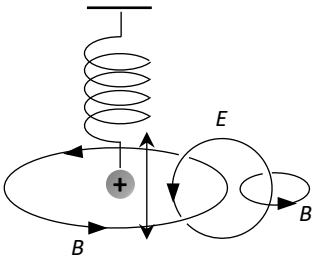
Symbol ∂_σ jest często używanym w fizyce matematycznej skrótowym zapisem pochodnej cząstkowej względem współrzędnej x^σ: $\frac{\partial}{\partial x^\sigma}$. Pierwsze z powyższych równań jest równoważne koniunkcji drugiego i trzeciego prawa Maxwella – są to tzw. równania jednorodne, w których nie występuje ani ładunek, ani prąd elektryczny. Drugie z kolei reprezentuje pozostałe dwa równania Maxwella (tzw. niejednorodne). Symbol J^μ oznacza czterowektor gęstości prądu, którego pierwsza składowa to gęstość ładunku elektrycznego, a trzy pozostałe reprezentują gęstość prądu w przestrzennych kierunkach x, y i z. Dopiero powyższe dwa równania stanowią kwintesencję unifikacji, zapoczątkowaną przez Maxwella. (Więcej informacji na ten temat znajdziecie w rozdziale 5 w paragrafach z gwiazdką.)

4.7. Fale elektromagnetyczne

Równania Maxwella – zarówno w wersji „tradycyjnej", jak i tensorowej – w piękny sposób dokonują syntezy naszej wiedzy na temat zjawisk elektromagnetycznych. Czy jednak na tym wyczerpuje się ich naukowa wartość? Czy prawa elektromagnetyzmu nie zawierają pewnej wartości dodanej, która poszerza naszą wiedzę w nieprzewidziany wcześniej sposób? Okazuje się, że tak właśnie jest. Sugerują one istnienie nowego i wcześniej zupełnie nieznanego zjawiska: rozchodzenia się fal elektromagnetycznych. Jest to jeden z bardziej spektakularnych przykładów z historii nauki, jak rozwój teorii może doprowadzić do sformułowania nowych przewidywań empirycznych, które następnie zostały w pełni potwierdzone doświadczalnie.

Rozumowanie Maxwella prowadzące do postawienia hipotezy o istnieniu fal elektromagnetycznych miało charakter ściśle matematyczny – przedstawimy je w ogólnych zarysach za chwilę, a bardziej szczegółowo w paragrafie z gwiazdką. Zanim jednak do tego dojdziemy, zauważmy, że idea rozchodzenia się „naprzemiennych" pól elektrycznych i magnetycznych może być łatwo wyprowadzona z przyczynowo zinterpretowanych praw Faradaya i Ampère'a-Maxwella. Wyobraźmy sobie pojedynczy ładunek elektryczny zawieszony na sprężynce, który „skacze" w górę i w dół ze zmienną prędkością (rys. 4.13). Zgodnie z prawem Ampère'a-Maxwella wokół tego ładunku pojawi się krążące pole magnetyczne, gdyż będzie on „przecinał" płaszczyznę prostopadłą do sprężynki, tworząc w ten sposób prąd elektryczny przechodzący przez tę płaszczyznę. Zauważmy jednak, że krążące pole magnetyczne będzie zmieniało swoje natężenie w czasie, gdyż jest ono uzależnione od wielkości prądu elektrycznego, czyli prędkości ładunku, a ta zmienia swoją wartość, a nawet kierunek. Zatem krążące pole magnetyczne będzie naprzemiennie rosło i malało w zależności od zmian prędkości ładunku. Wybierając dowolną powierzchnię prostopadłą do wektora pola magnetycznego w danym punkcie, zauważymy, że strumień pola przez tę powierzchnię zmienia się w czasie, co powoduje powstanie krążącego pola elektrycznego, zgodnie z prawem Faradaya. Ale znów, wartość tego pola będzie ulegała zmianie w zależności od szybkości zmiany strumienia pola magnetycznego, więc kolejne zastosowanie prawa Ampère'a-Maxwella implikuje powstanie „wtórnego" pola magnetycznego odpowiednio zorientowanego w przestrzeni. Proces ten jest kontynuowany w przestrzeni, tworząc coś w rodzaju łańcucha na choinkę, którego

„oczka” będą kolejnymi krążącymi polami elektrycznymi i magnetycznymi. Mamy więc jakościowy opis tego, co nazywa się falą elektromagnetyczną. Warto jeszcze dodać, że proces ten trwa nawet wtedy, gdy ładunek wytwarzający pierwotne pole magnetyczne zostanie usunięty. Daje to nam dodatkowy argument za realnością pól, które mają zdolność wzajemnego „przepychania się” przez pustą przestrzeń.



Rys. 4.13. Wyjaśnienie powstawania fal elektromagnetycznych

Spróbujmy teraz opisać to zjawisko nieco ściślej. Zacznijmy od matematycznego ujęcia zjawisk falowych. Falą nazwiemy dowolne zaburzenie pewnego ośrodka, które zmienia się w odpowiedni sposób w czasie. Takim zaburzeniem może być np. wychylenie drgającej struny w danym punkcie albo podniesienie czy obniżenie powierzchni wody w stosunku do położenia równowagi. Niech funkcja F reprezentuje owo zaburzenie. Załóżmy, że F jest następującą funkcją położenia x i czasu t: $F(x - vt)$. Wynika stąd, że np. w chwili zerowej $t = 0$, zaburzenie w dowolnym punkcie x będzie równe $F(x)$. Ponieważ zachodzi oczywista równość $F(x) = F[(x + vt) - vt]$, po czasie t zaburzenie w punkcie $x + vt$ będzie dokładnie takie samo, jakie było w punkcie x w czasie zero. Zatem funkcja $F(x - vt)$ opisuje rozchodzenie się pewnego zaburzenia z prędkością v w kierunku rosnących wartości x.

Zapiszmy teraz pewne matematyczne równanie, łączące drugie pochodne funkcji F względem zmiennej przestrzennej i zmiennej czasowej. Ci z Was, którzy znają podstawowe zasady różniczkowania funkcji, łatwo sprawdzą, że dowolna funkcja periodyczna $F(x - vt)$ będzie spełniać owo równanie:

$$\frac{\partial^2 F}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2 F}{\partial t^2}.$$

Jest to tak zwane równanie falowe w przypadku jednowymiarowym. Dla trzech wymiarów przestrzennych równanie falowe przyjmuje następującą postać:

$$\frac{\partial^2 F}{\partial x^2} + \frac{\partial^2 F}{\partial y^2} + \frac{\partial^2 F}{\partial z^2} = \frac{1}{v^2}\frac{\partial^2 F}{\partial t^2}. \tag{4.8}$$

Istnieje wiele sposobów na wyprowadzenie tego równania dla różnych szczególnych przypadków fal: dźwiękowych w powietrzu, fal naprężeń w metalach, a nawet fal pól fizycznych opisywanych odpowiednim lagrangianem. Aby się „oswoić” z powyższym różniczkowym równaniem drugiego rzędu, możemy o nim myśleć jako o pewnym wariancie drugiego prawa Newtona. Zauważmy, że prawa strona równania zawiera wyrażenie bardzo przypominające przyspieszenie (jeśli uznamy funkcję F za określającą położenie pewnego obiektu, np. pojedynczego ciężarka zaczepionego na długiej linie z podobnymi ciężarkami). Wtedy lewa

strona równania będzie mogła być zinterpretowana jako reprezentująca całkowitą siłę (na jednostkę masy) działającą na dany obiekt i pochodzącą od sąsiadujących obiektów. Nas jednak interesuje związek tego równania z równaniami Maxwella. Jeśli wprowadzimy założenie o niewystępowaniu ładunków elektrycznych i prądu elektrycznego ($\rho = 0$, $\boldsymbol{j} = 0$), równania Maxwella uproszczą się do postaci „symetrycznej" względem pól elektrycznych i magnetycznych:

$$\boldsymbol{\nabla} \cdot \boldsymbol{E} = 0$$

$$\boldsymbol{\nabla} \cdot \boldsymbol{B} = 0$$

$$\boldsymbol{\nabla} \times \boldsymbol{E} = -\frac{1}{c}\frac{\partial \boldsymbol{B}}{\partial t}$$

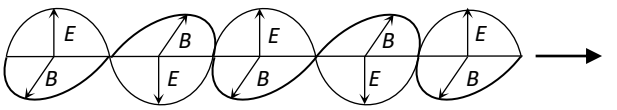
$$\boldsymbol{\nabla} \times \boldsymbol{B} = \frac{1}{c}\frac{\partial \boldsymbol{E}}{\partial t}.$$

Dokonując odpowiednich matematycznych operacji na powyższych równaniach i korzystając z praw łączących operacje rotacji i dywergencji, możemy pokazać, że zarówno wektor $\boldsymbol{E}$, jak i $\boldsymbol{B}$ spełniają równanie falowe (szczegóły w paragrafie z gwiazdką), przy czym w miejsce prędkości v pojawia się stała c:

$$\frac{\partial^2 \boldsymbol{E}}{\partial x^2} + \frac{\partial^2 \boldsymbol{E}}{\partial y^2} + \frac{\partial^2 \boldsymbol{E}}{\partial z^2} = \frac{1}{c^2}\frac{\partial^2 \boldsymbol{E}}{\partial t^2}$$

$$\frac{\partial^2 \boldsymbol{B}}{\partial x^2} + \frac{\partial^2 \boldsymbol{B}}{\partial y^2} + \frac{\partial^2 \boldsymbol{B}}{\partial z^2} = \frac{1}{c^2}\frac{\partial^2 \boldsymbol{B}}{\partial t^2}.$$

Równania te zdają się opisywać fale rozchodzące się z prędkością c. Do analizy sensu fizycznego stałej c powrócimy za chwilę, ale na razie wspomnijmy, że równania Maxwella dostarczają nam dodatkowych informacji na temat własności fizycznych owych fal (o ile występują one w rzeczywistości). Załóżmy, że nasza fala rozchodzi się w kierunku osi z. Jeśli rozpiszemy wektory natężenia pola elektrycznego i magnetycznego na trzy składowe wzdłuż osi x, y i z, możemy udowodnić, korzystając z założenia o zerowaniu się dywergencji, że składowe E_z oraz B_z będą wynosiły 0. Oznacza to, że zarówno wektor pola elektrycznego, jak i magnetycznego, są prostopadłe do kierunku rozchodzenia się fali. Fale tego rodzaju nazywamy poprzecznymi (podobnie jak fale na powierzchni wody albo na drgającej strunie).



Rys. 4.14. Fala elektromagnetyczna rozchodząca się w przestrzeni

Na koniec możemy zauważyć, że wektor pola magnetycznego $\boldsymbol{B}$ rozchodzącej się fali będzie zawsze prostopadły do wektora $\boldsymbol{E}$. Wynika to wprost z czwartego równania Maxwella – rotacja $\boldsymbol{\nabla} \times \boldsymbol{B}$ jest iloczynem wektorowym, a zatem jest prostopadła do $\boldsymbol{B}$, a z równania wynika, że musi ona być równa kierunkowi zmiany pola $\boldsymbol{E}$, a więc także kierunkowi samego

E. Zatem rozchodzenie się fal elektromagnetycznych można przedstawić jako dwie prostopadłe do siebie sinusoidy, których kierunek wychylenia jest prostopadły do kierunku propagacji, jak na rys. 4.14.

4.8. Światło jako fala elektromagnetyczna

Wróćmy teraz do kwestii interpretacji stałej *c*, która pojawia się w równaniach Maxwella, a także w równaniu falowym. W podręcznikach standardowo objaśnia się tę stałą jako prędkość światła w próżni. Jest to jednak ujęcie ahistoryczne, gdyż związek praw elektromagnetyzmu z prędkością światła został wykryty stosunkowo niedawno. Pierwotnie stałą *c* należało zinterpretować po prostu jako pewną stałą fizyczną, której wartość można ustalić doświadczalnie na podstawie odpowiednich pomiarów oddziaływań elektrostatycznych i magnetycznych. Fakt, że w równaniach elektromagnetyzmu nie występuje bezpośrednio prędkość światła jako taka, najlepiej widać, kiedy wyrazi się te równania w innych jednostkach – zamiast w tzw. jednostkach Gaussa (zwanych również CGS – centymetr, gram, sekunda), jak w powyższym ujęciu, w jednostkach SI (albo też MKSA: metr, kilogram, sekunda, amper). W równaniach Maxwella wyrażonych w jednostkach metrycznych pojawiają się dwie stałe: przenikalność dielektryczna próżni ε_0 oraz przenikalność magnetyczna próżni μ_0. Przenikalność dielektryczną próżni można wyznaczyć, korzystając z prawa Coulomba, gdyż wchodzi ona do współczynnika proporcjonalności $k = \frac{1}{4\pi\varepsilon_0}$. Jednakże z powodów praktycznych jej wartość określa się precyzyjnie z równania opisującego pojemność elektryczną kondensatora płaskiego o zadanej powierzchni i odległości między okładkami. Z kolei stała μ_0 ma przypisaną konwencjonalną wartość w jednostkach MKSA, z uwagi na to, że w tym układzie definiujemy jednostkę natężenia prądu (amper) za pomocą sił wywieranych przez oddziaływanie magnetyczne między przewodnikami z prądem. Niezależnie od tych szczegółów, które mogą być zbyt zawiłe dla mało praktycznych filozofów, należy podkreślić, że stałe występujące w równaniach elektromagnetyzmu w żaden sposób nie są definiowane przy pomocy prędkości światła.

Przechodząc z układu MKSA do CGS, dostajemy następującą zależność między używanymi w nich stałymi:

$$c = \frac{1}{\sqrt{\varepsilon_0 \mu_0}}.$$

Zatem stała *c* charakteryzuje własności elektryczne i magnetyczne próżni. Jednakże jeśli policzymy jej wartość w jednostkach m/s, to okaże się, że jest ona zdumiewająco bliska znanej wartości prędkości światła w próżni, wyznaczonej w różnych doświadczeniach (np. przez Ole Rømera już w XVII wieku, a później z dużo większą dokładnością w doświadczeniach Armanda Fizeau i Jeana Foucaulta). Do tego dochodzi fakt, że stała *c* pojawia się w wyprowadzonym przez Maxwella równaniu falowym w miejscu prędkości. Te dwa fakty nie mogą być zwykłą koincydencją – prawdopodobieństwo, że jest to przypadkowa zbieżność jest niezwykle niskie. Wyobraźmy sobie np., że badając pisma jakiegoś starożytnego autora natrafilibyśmy na tekst do złudzenia przypominający *Hamleta* Szekspira. Czy wzruszylibyśmy ramionami: „No cóż, to po prostu zwykła koincydencja, którą nie warto sobie zawracać głowy”? Jest oczywiste, że natychmiast poszukiwalibyśmy wyjaśnienia – najprawdopodob-

niej badane pisma są współczesną podróbką, a może też Szekspir skopiował swój dramat z utworu wcześniejszego twórcy. Wykluczając te hipotezy możemy wreszcie, zgodnie ze wskazówkami wielkiego detektywa Sherlocka Holmesa[11], przyjąć, że ktoś z przyszłości zbudował machinę czasu i z niewiadomego powodu przekazał owemu autorowi kopię *Hamleta*. Utwór ten zrobił na nim takie wrażenie, że odtworzył go w swojej twórczości. Jakiekolwiek rozwiązanie zagadki zostałoby ostatecznie przyjęte, nie ulega wątpliwości, że wykryta koincydencja nie byłaby zignorowana.

Podobnie rzecz się ma z koincydencją pomiędzy wartością pewnej stałej elektromagnetycznej a prędkością światła. Najbardziej naturalnym wyjaśnieniem tego faktu jest założenie, że fale elektromagnetyczne rozchodzą się w próżni z prędkością c, a po drugie, że światło jest falą elektromagnetyczną. Na tym jednak sprawa się nie kończy. Zaproponować wyjaśnienie danego faktu to jedno, a uzasadnić prawdziwość tego wyjaśnienia to drugie. Ptolemeusz pięknie wyjaśnił ruch retrogradacyjny planet swoimi epicyklami, ale wyjaśnienie to okazało się fałszywe. Skąd wiemy, że podobna sytuacja nie może się powtórzyć w rozpatrywanym wypadku? Zasady metodologii naukowej zobowiązują, aby nowe hipotezy poddać wszechstronnym testom empirycznym. Tak też się stało w wypadku hipotezy fal elektromagnetycznych. Jej pierwszy eksperymentalny test został przeprowadzony przez niemieckiego fizyka Heinricha Hertza. W swoim eksperymencie wykorzystał on przewodnik pierwotny, w którym za pomocą cewki produkującej wysokie napięcie wywołał drgania elektromagnetyczne, a następnie zauważył, że w obwodzie wtórnym umieszczonym w pewnej odległości pojawiają się wyładowania iskrowe o częstotliwości skorelowanej z częstotliwością pierwotnych drgań. Co więcej, Hertz potwierdził falowy charakter rozchodzącego się oddziaływania elektromagnetycznego przez utworzenie fali stojącej za pomocą metalowej płyty, odbijającej falę pochodzącą od obwodu pierwotnego. Późniejsze zastosowanie fal elektromagnetycznych do bezprzewodowej komunikacji radiowej przez Guglielmo Marconiego przypieczętowało sukces teorii Maxwella.

Twierdzenie, że światło widzialne jest falą elektromagnetyczną o częstotliwości z pewnego zakresu, jest dzisiaj truizmem porównywalnym być może z tezą o kulistości Ziemi. Jednakże trudno wskazać na jeden szczególny eksperymentalny fakt potwierdzający tę tezę. Raczej w grę wchodzi tutaj cały szereg zjawisk, ujawniających falową naturę światła (interferencja, dyfrakcja), które są spójne z opisem fal elektromagnetycznych sformułowanym przez Maxwella. Z doświadczenia wiemy, że światło jest falą poprzeczną, co zgadza się z przewidywaniami teorii Maxwella. Innym faktem potwierdzającym elektromagnetyczny charakter światła jest zjawisko polaryzacji, wykorzystywane powszechnie np. w okularach przeciwsłonecznych. Wreszcie teoria Maxwella pozwala na wyprowadzenie wielu obserwowalnych prawidłowości optycznych: prawa załamania, odbicia czy też faktów dotyczących rozpraszania światła przez ośrodki takie jak powietrze. Można zatem przyjąć, że potwierdzenie hipotezy o elektromagnetycznym charakterze promieniowania świetlnego, jakkolwiek pośrednie, jest mocne ze względu na jego szeroki zakres i różnorodność.

Jaki wpływ na naszą ocenę teorii elektromagnetyzmu powinien mieć epizod z odkryciem fal elektromagnetycznych? Powszechnie uważa się, że teoretyczne przewidzenie istnienia i własności fal elektromagnetycznych stanowiło ogromny sukces i ukoronowanie teorii Maxwella. Dzięki temu teoria uzyskała dodatkowe duże wsparcie empiryczne. Oczywiście

[11] Jak tłumaczył Holmes swojemu przyjacielowi, Watsonowi: Gdy odrzucisz to, co niemożliwe, pozostała hipoteza, choćby najbardziej nieprawdopodobna, musi być prawdą.

teoria Maxwella objaśnia i trafnie opisuje wiele znanych zjawisk elektrycznych i magnetycznych: oddziaływania między naładowanymi ciałami i między przewodnikami z prądem, zjawisko indukcji elektromagnetycznej, przepływy prądów w obwodach i wiele innych. Jednakże waga odkrycia fal elektromagnetycznych na podstawie teoretycznych rozważań Maxwella jest nieporównanie większa. Czy jedynym powodem tego faktu jest spektakularny charakter owego odkrycia i jego szerokie praktyczne zastosowania? Bez powszechnego użycia fal radiowych nie wyobrażamy sobie przecież współczesnej cywilizacji technicznej. Można jednak wskazać inne powody tego, że teorie przewidujące zachodzenie nowych, wcześniej nieznanych zjawisk powinny uzyskać znacznie mocniejszy stopień zaufania niż teorie, które jedynie odtwarzają znane fakty. Jest dość oczywiste, że na etapie budowy danej teorii wszystkie znane uprzednio dane i fakty powinny być wzięte pod uwagę i włączone do teoretycznego opisu. Nie było np. żadnym zaskoczeniem, że Ptolemeuszowskie epicykle dobrze wyjaśniały zjawisko retrogradacyjnego ruchu planet, ponieważ właśnie w tym celu zostały wprowadzone. Podobnie rzecz się ma z bogactwem faktów dotyczących zjawisk elektromagnetycznych, zbieranych przez pokolenia badaczy – wszystkie one posłużyły do skonstruowania odpowiedniej teorii, a zatem zostały włączone do niej na etapie tworzenia. Nie chcemy przez to powiedzieć, że ujęcie różnorodnych faktów i zjawisk w jedno syntetyczne matematyczne sformułowanie nie stanowi znaczącego naukowego osiągnięcia. Jednakże teoria uzyskana w taki sposób nadal jest potwierdzona dlatego, że tak właśnie została skonstruowana. Aby wzmocnić przekonanie, że teoria jest trafna, potrzebujemy czegoś zupełnie nowego. Takim nowym, niespodziewanym potwierdzeniem dla teorii Maxwella były właśnie fale elektromagnetyczne.

W ramce poniżej znajdziecie nieco inny argument formalny za tym, że wpływ nieznanych wcześniej świadectw na ocenę prawdopodobieństwa hipotezy powinien być większy niż wpływ faktów znanych. Jest to tak zwane Bayesowskie podejście do potwierdzania hipotez, oparte na prostych prawach rachunku prawdopodobieństwa, do których dołącza się interpretację pojęcia prawdopodobieństwa w kategoriach subiektywnego *stopnia przekonania* (ang. *degree of belief*). Na przykład mówiąc, że prawdopodobieństwo danej hipotezy jest równe ½, chcemy powiedzieć, że mamy równie mocne przekonanie co do jej prawdziwości, jak i fałszywości.

W rachunku prawdopodobieństwa można udowodnić proste twierdzenie znane jako twierdzenie Bayesa. Dla dowolnych dwóch zdarzeń *H* i *E*, prawdopodobieństwo zdarzenia *H* przy założeniu zajścia *E* jest równe prawdopodobieństwu zajścia *E* przy założeniu zajścia *H* razy prawdopodobieństwo *H*, podzielone przez prawdopodobieństwo zajścia *E*:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}.$$

Wzór ten można bezpośrednio wyprowadzić z definicji prawdopodobieństwa warunkowego: $P(A|B) = \frac{P(A \cap B)}{P(B)}$. Przyjmujemy, że *H* jest ogólną hipotezą (np. w naszym wypadku jest to teoria Maxwella), a *E* – pewnym potwierdzonym doświadczalnie faktem (istnienie fal elektromagnetycznych). Prawdopodobieństwo warunkowe $P(H|E)$ wyraża liczbowo stopień przekonania, że hipoteza *H* jest prawdziwa, przy założeniu, że zweryfikowaliśmy zajście faktu *E* (jest to tzw. prawdopodobieństwo hipotezy *a posteriori*). Z kolei prawdopodobieństwo $P(E|H)$ może być uznane za równe jedności w sytuacji, w której hipoteza *H* logicznie implikuje wystąpienie zdarzenia *E* (jak to ma miejsce w wypadku teo-

rii Maxwella i istnienia fal elektromagnetycznych). Prawdopodobieństwo $P(H)$ to tzw. prawdopodobieństwo *a priori* hipotezy (tj. stopień wiarogodności hipotezy, zanim pojawiło się świadectwo E).

Najważniejszym elementem całej układanki jest prawdopodobieństwo $P(E)$, które określa naszą ocenę szansy tego, że świadectwo E jest prawdziwe, niezależnie od prawdziwości hipotezy H. Parametr ten ostro odróżnia fakty uprzednio znane of faktów nowych, niespodziewanych. Na przykład zjawisko indukcji elektromagnetycznej było znane, zanim Maxwell sformułował cały zestaw równań, a zatem można przyjąć, że stopień pewności, z jakim uznajemy wystąpienie owego zjawiska, wynosi jeden. W takiej sytuacji prawdopodobieństwo *a posteriori* teorii przy założeniu występowania zjawiska indukcji elektromagnetycznej jest równe prawdopodobieństwu *a priori*, czyli zachodzenie zjawiska indukcji nie wpływa na zmianę wiarogodności naszej hipotezy. Natomiast istnienie fal elektromagnetycznych nie było w żaden sposób znane przed sformułowaniem hipotezy Maxwella, czyli możemy przyjąć, że wiarogodność $P(E)$ była stosunkowo niska (lecz nie niższa niż wiarogodność samej hipotezy $P(H)$). Zatem stosunek $\frac{P(H)}{P(E)}$ może być odpowiednio dużą liczbą, w granicy zbliżającą się do jedności. To pokazuje, że zajście nieznanego wcześniej, lecz przewidywanego przez hipotezę zdarzenia E znacznie wzmacnia nasz stopień zaufania do hipotezy.

Wyprowadzenie z równań Maxwella matematycznego opisu fal elektromagnetycznych było niewątpliwym sukcesem, ale paradoksalnie zawierało element destrukcji, który doprowadził w końcu do wysadzenia w powietrze fundamentu fizyki klasycznej – koncepcji czasu i przestrzeni Galileusza i Newtona. Iskrą zapalną stała się tutaj interpretacja prędkości c w równaniu falowym, opisującym rozchodzenie się fal elektromagnetycznych. Jak wiadomo od Galileusza, pojęcie prędkości wymaga zawsze relatywizacji, czyli określenia, względem czego mierzymy daną prędkość. W wypadku równań opisujących fale mechaniczne nie jest to problemem, gdyż prędkość fali jest zawsze mierzona względem ośrodka, w którym rozchodzi się dana fala – powietrza, wody czy metalu. Kłopot pojawia się przy opisie fal elektromagnetycznych, dla których nie ma jednoznacznie określonego ośrodka ich rozchodzenia. „Zaburzeniem" dla fali elektromagnetycznej jest wartość pola elektrycznego i pola magnetycznego w danym punkcie przestrzeni, a nie odpowiedni fizyczny stan jakiejś substancji (np. ciśnienie czy gęstość niewielkiej porcji gazu). Fale elektromagnetyczne bez problemu rozchodzą się w próżni, w której przecież nie ma żadnego ośrodka fizycznego.

Czy aby na pewno? Wielu fizyków dziewiętnastowiecznych, włącznie z Maxwellem, uważało, że oddziaływania elektryczne i magnetyczne przenoszą się za pośrednictwem niewidzialnej substancji, zwanej eterem. Eter należy do kategorii fluidów, z których wcześniej poznaliśmy cieplik jako domniemany nośnik zjawisk termicznych. Pojęcie eteru pojawiło się już w starożytnych systemach astronomicznych (w szczególności u Arystotelesa), jako reprezentujące substancję wypełniającą świat nadksiężycowy (w średniowieczu zwaną również *quinta essentia*). Było ono wykorzystywane między innymi przez Newtona do opisu zjawisk optycznych. W wersji nowożytnej eter miał stanowić substancjalną podstawę dla pól elektrycznych i magnetycznych, które w takim ujęciu można było traktować jako „zaburzenie" pewnego fizycznego ośrodka. Jako ciekawostkę dodajmy, że Maxwell stworzył bardzo pomysłowy model eteru elektromagnetycznego oparty na idei przenoszenia oddziaływań za pomocą czegoś w rodzaju zazębiających się kół, czy też wirów. Była to zatem próba sprowa-

dzenia zjawisk elektromagnetycznych do mechanicznego oddziaływania przez kontakt. Zwróćmy uwagę, że przyjęcie modelu eteru Maxwella jako adekwatnie opisującego rzeczywistość rozwiązuje problem działania na odległość (nielokalności), o którym mówiliśmy w poprzednich paragrafach. Na przykład działanie prądu w przewodniku na odległą igłę magnetyczną może być wyjaśnione tym, że poruszające się elektrony w przewodniku pobudzają do ruchu sąsiadujące z nim koła (wiry) eteru, które znów odpowiednio przenoszą ten ruch, aż dotrze on do igły magnetycznej.

Koncepcja mechanicznego eteru przenoszącego oddziaływania elektromagnetyczne, a także umożliwiającego rozchodzenie się fal elektromagnetycznych, natrafiła na spore trudności teoretyczne. Jak pamiętamy, fale elektromagnetyczne są falami poprzecznymi. Z doświadczenia wiemy, że fale poprzeczne rozchodzą się w ośrodkach sprężystych, takich jak metale. Ośrodki niesprężyste, jak powietrze, przenoszą głównie fale podłużne (zagęszczanie i rozrzedzanie ośrodka w kierunku propagacji fal). Jednakże trudno sobie wyobrazić mechaniczny ośrodek, który byłby wystarczająco sprężysty, aby umożliwić propagację fal poprzecznych, a jednocześnie nie stawiał zauważalnego oporu poruszającym się w nim ciałom fizycznym. Sceptycyzm co do hipotezy eteru stopniowo narastał, ale ostateczny cios tej hipotezie zadały doświadczenia wyznaczające prędkość rozchodzenia się światła w różnych układach odniesienia. Omówimy te doświadczenia szczegółowo w następnym rozdziale, poświęconym wprowadzeniu do szczególnej teorii względności.

4.9.* Matematyczne podstawy teorii Maxwella

Jak zapewne zauważyliście, teoria elektromagnetyzmu posługuje się dość zaawansowanym aparatem matematycznym, którego podstawowym elementem są wektory oraz operacje na wektorach, w tym operacje odwołujące się do rachunku różniczkowego i całkowego. Spróbujmy przyjrzeć się dokładniej tym pojęciom, zaczynając od dwóch fundamentalnych operacji: iloczynu skalarnego i wektorowego. Będziemy przy tym rozważać wektory w trójwymiarowej przestrzeni, czyli posiadające trzy składowe wzdłuż dowolnie wybranych prostopadłych osi x, y i z. Niech $\boldsymbol{A}$ i $\boldsymbol{B}$ będą wektorami o składowych A_x, A_y, A_z oraz B_x, B_y, B_z (pamiętamy, że składowe wektora to zwykłe liczby). Iloczyn skalarny wektorów $\boldsymbol{A}$ i $\boldsymbol{B}$, który oznaczamy „kropką”, to następująca liczba:

$$\boldsymbol{A} \cdot \boldsymbol{B} = A_x B_x + A_y B_y + A_z B_z. \tag{4.9}$$

Można udowodnić, choć nie będziemy tego robić, że wzór ten jest równoważny znanej ze szkolnego kursu formule „iloczyn długości obu wektorów razy *cosinus* kąta między nimi”. Jest niezmiernie istotne, że iloczyn skalarny umożliwia nam zdefiniowanie pojęcia prostopadłości. Dwa wektory są prostopadłe, gdy ich iloczyn skalarny wynosi zero. Można to udowodnić ogólnie, używając nieco skomplikowanych funkcji trygonometrycznych, lub też zastosować następujący trik: ponieważ wybór układu współrzędnych jest sprawą konwencji, zawsze wolno wybrać trzy osie tak, aby tylko jedna współrzędna wektora $\boldsymbol{A}$ była niezerowa, np. A_x. Wtedy iloczyn skalarny $\boldsymbol{A}$ z dowolnym wektorem $\boldsymbol{B}$ redukuje się do iloczynu $A_x B_x$, a więc jego zerowanie oznacza, że $B_x = 0$. Zatem jedyne możliwe niezerowe składowe wektora $\boldsymbol{B}$ to B_y i B_z, czyli jest to wektor prostopadły do $\boldsymbol{A}$ (leży on w płaszczyźnie yz, prostopadłej do osi x).

Iloczyn wektorowy dwóch wektorów jest nieco bardziej skomplikowany. Przede wszystkim, jak sama nazwa wskazuje, jest to wektor, a nie liczba (skalar). Współrzędne tego wektora są następujące:

$$\begin{aligned}(\boldsymbol{A}\times\boldsymbol{B})_x &= A_yB_z - B_yA_z\,,\\ (\boldsymbol{A}\times\boldsymbol{B})_y &= A_zB_x - B_zA_x\,,\\ (\boldsymbol{A}\times\boldsymbol{B})_z &= A_xB_y - B_xA_y\,.\end{aligned} \tag{4.10}$$

Jak widać, każda współrzędna iloczynu wektorowego jest dana w postaci pewnej kombinacji pozostałych dwóch współrzędnych mnożonych wektorów. Nietrudno pokazać, że wektor $\boldsymbol{A}\times\boldsymbol{B}$ będzie prostopadły zarówno do $\boldsymbol{A}$, jak i $\boldsymbol{B}$. Sprawdźcie, proszę, sami, korzystając z formuł (4.9) i (4.10), że zachodzą następujące równości:

$$(\boldsymbol{A}\times\boldsymbol{B})\cdot\boldsymbol{A} = 0.$$

$$(\boldsymbol{A}\times\boldsymbol{B})\cdot\boldsymbol{B} = 0,$$

Wprowadźmy teraz ważne różniczkowe operacje dywergencji i rotacji. Kluczem do tych pojęć jest potraktowanie operacji różniczkowania względem trzech współrzędnych x, y, i z jako wektora. Jest to oczywiście pewne „nadużycie" notacji, ale okazuje się ono bardzo wygodne. Zatem zdefiniujemy następujący „wektor" różniczkowy $\boldsymbol{\nabla}$, podając jego składowe

$$\nabla_x = \frac{\partial}{\partial x},$$

$$\nabla_y = \frac{\partial}{\partial y},$$

$$\nabla_z = \frac{\partial}{\partial z}.$$

lub też „zbiorczo":

$$\boldsymbol{\nabla} = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right).$$

Składowe wektora $\boldsymbol{\nabla}$ (zwanego „nabla") nie są liczbami, a operatorami. Jednakże można użyć go do zwykłych operacji iloczynu skalarnego i wektorowego, opartych na mnożeniu liczb, jeśli zastąpimy mnożeniem „zadziałaniem" operatora na odpowiednim wektorze. Żeby to zrobić poprawnie, musimy jednak dysponować nie jednym wektorem, a całym polem wektorowym. Pole wektorowe to nic innego jak funkcja, która każdemu punktowi w przestrzeni przypisuje pewien wektor. Matematycznie najlepiej przedstawić taką funkcję jako składającą się z trzech funkcji o trzech argumentach rzeczywistych, z których każda przypisuje trzem współrzędnym x, y, z składową wektora w danym kierunku. Zatem ogólnie pole wektorowe $\boldsymbol{A}$ jest dane następująco:

$$\boldsymbol{A}(x,y,z) = \Big(A_x(x,y,z), A_y(x,y,z), A_z(x,y,z)\Big).$$

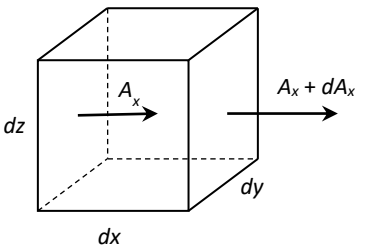
W praktyce często opuszcza się argumenty x, y i z i przedstawia dane pole wektorowe przy pomocy jednego symbolu $\boldsymbol{A}$. Pamiętać jednak należy, że nie jest to jeden wektor, a cała ich nieskończona grupa, z których każdy jest przypisany do jednego punktu.

Możemy teraz zdefiniować operacje dywergencji i rotacji na polu wektorowym $\boldsymbol{A}$, stosując formuły (4.9) i (4.10):

$$\boldsymbol{\nabla} \cdot \boldsymbol{A} = \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z} \tag{4.11}$$

$$\boldsymbol{\nabla} \times \boldsymbol{A} = \left(\frac{\partial A_y}{\partial z} - \frac{\partial A_z}{\partial y}; \frac{\partial A_z}{\partial x} - \frac{\partial A_x}{\partial z}; \frac{\partial A_x}{\partial y} - \frac{\partial A_y}{\partial x} \right) \tag{4.12}$$

Pamiętajmy, że wyrażenia A_x, A_y, A_z oznaczają funkcje trzech argumentów, więc można je różniczkować po każdym argumencie z osobna. W ten sposób otrzymamy dwie nowe funkcje, określone na punktach (trójkach liczb rzeczywistych): funkcję dywergencji, która każdemu punktowi przypisuje liczbę (jest to zatem pole skalarne), oraz funkcję rotacji, przypisującą punktom odpowiedni wektor.

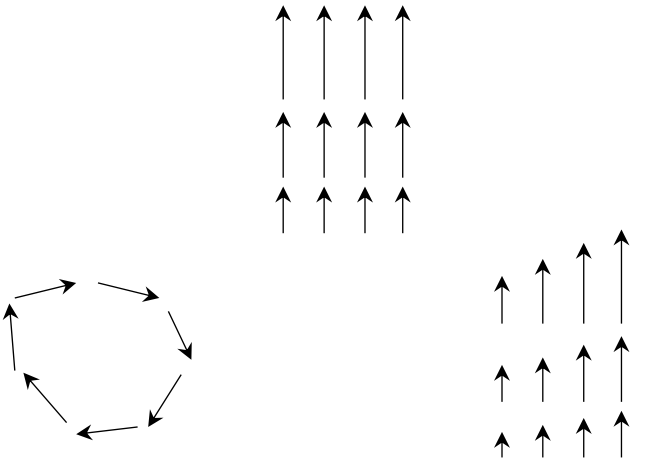


Rys. 4.15. Wytłumaczenie pojęcia dywergencji

Jaki jest sens fizyczny obu operacji? Zacznijmy od dywergencji. Rozważmy „infinitezymalny" sześcian o bokach *dx*, *dy* i *dz* (rys. 4.15) Wyobraźmy sobie, że pole wektorowe $\boldsymbol{A}$ obrazuje przepływ pewnej substancji (może to być pole prędkości przepływu wody w strumieniu). Niech dA_x oznacza zmianę składowej A_x pola przy przesunięciu wzdłuż osi *x* o wartość *dx*. Mnożąc tę wielkość przez pole powierzchni bocznej sześcianu prostopadłej do *x* otrzymamy $dA_x dydz$. Jest to, jak się łatwo domyślić, miara ilości wypływu danej substancji z sześcianu przez ścianki prostopadłe do *x* na jednostkę czasu (np. jeśli składowa A_x się w ogóle nie zmienia, czyli $dA_x = 0$, sumaryczny wypływ będzie zerowy: tyle samo wpływa, co wypływa). Jeśli chcemy policzyć tę ilość relatywnie do jednostki objętości, musimy podzielić wielkość $dA_x dydz$ przez element objętości $dxdydz$. Ale w ten sposób dostaniemy szybkość zmiany składowej A_x wzdłuż *x*: $\frac{dA_x}{dx}$, a raczej $\frac{\partial A_x}{\partial x}$, biorąc pod uwagę, że A_x zależy od trzech zmiennych. Analogiczne rozumowanie pokazuje, że pochodne cząstkowe $\frac{\partial A_y}{\partial y}$ i $\frac{\partial A_z}{\partial z}$ reprezentują wypływy danej substancji wzdłuż osi *y* i *z* na jednostkę czasu i objętości. Sumując te trzy liczby, otrzymujemy całkowity wypływ pola na jednostkę objętości z infinitezymalnego sześcianu.

Rotacja, jak nazwa wskazuje, przedstawia tendencję pola wektorowego do obracania się przy niewielkich przesunięciach. Nie będziemy tego uzasadniać ogólnie, aby nie przeładować naszego wywodu różnymi funkcjami trygonometrycznymi. Możemy jednak zauważyć,

że jeśli pole $\boldsymbol{A}$ ma stały kierunek, a jego wielkość nie zmienia się w kierunkach prostopadłych (por rys. 4.16), to jego rotacja będzie zerowa. Załóżmy, że $\boldsymbol{A}$ ma tylko jedną niezerową składową A_x, która oczywiście może zmieniać swoją wartość, ale tylko w kierunku osi x. W takiej sytuacji wszystkie pochodne cząstkowe występujące w definicji rotacji $\boldsymbol{\nabla} \times \boldsymbol{A}$ (4.12) będą równe zeru. Pochodne zerowych składowych A_y, A_z są z oczywistego powodu zerowe, natomiast jedyne pochodne składowej A_x wchodzące do równania są pochodnymi w kierunkach y i z, w których, jak założyliśmy, składowa ta się nie zmienia, czyli rezultat jest również zerowy. Aby rotacja pola wektorowego była niezerowa, pole to musi zmieniać swój kierunek przy niewielkich przesunięciach lub też zmieniać swoją wartość przy przesunięciach w kierunku prostopadłym do danego wektora, jak na rysunku.



Rys. 4.16. Przykład pola wektorowego z zerową rotacją (góra) i pól wektorowych z niezerową rotacją (dół)

Jak pamiętamy, operatory dywergencji i rotacji występują w różniczkowej wersji równań Maxwella. Jednakże bardziej intuicyjne z fizycznego punktu widzenia są wersje całkowe, które opisują dobrze znane zjawiska – indukcji elektromagnetycznej czy wytwarzania pola magnetycznego przez przewodniki z prądem. W jaki sposób uzyskać z całkowych praw elektromagnetyzmu ich „abstrakcyjne" postaci różniczkowe? Pomocne są tutaj dwa ważne twierdzenia matematyczne: twierdzenie Gaussa i twierdzenie Stokesa. Ogólnie rzecz ujmując, twierdzenia te umożliwiają utożsamienie całki po pewnej n-wymiarowej rozmaitości z całką po rozmaitości o wymiarze o jeden większym (n+1). Na przykład twierdzenie Gaussa łączy całkę po dwuwymiarowej powierzchni zamkniętej z całką po trójwymiarowym obszarze zamkniętym tą powierzchnią. Twierdzenie Stokesa z kolei porównuje całkę po jednowymiarowej zamkniętej krzywej z całką po dwuwymiarowym obszarze zamkniętym tą krzywą.

Do tej samej klasy twierdzeń (niektórzy uważają, że jest to tak naprawdę jedno ogólne twierdzenie) należy tzw. podstawowe twierdzenie rachunku różniczkowego i całkowego. Zapewne wielu z was zetknęło się z hasłem, że różniczkowanie jest „odwrotnością" całkowania. Oznacza to, że jeśli weźmiemy pochodną danej funkcji, a następnie obliczymy z tej pochodnej całkę, to uzyskamy z powrotem tę samą funkcję:[12]

[12] Ściśle rzecz biorąc, do funkcji $f(x)$ po prawej stronie równania należy dodać dowolną stałą C. Stałą tę można wyeliminować przez obliczenie całki oznaczonej w granicach of a do b, co da nam

$$\int \frac{df(x)}{dx} dx = f(x).$$

Całka po lewej stronie równania jest jednowymiarowa, co widać po infinitezymalnym elemencie długości *dx*. Natomiast funkcję po prawej stronie można zinterpretować jako całkę po zero-wymiarowej powierzchni, czyli punkcie. Skoro całka po danym obszarze z funkcji jest sumą iloczynów wartości tej funkcji razy mały element obszaru, to w wypadku punktu taka „zerowa całka" będzie po prostu samą wartością tej funkcji.

Zauważmy pewną ciekawą prawidłowość – aby wyrazić całkę „niższego rzędu" za pomocą całki o jeden stopień wyższej, należy zastosować w tej ostatniej operator różniczkowy. Oto schemat ogólnego twierdzenia: całka *n*-wymiarowa z dowolnej funkcji jest równa całce *n*+1-wymiarowej z pewnego operatora różniczkowego z tej funkcji. Pozostaje tylko określić, jaki to będzie operator. Weźmy przypadek przejścia z całki jednowymiarowej do dwuwymiarowej. Niech *O* będzie pewną krzywą zamkniętą, a *S* – powierzchnią wewnątrz tej krzywej. Rozważmy całkę po krzywej *O* z pewnego pola wektorowego $\boldsymbol{A}$. Wtedy odpowiednim operatorem różniczkowym będzie rotacja pola $\boldsymbol{A}$, a twierdzenie Stokesa, wyrażające połączenie między całkami po *O* i *S*, będzie miało postać:

$$\oint_O \boldsymbol{A} \cdot d\boldsymbol{l} = \int_S (\boldsymbol{\nabla} \times \boldsymbol{A}) \cdot d\boldsymbol{S}. \tag{4.13}$$

Zwróćmy uwagę, że w wyrażeniach pod całkami występuje znany nam już iloczyn skalarny wektorów, co oznacza, że będziemy całkować funkcje liczb rzeczywistych. W wypadku wyższych wymiarów operatorem umożliwiającym nam „przejście" z dwuwymiarowej powierzchni do trójwymiarowej objętości jest oczywiście dywergencja. Stąd mamy twierdzenie Gaussa (zwane również twierdzeniem Greena-Ostrogradzkiego-Gaussa):

$$\int_S \boldsymbol{A} \cdot d\boldsymbol{S} = \int_V \boldsymbol{\nabla} \cdot \boldsymbol{A} dV. \tag{4.14}$$

Jak zwykle, obszary całkowania są ze sobą powiązane w ten sposób, że objętość *V* jest zawarta w zamkniętej powierzchni *S*.

Możemy teraz zastosować powyższe zamienniki do równań Maxwella w wersji całkowej. Zacznijmy od pierwszego prawa elektrostatyki, zwanego również, choć myląco, prawem Gaussa.[13]

$$\int_S \boldsymbol{E} \cdot d\boldsymbol{S} = 4\pi Q.$$

Przypomnijmy: wyrażenie po lewej stronie oznacza całkowity strumień pola elektrycznego przepływający przez zamkniętą powierzchnię *S*, natomiast po prawej stronie mamy całkowity ładunek elektryczny *Q* zawarty w powierzchni *S*. Możemy teraz zastąpić całkę po powierzchni całką po objętości z dywergencji pola $\boldsymbol{E}$, wykorzystując tw. Gaussa (4.14). Natomiast ładunek *Q* da się obliczyć, całkując gęstość ładunku ρ po całej objętości *V*. Dostaniemy wtedy:

---

równanie $\int_a^b \frac{df(x)}{dx} dx = f(a) - f(b)$.

[13] Pamiętajmy, że twierdzenie Gaussa jest tezą matematyki, czyli prawdą aprioryczną, natomiast prawo Gaussa to teza empiryczna, która mogłaby być fałszywa w świecie fizycznym.

$$\int_V \boldsymbol{\nabla} \cdot \boldsymbol{E} dV = 4\pi \int_V \rho dV.$$

Zrównując funkcje pod całkami (co odpowiada „przechodzeniu do punktu"), otrzymujemy pierwsze równanie Maxwella w wersji różniczkowej (por. 4.7):

$$\boldsymbol{\nabla} \cdot \boldsymbol{E} = 4\pi\rho.$$

Pokażmy, jak wygląda analogiczne przejście w wypadku prawa Faradaya (4.3):

$$\frac{1}{c}\frac{d}{dt}\int_S \boldsymbol{B} \cdot d\boldsymbol{S} = -\oint_O \boldsymbol{E} \cdot d\boldsymbol{l}\,.$$

Korzystając z twierdzenia Stokesa (4.13), zastępujemy całkę po prawej stronie:

$$\frac{1}{c}\frac{d}{dt}\int_S \boldsymbol{B} \cdot d\boldsymbol{S} = -\int_S (\boldsymbol{\nabla} \times \boldsymbol{E}) \cdot d\boldsymbol{S}.$$

Wprowadzając operację różniczkowania po czasie „pod" całkę i jednocześnie zamieniając ją na pochodną cząstkową (ponieważ funkcja $\boldsymbol{B}$ zależy również od współrzędnych przestrzennych), dostajemy po opuszczeniu całek:

$$\boldsymbol{\nabla} \times \boldsymbol{E} = -\frac{1}{c}\frac{\partial \boldsymbol{B}}{\partial t}.$$

czyli kolejne równanie Maxwella w wersji różniczkowej. W analogiczny sposób otrzymamy pozostałe równania.

Przejdźmy teraz do wyprowadzenia równania falowego z praw Maxwella przy założeniu, że ładunki i prądy są zerowe. Oto te prawa dla przypomnienia:

$$\begin{aligned} \boldsymbol{\nabla} \cdot \boldsymbol{E} &= 0 \\ \boldsymbol{\nabla} \cdot \boldsymbol{B} &= 0 \\ \boldsymbol{\nabla} \times \boldsymbol{E} &= -\frac{1}{c}\frac{\partial \boldsymbol{B}}{\partial t} \\ \boldsymbol{\nabla} \times \boldsymbol{B} &= \frac{1}{c}\frac{\partial \boldsymbol{E}}{\partial t} \end{aligned} \tag{4.15}$$

Rozważmy czwarte z powyższych równań i zastosujmy do jego obu stron operację różniczkowania po czasie. Korzystając z faktu, że różniczkowanie po czasie jest przemienne z rotacją, otrzymujemy:

$$\boldsymbol{\nabla} \times \frac{\partial \boldsymbol{B}}{\partial t} = \frac{1}{c}\frac{\partial^2 \boldsymbol{E}}{\partial t^2}.$$

Podstawiając wyrażenie na $\frac{\partial \boldsymbol{B}}{\partial t}$ z trzeciego prawa, dostaniemy

$$-\boldsymbol{\nabla} \times (\boldsymbol{\nabla} \times \boldsymbol{E}) = \frac{1}{c^2}\frac{\partial^2 \boldsymbol{E}}{\partial t^2}. \tag{4.16}$$

Musimy teraz zastosować pewną matematyczną zależność między operacjami rotacji, dywergencji a jeszcze jedną operacją, zwaną operacją Laplace'a (w skrócie nazywaną laplasjanem). Definiuje się ją następująco:

$$\boldsymbol{\nabla}^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}.$$

Wspomniana zależność matematyczna wygląda tak:

$$\boldsymbol{\nabla} \times (\boldsymbol{\nabla} \times \boldsymbol{E}) = \boldsymbol{\nabla}(\boldsymbol{\nabla} \cdot \boldsymbol{E}) - \boldsymbol{\nabla}^2 \boldsymbol{E}. \tag{4.17}$$

Rozpisując pracowicie obie strony równania na współrzędne, przy pomocy powyższych definicji dywergencji i rotacji (4.11) i (4.12), można przekonać się, że istotnie taka równość zachodzi. Zostawmy tę kwestię dociekliwym czytelnikom do weryfikacji, natomiast zwróćmy może uwagę na jeden drobny fakt: operacja $\boldsymbol{\nabla}(\boldsymbol{\nabla} \cdot \boldsymbol{E})$ po prawej stronie równania nie jest tożsama z laplasjanem $\boldsymbol{\nabla}^2 \boldsymbol{E}$, mimo powierzchownego podobieństwa. Dywergencja $\boldsymbol{\nabla} \cdot \boldsymbol{E}$, jak wiadomo, daje nam skalar o postaci $\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z}$, natomiast zastosowanie do niego operatora $\boldsymbol{\nabla}$ (czyli operacji gradientu) produkuje wektor, którego przykładowa składowa *x* wygląda następująco:

$$\frac{\partial^2 E_x}{\partial x^2} + \frac{\partial^2 E_y}{\partial x \partial y} + \frac{\partial^2 E_z}{\partial x \partial z}.$$

Jest to oczywiście wynik różny od składowej *x* wektora $\boldsymbol{\nabla}^2 \boldsymbol{E}$, która ma postać:

$$\frac{\partial^2 E_x}{\partial x^2} + \frac{\partial^2 E_x}{\partial y^2} + \frac{\partial^2 E_x}{\partial z^2}.$$

Wracając do matematycznej zależności (4.17), zauważmy, że z pierwszego równania bezładunkowych i bezprądowych równań Maxwella (4.15) wynika, że $\boldsymbol{\nabla}(\boldsymbol{\nabla} \cdot \boldsymbol{E}) = 0$. Zatem podwójna rotacja z natężenia pola elektrycznego będzie po prostu równa laplasjanowi z minusem. Możemy więc wstawić ten laplasjan do wzoru (4.16), uzyskując w rezultacie:

$$\boldsymbol{\nabla}^2 \boldsymbol{E} = \frac{1}{c^2} \frac{\partial^2 \boldsymbol{E}}{\partial t^2}$$

Jest to nic innego jak równanie falowe dla wektora $\boldsymbol{E}$. W analogiczny sposób uzyskamy równanie dla wektora pola magnetycznego. Zatem istotnie równania Maxwella zawierają w sobie opis nowego zjawiska: fali elektromagnetycznej.

Można zatem rozpisać składowe pola elektrycznego (i magnetycznego) w danym punkcie jako pewne zaburzenie *F*, rozchodzące się w kierunku osi *z* i zależne od czasu *t*, jak to robiliśmy w poprzednim paragrafie:

$$E_x = \mathcal{E}_x F(z - vt),$$

$$E_y = \mathcal{E}_y F(z - vt),$$

$$E_z = \mathcal{E}_z F(z - vt).$$

Funkcja *F* może być np. dobrze znaną funkcją okresową *sinus* lub *cosinus*, a w takim wypadku liczby $\mathcal{E}_x$, $\mathcal{E}_y$, $\mathcal{E}_z$ będą odpowiednio amplitudami w kierunkach *x*, *y* i *z*. Zauważmy, że ponieważ składowe pola w kierunkach *x* i *y* zależą tylko od kierunku *z*, ich pochodne względem *x* i *y* są równe zeru:

$$\frac{\partial E_x}{\partial x} = 0,$$

$$\frac{\partial E_y}{\partial y} = 0.$$

Pamiętajmy jednak, że dywergencja pola $\boldsymbol{E}$ jest również równa zeru (pierwsze równanie w 4.15):

$$\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} = 0.$$

Wnioskujemy stąd, że

$$\frac{\partial E_z}{\partial z} = 0.$$

Jest to możliwe tylko wtedy, gdy $\mathcal{E}_z = 0$, a zatem cała składowa E_z jest zerowa. Wyprowadzamy stąd wniosek, że wektor pola elektrycznego jest zawsze prostopadły do kierunku propagacji fali. To samo dotyczy wektora pola magnetycznego, a dodatkowo wiemy, że wektor $\boldsymbol{E}$ musi być prostopadły do $\boldsymbol{B}$. Zatem przebieg fali elektromagnetycznej wygląda tak, jak to pokazano na rys 4.14.

Pytania i problemy

1. Czy fakt istnienia dwóch rodzajów sił elektrostatycznych (przyciągających i odpychających) implikuje istnienie dokładnie dwóch rodzajów ładunku elektrycznego? Rozważ teoretyczną możliwość istnienia trzech rodzajów ładunku, z których każde dwa różne się przyciągają, a takie same – odpychają. Jakie fakty przemawiają za fałszywością takiej hipotezy?

2. Wyjaśnij pojęcie pola elektrostatycznego. Jaki jest związek między przyjęciem realności pól elektromagnetycznych a problemem nielokalności oddziaływań elektrostatycznych?

3. Porównaj problem realności pola elektrostatycznego z filozoficznym problemem realności własności dyspozycyjnych. Jakiego rodzaju własnością dyspozycyjną mogłoby być pole?

4. Omów podstawowe argumenty za nierealnością potencjału elektrostatycznego oraz linii sił pola.

5. W jaki sposób można wyprowadzić prawo Coulomba (zależność natężenia pola od odwrotności kwadratu odległości) z fałszywego założenia, że pole elektryczne powstaje w wyniku wypływu nieściśliwego płynu z ładunków?

6. Czy istnieje prawo analogiczne do prawa Coulomba z elektrostatyki, opisujące oddziaływania magnetyczne? Co jest przyczyną odmiennego traktowania pola elektrycznego i magnetycznego?

7. W jaki sposób pole magnetyczne wpływa na ładunki elektryczne? Jaki jest warunek konieczny, aby ładunek elektryczny „odczuł" działającą na niego siłę magnetyczną?

8. Omów poglądowo pojęcia krążenia pola wzdłuż krzywej zamkniętej oraz strumienia pola przez powierzchnię. Czy zerowe krążenie wokół danej krzywej oraz zerowy strumień przez powierzchnię zamkniętą oznaczają, że pole jest również zerowe?

9. Omów zjawisko indukcji elektromagnetycznej oraz opisujące je prawo Faradaya. Czy matematyczna forma prawa Faradaya zawiera informację na temat związków przyczynowych między polem magnetycznym a elektrycznym?

10. Dlaczego prawo Ampère'a, opisujące powstanie pola magnetycznego wywołanego przepływem ładunków wymaga modyfikacji? Jakie są dwa możliwe źródła pola magnetycznego według prawa Ampère'a-Maxwella?

11. Przedstaw treść wszystkich czterech praw elektromagnetyzmu Maxwella. Jaka jest różnica między wersją całkową a różniczkową równań Maxwella?

12. Czy istnienie praw łączących zjawiska elektryczne i magnetyczne oznacza, że elektryczność i magnetyzm powinny być zunifikowane? Co oznacza unifikacja pewnego rodzaju zjawisk?

13. Jak zachowują się pola elektryczne i magnetyczne przy przejściu z jednego układu inercjalnego do innego? Podaj przykłady takiego zachowania. Jaki to ma związek z problemem unifikacji tych pól?

14. W jaki sposób można matematycznie opisać pole elektromagnetyczne? Jak taki opis rozwiązuje problem zmienności pól pod wpływem zmiany układu odniesienia?

15. Przedstaw w ogólnych zarysach rozumowanie prowadzące do hipotezy fal elektromagnetycznych. W jaki sposób hipoteza ta została zweryfikowana? Jaki wpływ na status teorii elektromagnetyzmu miało odkrycie istnienia fal elektromagnetycznych?

16. Czy istnienie fal elektromagnetycznych dostarcza argumentu w sporze pomiędzy realistyczną a instrumentalistyczną interpretacją pól?

17. Jakie są podstawy dla twierdzenia, że światło jest falą elektromagnetyczną? Jaka trudność pojawia się przy próbie interpretacji stałej *c* jako prędkości światła?

Literatura uzupełniająca

Historia rozwoju pojęć i teorii elektromagnetyzmu, ze szczególnym uwzględnieniem aspektu doświadczalnego, została pięknie opisana w rozdziałach 10. i 13. cytowanego już dzieła: A.K. Wróblewski, *Historia fizyki*, PWN, Warszawa 2007.

Bardzo dobre omówienie teorii elektromagnetyzmu i jej implikacji filozoficznych znajduje się w książce: M. Lange, *An Introduction to the Philosophy of Physics. Locality, Fields, Energy and Mass*, Blackwell, Oxford 2002.

Godna polecenia jest analiza podstawowych praw elektromagnetyzmu wraz z przejrzystym wytłumaczeniem ich aparatu matematycznego w „kultowym" podręczniku *Feynmana wykłady z fizyki* (R.P. Feynman, R.B. Leighton, M. Sands), tom II, część 1, PWN, Warszawa 1974.

Opis pól elektrycznych i magnetycznych i ich związek z teorią względności jest przedmiotem rozdziału 3. w książce: A. Einstein, L. Infeld, *Ewolucja fizyki*, PWN, Warszawa 1962.

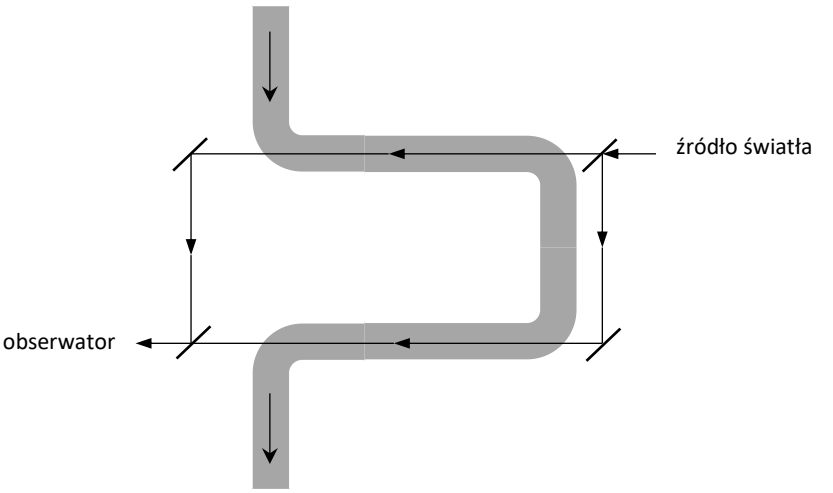
ROZDZIAŁ 5. SZCZEGÓLNA TEORIA WZGLĘDNOŚCI

5.1. Doświadczalne testy hipotezy eteru

Jak pamiętamy z poprzedniego rozdziału, najbardziej naturalną interpretacją prędkości fal elektromagnetycznych pojawiającej się w równaniach Maxwella jest prędkość względem ośrodka ich rozchodzenia się, czyli względem eteru (czasem zwanego również eterem światłonośnym). Jednakże niezależne potwierdzenie hipotezy eteru okazało się problematyczne. Spodziewanym efektem istnienia eteru powinna być zależność mierzonej prędkości światła od stanu ruchu obserwatora. Rozważmy przykład fal dźwiękowych rozchodzących się w powietrzu. Dla obserwatora spoczywającego względem powietrza prędkość dźwięku jest równa ok. 340 m/s. Natomiast obserwator poruszający się w powietrzu zarejestruje inną prędkość – większą lub mniejszą, w zależności od kierunku ruchu. W szczególności podczas przekraczania bariery dźwięku przez ponaddźwiękowy odrzutowiec fala dźwiękowa rozchodząca się w kierunku lotu będzie miała prędkość zerową względem samolotu. Wszystko to wynika oczywiście ze znanego prawa składania prędkości: aby obliczyć prędkość danego obiektu relatywnie do poruszającego się układu odniesienia, należy odjąć wektorowo prędkość tego układu od prędkości obiektu względem układu spoczywającego.

W przypadku eteru nie umiemy jednak wyznaczyć naszej prędkości względem tej substancji, gdyż jest ona niewidoczna i nie oddziałuje bezpośrednio na nasze instrumenty pomiarowe. Musimy zatem odwołać się do metod pośrednich, wykorzystujących promienie świetlne, które będą naszym „detektorem". Pierwszym pytaniem jest, czy poruszające się ciała fizyczne są w stanie „pociągać" za sobą niewidzialny eter. To ważne, gdyż jeśli efekt pociągania występuje, mogą być problemy z zaobserwowaniem zmienności prędkości światła w zależności od stanu ruchu obserwatora, skoro za każdym razem warstwa eteru będzie podróżować razem z obserwatorem. Hipoteza pociągania eteru została poddana eksperymentalnemu testowi w tzw. doświadczeniu Fizeau. Pytaniem postawionym przez Fizeau było, czy ruch wody ma wpływ na eter. Jeśli woda pociąga za sobą eter, to prędkość światła przepuszczanego „pod prąd" powinna być inna niż prędkość światła „z prądem". Skoro światło porusza się z prędkością c względem eteru, a on podróżuje razem z wodą z prędkością v, to prędkość światła w laboratorium powinna wynosić $c - v$ dla przypadku „pod prąd" i $c + v$ w sytuacji „z prądem".

Układ eksperymentalny wykorzystany w doświadczeniu Fizeau jest przedstawiony na rys. 5.1. Dwa promienie świetlne zostały przepuszczone przez rurki z płynącą wodą – raz w kierunku zgodnym z ruchem wody, a raz w przeciwnym. Ze względu na ogromną różnicę między prędkością wody v a prędkością światła c, różnice czasów przejścia dla obu promieni były znikome, jednak Fizeau wykorzystał tu bardzo czułe zjawisko interferencji. Rezultat doświadczenia był zasadniczo negatywny – przewidywany teoretycznie efekt pociągania eteru nie został potwierdzony. Można zatem przyjąć, że eter nie wykazuje skłonności do oddziaływania z ciałami materialnymi, inaczej niż gęstsze ośrodki jak woda czy powietrze.



Rys. 5.1. Schemat doświadczenia Fizeau

Drugą możliwość w kwestii zachowania eteru wyraża tzw. hipoteza morza eteru. Zakłada ona, że eter jest jedną całością, której fragmenty nie poruszają się względem siebie, a przedmioty materialne po prostu „przenikają" przez niego, nie zmieniając jego stanu ruchu. W takim ujęciu eter byłby czymś w rodzaju absolutnej przestrzeni Newtona, jako że wyznaczałby stan absolutnego spoczynku. Jeśli hipoteza ta jest prawdziwa – a w sytuacji nieistnienia efektu „pociągania" wydaje się ona jedyną logiczną możliwością – to powinniśmy być w stanie zaobserwować zmienność prędkości światła, zmieniając prędkość ruchu urządzenia pomiarowego (np. zmieniając jego kierunek ruchu). Ta obserwacja stała się podstawą słynnego doświadczenia Michelsona-Morleya. Przypatrzmy się bliżej jego teoretycznym podstawom, gdyż okażą się one ważne w dalszej części dyskusji. Rozważmy sytuację, w której promień świetlny wysłany z pewnego źródła dobiega do zwierciadła odległego o pewien dystans l, a następnie po odbiciu wraca do punktu wyjścia. Spróbujemy obliczyć czas, w jakim promień świetlny wykona całą podróż, w zależności od relatywnego ruchu układu względem eteru.

Pierwszy analizowany przypadek zakłada, że całe urządzenie porusza się względem eteru równolegle do biegu promienia z prędkością v (rys. 5.2). W takiej sytuacji prędkość promienia biegnącego w kierunku do lustra będzie wynosić $c - v$ (pamiętamy, że z założenia c jest prędkością światła względem eteru, a nie względem laboratorium), a w kierunku przeciwnym $c + v$. Łatwo zatem wyliczyć, że łączny czas przebiegu będzie następujący:

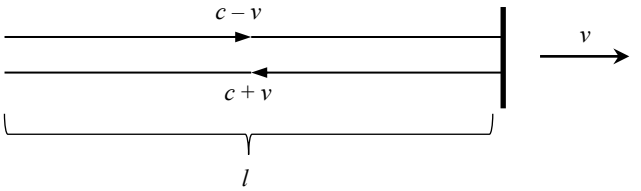
$$T_{\parallel} = \frac{l}{c-v} + \frac{l}{c+v} = \frac{2l}{c}\frac{1}{1-\frac{v^2}{c^2}}. \tag{5.1}$$

Drugi przypadek to sytuacja, w której promień świetlny porusza się prostopadle do kierunku ruchu względem eteru (rys. 5.3). Wtedy tor promienia z perspektywy eteru będzie wyglądał jak na rysunku (promień będzie biegł pod kątem w stosunku do pionowej linii). Otrzymujemy tutaj trójkąt prostokątny, którego przyprostokątne wynoszą *l* i *vt*, a przeciwprostokątna *ct* (gdyż jest to droga promienia względem eteru). Z twierdzenia Pitagorasa dostajemy

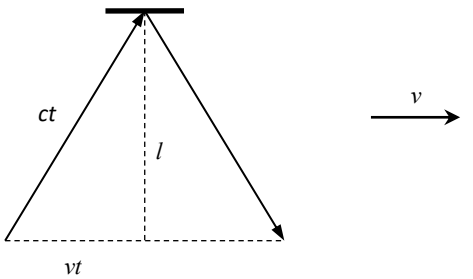
$$l^2 + (vt)^2 = (ct)^2.$$

Można stąd wyliczyć *t*, które jest połową czasu całkowitej podróży tam i z powrotem. Zatem całkowity czas $T_{\perp} = 2t$ wynosi:

$$T_{\perp} = \frac{2l}{c}\frac{1}{\sqrt{1-\frac{v^2}{c^2}}}. \tag{5.2}$$



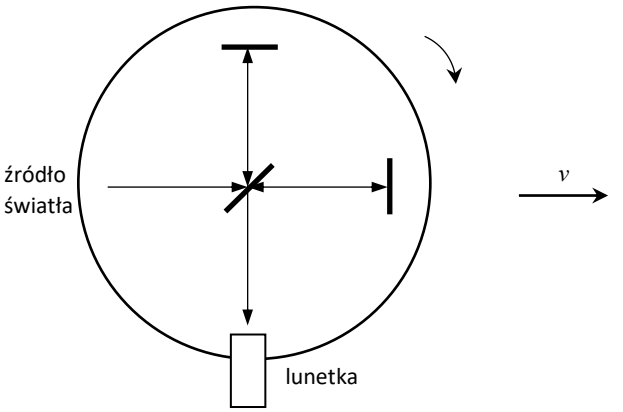
Rys. 5.2. Promień świetlny biegnący równolegle do ruchu układu względem eteru



Rys. 5.3. Promień świetlny biegnący prostopadle do kierunku ruchu względem eteru

Jak widać, czasy $T_{\parallel}$ i $T_{\perp}$ różnią się od siebie. Doświadczenie Michelsona-Morleya zostało zaprojektowane z myślą o zarejestrowaniu tej różnicy. Schemat układu doświadczalnego – tzw. interferometru Michelsona-Morleya – jest poglądowo przedstawiony na rys. 5.4. Promień świetlny wychodzący ze źródła przechodzi przez półprzepuszczalne zwierciadło, które rozdziela go na dwa promienie biegnące prostopadle do siebie. Po odbiciu od zwierciadeł dwa niezależne promienie świetlne ponownie się łączą, a efekt ich nałożenia się na siebie może być obserwowany za pomocą lunetki. Ponieważ światło jest falą, w efekcie nałożenia

dwóch ciągów fal zaobserwujemy zjawisko interferencji – wzmacniania i osłabiania natężenia w zależności od różnicy dróg optycznych. Istotne jest, że efekt interferencyjny bardzo mocno zależy od czasu dojścia obu promieni. Nawet niewielka różnica w czasie dojścia między ramionami interferometru spowoduje widoczne przesunięcie prążków interferencyjnych.



Rys. 5.4. Schemat interferometru Michelsona-Morley'a

Michelson i Morley umieścili swoje urządzenie na obrotowej podstawie, która umożliwiała płynną zmianę położenia obu ramion interferometru w stosunku do otoczenia. Ich rozumowanie było następujące: jeśli laboratorium porusza się względem eteru, to obracając całym układem, powinniśmy zmieniać czasy przejścia promieni w obu prostopadłych do siebie kierunkach, w zależności od ich położenia względem kierunku ruchu laboratorium. Te zmiany z kolei powinny być zaobserwowane w postaci przesunięcia prążków interferencyjnych. Jednakże mimo wielokrotnie powtarzanych obserwacji nie udało się wykryć żadnej zmiany w położeniu prążków interferencyjnych, a zatem także żadnej różnicy między czasem dojścia promienia świetlnego biegnącego wzdłuż kierunku ruchu układu a czasem dojścia promienia prostopadłego do kierunku ruchu.

Logiczna struktura doświadczenia Michelsona-Morleya jest następująca. Zakładamy, że eter jako całość spoczywa względem pewnego układu odniesienia (hipoteza morza eteru), a Ziemia razem z całym laboratorium i interferometrem porusza się względem niego. Z tej przesłanki wyprowadzamy wniosek, że czasy $T_{\parallel}$ i $T_{\perp}$ różnią się od siebie. Mimo to eksperymentalnie nie stwierdzamy żadnej różnicy. Jak zatem wytłumaczyć negatywny rezultat doświadczenia? Możliwych jest kilka strategii, które pokrótce omówimy. Najprostszym wyjaśnieniem jest odrzucenie przesłanki o ruchu Ziemi względem morza eteru. To jednak oznacza powrót do tezy o uprzywilejowanym statusie Ziemi w stosunku do innych ciał niebieskich, gdyż planety, Słońce i gwiazdy znajdują się w ruchu w stosunku do Ziemi, a więc także w stosunku do hipotetycznego eteru. Poza tym takie rozwiązanie można byłoby poddać definitywnej próbie eksperymentalnej. W tym celu należałoby przeprowadzić doświadczenie Michelsona-Morleya w laboratorium poruszającym się względem Ziemi – np. w samolocie albo na stacji kosmicznej okrążającej naszą planetę. Z tego co mi wiadomo, nikt nigdy nie wpadł na pomysł takiego eksperymentu, zapewne z powodu znikomej szansy na to, aby wykrył on ruch względem eteru.

Druga możliwość to powrót do hipotezy pociągania eteru przez ciała materialne, w tym przez powierzchnię Ziemi. Doświadczenie Fizeau pokazało, że hipoteza ta jest wysoce niewiarygodna, ale jest możliwe logicznie, że Ziemia jako całość działa w inny sposób na eter

niż woda przepływająca w rurce. Z drugiej strony, skoro większość powierzchni Ziemi zajmują oceany, byłoby dziwne, że eter „przykleja się" do nich podczas dobowego i rocznego ruchu naszej planety, natomiast nie wykazuje takiej tendencji w przypadku przepływu wody w laboratorium. Dodajmy, że takie rozwiązanie może być poddane kolejnemu testowi. W tym celu należałoby skonstruować interferometr, którego zwierciadła byłyby w stanie ruchu względem źródła światła. Wtedy eter byłby pociągany z inną prędkością w pobliżu zwierciadeł i przy źródle światła, a zatem promień świetlny musiałby przechodzić przez „warstwy" eteru o różnych względnych prędkościach, co dałoby obserwowalny rezultat w postaci zmiennych czasów dojścia. Nie sądzę jednak, aby jakikolwiek fizyk chciał przeprowadzić taki eksperyment, skoro istnieje znakomicie potwierdzona odpowiedź na negatywny rezultat doświadczenia Michelsona-Morleya w postaci szczególnej teorii względności.

Zanim jednak do tego przejdziemy, wspomnijmy na koniec o trzecim możliwym rozwiązaniu, które było poważnie rozważane przez fizyków. Chodzi tutaj o hipotezę skrócenia Lorentza-FitzGeralda. Dotyczy ona wpływu eteru na ciała w nim się poruszające. Być może eter wywołuje efekt „spłaszczenia" poruszających się w nim ciał w taki sposób, że droga promienia świetlnego w kierunku ruchu względem eteru ulega skróceniu o odpowiedni czynnik. Natomiast ramię interferometru, które jest prostopadłe do kierunku ruchu, nie zmienia swojej długości. Załóżmy, że w wyniku efektu skrócenia ciało o długości l, poruszające się z prędkością v względem eteru, będzie miało długość daną następującym wzorem:

$$l' = l\sqrt{1 - \frac{v^2}{c^2}}.$$

Wstawiając to wyrażenie zamiast l we wzorze (5.1) na czas „równoległy" $T_{\parallel}$, otrzymamy dokładnie taką samą formułę co w wypadku czasu „prostopadłego" $T_{\perp}$ (5.2). Zatem uzyskujemy zgodność z doświadczeniem.

Czy jednak jest to satysfakcjonujące rozwiązanie problemu? Nietrudno zauważyć, że hipoteza skrócenia ma poważne wady. Po pierwsze, jest to ewidentnie hipoteza *ad hoc*, wymyślona po to i w takiej formie, żeby „się zgadzało". Nie mamy żadnych niezależnych danych przemawiających za istnieniem fizycznego efektu skrócenia poza negatywnym rezultatem doświadczenia Michelsona-Morleya. Po drugie, jest niezwykle mało prawdopodobne, aby efekt opisany powyższym równaniem dotykał jednakowo wszystkie fizyczne obiekty, niezależnie od ich budowy, składu chemicznego itd. Sama idea wpływania eteru na poruszające się w nim obiekty nie jest zupełnie pozbawiona podstaw naukowych. Jeśli eter umożliwia przekazywanie oddziaływań elektromagnetycznych między ciałami naładowanymi elektrycznie, a przedmioty materialne są przecież w większości zbudowane z dodatnio naładowanych protonów i ujemnie naładowanych elektronów, to nie powinno być zaskoczeniem, że eter będzie oddziaływał na poruszające się w nim ładunki, a zatem także na makroskopowe ciała materialne. Jednakże kumulatywny efekt takiego oddziaływania w postaci obserwowalnych deformacji kształtu poruszających się obiektów powinien silnie zależeć od ich szczegółów budowy: gęstości, sposobu rozłożenia ładunków (np. tego, czy znajdują się w nim wolne elektrony, czy też pozostają one na powłokach atomowych), rodzaju wiązań chemicznych i wielu innych. Gdyby efekt skrócenia Lorentza-FitzGeralda rzeczywiście istniał, należałby on do rodzaju tzw. efektów uniwersalnych, które wymagają specyficznych wyjaśnień teoretycznych (z podobnym przykładem zetkniemy się przy okazji analizy oddziaływań grawitacyjnych w ogólnej teorii względności).

Warto w tym momencie wyjaśnić możliwe nieporozumienie. Zapewne wielu czytelników słyszało o efekcie skrócenia długości, wynikającym ze szczególnej teorii względności (skrócenie to również określa się w literaturze mianem skrócenia Lorentza lub Lorentza-FitzGeralda, co jest niestety mylące). Matematycznie efekt relatywistycznego skrócenia opisuje się powyższym równaniem, lecz jego sens fizyczny jest zupełnie inny. Po pierwsze, prędkość v występująca w równaniu relatywistycznym jest prędkością ciała względem dowolnego układu, a nie względem eteru. Po drugie, skrócenie obiektu jest obserwowalne wyłącznie z perspektywy tego układu, względem którego dany obiekt się porusza. W układzie poruszającym się razem z obiektem nie występują żadne fizyczne zjawiska skrócenia, które mogłyby być w jakikolwiek sposób zaobserwowane. Będziemy o tym szerzej mówić w następnych paragrafach rozdziału.

Po wyeliminowaniu hipotezy skrócenia zostaje nam zasadniczo tylko jedno wyjście: należy przyjąć, że eter nie istnieje, skoro jego przewidywany wpływ na prędkość rozchodzenia się fal elektromagnetycznych nie występuje. Mamy tu do czynienia z prostym rozumowaniem: jeśli zjawisko A z konieczności wywołuje efekt B, a nie istnieją żadne dodatkowe czynniki uniemożliwiające zajście B, to z faktu, że B nie zachodzi, wyciągamy logiczny wniosek, że A nie może zachodzić. W naszym wypadku A jest faktem istnienia eteru, a B występowaniem przesunięcia prążków interferencyjnych w doświadczeniu Michelsona-Morleya. W takiej jednak sytuacji wracamy do punktu wyjścia: względem czego należy mierzyć prędkość fal elektromagnetycznych, występującą w równaniach Maxwella? W tym momencie pojawia się niezwykła hipoteza Einsteina, że jest to prędkość względem *dowolnego* układu odniesienia. Einstein sugeruje, aby przyjąć, że w każdym układzie odniesienia prędkość światła jest jednakowa.

Zastanówmy się, jakie konsekwencje płyną z takiego zaskakującego założenia. Implikuje ono, że przejście z jednego układu odniesienia do drugiego, poruszającego się z pewną prędkością, nie zmienia obserwowalnej prędkości światła. Jest to jawnie niezgodne z intuicyjną regułą składania prędkości, wynikającą wprost z transformacji Galileusza, którą wyprowadziliśmy w paragrafie 2.4. Zatem aby zachować założenie Einsteina o stałości prędkości światła we wszystkich układach odniesienia, należy zmodyfikować reguły transformacji współrzędnych między inercjalnymi układami odniesienia. Zajmiemy się tym problemem w następnym paragrafie.

5.2. Względność równoczesności i transformacja Lorentza

Fundamentalnym założeniem fizyki klasycznej, widocznym w transformacji Galileusza ($t' = t$), jest teza, że współrzędna czasowa nie zależy od wybranego układu odniesienia. Mówiąc swobodnie, czas płynie tak samo, niezależnie od stanu ruchu obserwatora. Milcząco przyjmuje się ponadto, że określenie współrzędnej czasowej zdarzeń odległych od siebie przestrzennie nie nastręcza zasadniczych trudności. To założenie zostało jednak podważone w słynnym rozumowaniu Einsteina. Aby określić współrzędną czasową zdarzenia oddalonego od nas o lata świetlne, musimy dysponować uniwersalną metodą określenia równoczesności (w celu synchronizacji zegarów umożliwiających lokalizację czasową zdarzeń w dowolnym miejscu przestrzeni). Wiemy, że ze względu na skończoną prędkość rozchodzenia się światła, obserwowane zdarzenie nie jest równoczesne z aktem obserwacji, tylko od niego wcześniejsze. Na przykład widząc eksplozję supernowej w odległej galaktyce wiemy, że

miała ona miejsce setki tysięcy, a nawet miliony lat temu, gdyż tyle czasu zajmuje dotarcie do nas promienia świetlnego niosącego informację o wybuchu. Jak zatem możemy ustalić, które zdarzenia są naprawdę równoczesne z aktem obserwacji?

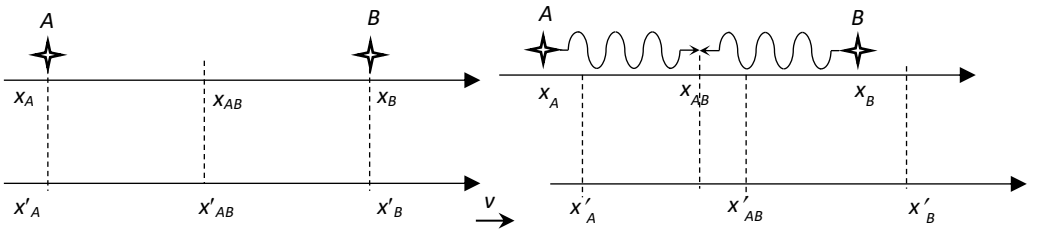
Na pierwszy rzut oka wydawać się może, że wystarczy w tym celu wziąć poprawkę na czas dojścia sygnału świetlnego. Znamy przecież prędkość rozchodzenia się światła – zarówno z doświadczenia, jak i teoretycznie, na podstawie analizy zjawisk elektromagnetycznych. Mierząc odległość między danym zdarzeniem (np. wspomnianym wcześniej wybuchem supernowej) a aktem obserwacji, możemy obliczyć opóźnienie i odjąć je od wskazań naszego zegara, uzyskując w ten sposób „prawdziwy" czas obserwowanego zjawiska. Na podstawie podobnego rozumowania stwierdziliśmy zresztą w poprzednim akapicie, że wybuch supernowej musiał nastąpić dużo wcześniej niż pojawienie się eksplozji na naszym niebie. Niestety, zaproponowana metoda ma jedną zasadniczą wadę. Jak pamiętamy z rozważań poświęconych pojęciom czasu i przestrzeni w fizyce newtonowskiej (w czasoprzestrzeni Galileusza), odległość przestrzenna między dwoma zdarzeniami nie jest w ogólności niezmiennicza względem transformacji Galileusza, jeśli tylko zdarzenia te nie są równoczesne. W jednym układzie odległość przestrzenna między nierównoczesnymi zdarzeniami może wynosić miliony lat świetlnych, a w drugim, równie dobrym, zero! Aby zastosować odpowiednią poprawkę, musielibyśmy już wiedzieć, że zdarzenia są równoczesne, ale to właśnie chcemy ustalić. Popadamy zatem w błędne koło.[1]

Mamy zatem impas. Z jednej strony wierzymy, że dla każdych zdarzeń odseparowanych przestrzennie jest obiektywnym faktem, czy są one równoczesne, czy też nie. Z drugiej strony nie dysponujemy metodą umożliwiającą stwierdzenie, czy taka równoczesność zachodzi. Einstein uważał, że zakładanie pojęcia, co do którego nie wiemy, jak je zastosować w praktyce, nie jest zgodne z zasadami metodologii naukowej. Dla tak fundamentalnego pojęcia, jakim jest pojęcie równoczesności, musimy znaleźć odpowiednie kryterium, rozstrzygające kwestię porównania czasu zachodzenia zdarzeń niezależnie od dzielących je odległości. Zaproponowane przez Einsteina kryterium sygnałowe jest bardzo proste i na pierwszy rzut oka intuicyjne, ale okazuje się mieć bardzo nieintuicyjną konsekwencję. Kryterium to jest następujące: dwa zdarzenia uznamy za równoczesne, gdy promienie świetlne wysłane z tych zdarzeń spotkają się dokładnie w połowie drogi między nimi. Wydaje się to bardzo zdroworozsądkowe. Jednak nie jest zdroworozsądkowe założenie, że kryterium sygnałowe możemy stosować niezależnie od tego, w jakim układzie się znajdujemy i z jaką prędkością się poruszamy.

Definicja równoczesności Einsteina jest jednym z koronnych przykładów tzw. definicji operacyjnych w nauce. Definicja operacyjna danego pojęcia polega na wskazaniu konkretnej metody jego zastosowania w postaci pewnej operacji, najczęściej pomiarowej. Istnieje kierunek w metodologii nauki zwany operacjonizmem – zapoczątkowany przez amerykańskiego fizyka P.W. Bridgmana – zgodnie z którym jedynie operacyjnie zdefiniowane terminy mają sens empiryczny i mogą być zaakceptowane w nauce. Pojęcia, dla

[1] Jeśli dwa porównywane zdarzenia zachodzą „na" przedmiotach trwających w czasie, które nie poruszają się względem siebie, to możemy przyjąć, że odległość przestrzenna między tymi zdarzeniami jest wyznaczona względem układu, w którym oba te przedmioty spoczywają. Co jednak z sytuacją, w której przedmioty są we względnym ruchu lub też zdarzenia w ogóle nie są związane z żadnymi trwającymi w czasie ciałami? Na przykład zdarzeniami mogą być błyski światła w próżni, bez żadnej „podstawy" materialnej. W takiej sytuacji każdy układ odniesienia wydaje się równie dobry.

których nie podano odpowiedniej metody zastosowania, nie powinny być przyjmowane. W wypadku pojęcia równoczesności zdarzeń w fizyce newtonowskiej nie istnieje dobrze określona metoda weryfikacji, czy dane dwa zdarzenia są równoczesne, a zatem pojęcie to nie spełnia warunku naukowości. Natomiast definicja Einsteina podaje taką metodę w postaci odpowiedniej operacji, której zastosowanie umożliwia wyznaczenie zdarzeń równoczesnych. Jednakże teza, iż pojęcia naukowe są jednoznacznie wyznaczone przez odpowiednie operacje pomiarowe, wydaje się zbyt radykalna. Wynika z niej, że dwie różne operacje pomiarowe definiują dwa różne pojęcia, nawet jeśli rezultaty takich pomiarów będą takie same. Na przykład pomiary długości za pomocą linijki i za pomocą miernika laserowego wyznaczają dwa różne pojęcia długości, co wydaje się nieintuicyjne.



Rys. 5.5. Względność równoczesności według Einsteina

Jak łatwo się przekonać, wyznaczenie punktu dzielącego na połowę odległość między dwoma zdarzeniami zależy od przyjętego układu odniesienia. Rozważmy dwa układy poruszające się względem siebie oraz dwa zdarzenia A i B (por. rys. 5.5). W jednym układzie połowa odległości między A i B wypada w punkcie o współrzędnej $x_{AB} = \frac{x_A + x_B}{2}$. W układzie poruszającym się względem pierwszego, współrzędna połowy odległości to $x'_{AB} = \frac{x'_A + x'_B}{2}$. Jednakże punkt x'_{AB} na osi drugiego układu nie będzie się przez cały czas pokrywał z punktem x_{AB}, gdyż drugi układ porusza się w prawo. W rezultacie jeśli promienie świetlne wychodzące z A i B zetkną się w punkcie czasoprzestrzennym o współrzędnej x_{AB}, punkt ten nie będzie miał współrzędnej x'_{AB} w układzie poruszającym się, lecz będzie „przesunięty" w kierunku położenia zdarzenia A (x'_A). Stosując kryterium Einsteina musimy więc przyznać, że zdarzenia A i B są równoczesne w pierwszym układzie, ale nie w drugim. Z punktu widzenia poruszającego się układu zdarzenie A zaszło wcześniej, gdyż promień świetlny wysłany z A przebył dłuższą drogę niż promień z B, zanim oba promienie się spotkały.

Warto podkreślić, odpowiadając na ewentualne zarzuty w stosunku do powyższego argumentu, że kryterium sygnałowe Einsteina opiera się na założeniu stałości prędkości światła we wszystkich układach odniesienia. To dzięki temu założeniu mamy prawo twierdzić, że kryterium to może być zastosowane w każdym układzie odniesienia. Gdyby do światła stosowało się zwykłe prawo składania prędkości (jak np. w wypadku fal dźwiękowych), należałoby odrzucić obserwację w poruszającym się układzie, jako że prędkości sygnałów świetlnych z A i B różniłyby się między sobą (sygnał dochodzący z A byłby szybszy, a z B wolniejszy). W takiej sytuacji kryterium sygnałowe mogłoby być stosowane tylko w wyróżnio-

nym układzie, w którym prędkość światła we wszystkich kierunkach wynosi c. Jak jednak już wiemy m.in. z doświadczenia Michelsona-Morleya, nie jesteśmy w stanie wyznaczyć takiego uprzywilejowanego układu odniesienia. Albo zatem trzymamy się klasycznego kryterium równoczesności, z którego nie będziemy mogli korzystać w praktyce, albo też „zadowolimy się" możliwym do zastosowania kryterium, które jednak implikuje względność równoczesności.

Przyjmując to drugie rozwiązanie, widzimy, że nie jest do utrzymania prosta tożsamościowa transformacja czasu z fizyki klasycznej: $t' = t$. Skoro dwa zdarzenia w jednym układzie odniesienia są równoczesne, czyli $t_A = t_B$, a w drugim nie ($t_A' \neq t_B'$), to współrzędna czasowa w układzie poruszającym się nie może być taka sama jak w układzie „spoczywającym". Potrzebujemy nowych reguł transformacji z jednego układu do drugiego. Odpowiednie reguły noszą nazwę transformacji Lorentza – stanowią one podstawę nowej teorii, czyli szczególnej teorii względności. Istnieje wiele interesujących metod wyprowadzenia tych transformacji z pewnych ogólnych zasad, włączając w to zasadę stałości prędkości światła. Nie będziemy jednak tego robić – zamiast tego podamy je w gotowej postaci, a następnie przekonamy się, że prowadzą do pożądanych rezultatów. Oto matematyczna postać transformacji Lorentza:

$$x' = \frac{x - vt}{\sqrt{1 - \frac{v^2}{c^2}}}$$
$$t' = \frac{t - \frac{vx}{c^2}}{\sqrt{1 - \frac{v^2}{c^2}}}. \tag{5.3}$$

Jak zwykle zakładamy, że układ odniesienia (x', t') porusza się względem układu (x, t) z prędkością v w kierunku rosnących wartości x. Zauważmy, że w obu równaniach pojawia się czynnik $\sqrt{1 - \frac{v^2}{c^2}}$, znany wcześniej z analizy doświadczenia Michelsona-Morleya. Jest on w pewnym sensie znakiem firmowym teorii względności – jego obecność zawsze wskazuje, że opuszczamy dobrze znany teren fizyki klasycznej i zapuszczamy się do świata przedziwnych efektów relatywistycznych. Liczbowa wartość tego czynnika zależy od tego, jak duża jest prędkość v w stosunku do prędkości światła. Jeśli prędkość v jest niewielka – taka jak prędkość poruszających się wokół nas obiektów (samochodów, pociągów, a nawet samolotów odrzutowych czy rakiet kosmicznych) – to współczynnik ten jest bardzo bliski jedności. W takiej sytuacji pierwsze równanie przechodzi w dobrze znaną regułę transformacji Galileusza: $x' = x - vt$.

A co z drugim równaniem, charakteryzującym współrzędną czasową? W liczniku pojawia się jeszcze czynnik $\frac{vx}{c^2}$. Można powiedzieć, że iloraz $\frac{v}{c^2}$ dla niewielkich prędkości v będzie niezmiernie małą liczbą (prędkość światła podniesiona do kwadratu jest ogromną wielkością), a zatem możemy pominąć całe wyrażenie. To jednak jest niezupełnie poprawne rozumowanie. Odpowiednio duża wartość x może „zrekompensować" ogromną liczbę w mianowniku, co spowoduje, że czas t' w układzie poruszającym się będzie znacząco różny od czasu t. Zatem nawet gdy przechodzimy między układami poruszającymi się relatywnie

wolno, współrzędne czasowe zdarzeń mogą się od siebie istotnie różnić, jeśli tylko rozważymy zdarzenia bardzo od siebie odległe.

Podsumowując powyższe obserwacje, możemy stwierdzić, że jeśli ograniczymy się do wolno poruszających się układów i nieodległych zdarzeń, reguły transformacji Lorentza przechodzą w transformacje Galileusza. Tłumaczy to, dlaczego przez wieki ludzie nie byli w stanie zaobserwować żadnego odstępstwa od zasad fizyki klasycznej. Widzimy także, że szczególna teoria względności nie miała szans powstać jako proste uogólnienie wielu obserwacji. Stosując zasadę indukcji, musielibyśmy uznać, że skoro mechanika klasyczna sprawdza się w ogromnej liczbie przypadków, jest ona uniwersalnie prawdziwa. Dopiero głęboki kryzys pojęciowy u podstaw teorii elektromagnetyzmu pokazał, że potrzebna była rewizja fundamentalnych praw dotyczących czasu i przestrzeni.

Pokażmy teraz, że nowe reguły transformacji potwierdzają zasadę stałości prędkości światła. W tym celu wyprowadzimy nowe prawo składania prędkości, umożliwiające obliczenie prędkości danego obiektu (w tym światła) w układzie poruszającym się z pewną prędkością. Niech x/t będzie prędkością ciała równą V wyznaczoną w jednym układzie. Obliczmy teraz stosunek x'/t', reprezentujący prędkość tego ciała w innym układzie odniesienia, poruszającym się względem pierwszego z prędkością v:

$$\frac{x'}{t'} = \frac{x - vt}{t - \frac{vx}{c^2}} = \frac{\frac{x}{t} - v}{1 - \frac{vx}{c^2 t}}.$$

Zatem relatywistyczna reguła składania prędkości będzie następująca:

$$V' = \frac{V - v}{1 - \frac{vV}{c^2}}.$$

Załóżmy, że prędkość V względem danego układu jest równa c. Otrzymujemy wtedy:

$$V' = \frac{c - v}{1 - \frac{v}{c}} = c\frac{c - v}{c - v} = c\,.$$

Prędkość światła w układzie poruszającym się z prędkością v będzie nadal wynosiła c. Światło rozchodzi się ze stałą prędkością, niezależnie od układu odniesienia.

Zauważmy jeszcze jedną matematyczną konsekwencję transformacji Lorentza, która ma doniosłe znaczenie fizyczne. Co się stanie, jeżeli założymy, że układ u' porusza się z prędkością światła? Podstawienie za v prędkości c we wzorach (5.3) na x' i t' daje zero w mianowniku, czyli niewykonalną matematyczną operację (dzielenie przez zero). Zwykle interpretuje się ten fakt jako pokazujący, że niemożliwe jest istnienie układu odniesienia poruszającego się z prędkością światła. Ciała materialne mogą zbliżać się dowolnie do prędkości światła, ale nie mogą jej osiągnąć.

5.3. Dwa relatywistyczne efekty

Jestem pewien, że każdy czytelnik musiał słyszeć o dwóch zdumiewających efektach przewidywanych przez teorię względności: dylatacji czasu i skróceniu długości. Terminy te przeniknęły do popularnej kultury i literatury *science-fiction*. Niestety, co za tym często idzie, pojawiła się masa nieporozumień i błędnych interpretacji w kwestii natury owych efektów.

W niniejszym paragrafie pokażemy, jak istnienie obu efektów można wyprowadzić bezpośrednio z transformacji Lorentza. Zobaczymy też dokładniej, na czym polega owo wydłużenie czasu i skrócenie długości.

Zacznijmy od zjawiska dylatacji czasu. Wyobraźmy sobie pewien proces trwający przez określony interwał od 0 do t – na przykład przesypywanie piasku z jednej komory klepsydry do drugiej. Ponieważ wiemy, że pomiary czasu i przestrzeni muszą być jawnie zrelatywizowane do jakiegoś układu, wybierzmy układ, w którym klepsydra spoczywa. Czas dla danego obiektu mierzony w układzie, w którym ten obiekt spoczywa, nazywamy *czasem własnym* owego obiektu. Rozważmy teraz ten sam proces z perspektywy układu poruszającego się względem klepsydry. Proces ten będzie trwał od chwili 0 do t', gdzie czas t' jest powiązany ze współrzędnymi w pierwszym układzie drugim równaniem transformacji Lorentza (5.3). Dzięki założeniu o spoczywaniu klepsydry w pierwszym układzie możemy przyjąć, że położenie zarówno początku, jak i końca procesu wynosi $x = 0$. Zatem transformacja czasu przyjmuje prostą postać:

$$t' = \frac{t}{\sqrt{1 - \frac{v^2}{c^2}}}.$$

Czas mierzony w układzie poruszającym się względem danego ciała będzie zawsze dłuższy od czasu własnego tego ciała, ponieważ dzielimy liczbę t przez czynnik mniejszy od jedności. Efekt dylatacji czasu został potwierdzony doświadczalnie m.in. w eksperymentach z nietrwałymi cząstkami elementarnymi. Na podstawie pomiarów w laboratorium określono czas życia dla cząstki danego typu spoczywającej względem układu pomiarowego. Taką samą cząstkę następnie zidentyfikowano w promieniowaniu kosmicznym, poruszającym się z ogromną prędkością w stosunku do Ziemi. Czas, jaki upłynął do momentu rozpadu dla cząstki w pędzącym strumieniu, okazał się dłuższy od uprzednio zmierzonego czasu życia w stanie spoczynku, zgodnie z powyższą formułą.

Wyprowadzenie wzoru na skrócenie długości wymaga nieco ostrożności. Przede wszystkim zdefiniujmy precyzyjnie, na czym polega pomiar długości np. pręta. W układzie, w którym pręt spoczywa, sprawa jest oczywista – stosujemy zwykłą miarkę, aby określić jego długość w centymetrach, calach czy jakichkolwiek jednostkach. Natomiast problemem jest, jak zmierzyć długość pręta poruszającego się względem nas. Nie możemy go po prostu zatrzymać, bo w ten sposób wrócimy do przypadku układu, w którym pręt spoczywa. Musimy wymyślić jakąś metodę mierzenia długości pręta bez wpływania na jego ruch. Rozwiązaniem może być wyznaczenie momentalnego położenia początku i końca pręta w naszym układzie np. za pomocą fotokomórki. Jednakże kluczowe jest tutaj, aby oba pomiary dokonały się równocześnie. Nic nam nie przyjdzie z tego, jeśli w jednej chwili odznaczymy położenie końca pręta, a za parę sekund położenie początku – uzyskany wynik nie będzie odzwierciedlał długości, bo przecież pręt się porusza. Przyjmijmy zatem, że współrzędna czasowa obu pomiarów w naszym układzie wynosi $t = 0$. Konwencjonalnie możemy również założyć, że zarejestrowana współrzędna przestrzenna początku pręta x_0 jest równa 0, a wtedy współrzędna końca x_1 będzie po prostu zmierzoną długością pręta w ruchu (nazwijmy ją l_r). Z kolei w układzie, w którym pręt spoczywa (a zatem jest to układ, który porusza się z prętem względem nas), współrzędne przestrzenne początku i końca pręta mogą być ustalone na $x_0' = 0$ i $x_1' = l_s$, gdzie l_s – długość spoczywającego pręta (rys. 5.6). Z pierwszej transformacji Lorentza

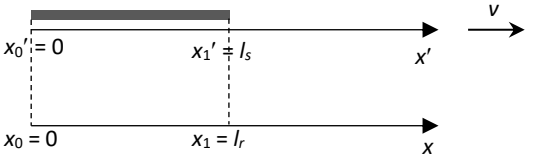
(5.3) mamy następujący wzór, łączący współrzędną x_1' ze współrzędnymi w naszym układzie (zgodnie z założeniem kładziemy $t = 0$):

$$x_1' = \frac{x_1}{\sqrt{1 - \frac{v^2}{c^2}}}.$$

Stąd dostajemy wprost wzór na długość pręta w ruchu:

$$l_r = l_s \sqrt{1 - \frac{v^2}{c^2}}.$$

Widać więc, że wynik pomiaru w ruchu będzie mniejszy od długości pręta w spoczynku. Pręt ulega „skróceniu".



Rys. 5.6. Pomiar długości pręta w ruchu

Czy zjawiska dylatacji czasu i skrócenia długości są realne? Ale co to znaczy, że zjawisko jest realne? Efekty skrócenia długości przedmiotów i wydłużenia czasu trwania procesów pojawiają się podczas przeprowadzania dokładnych i obiektywnych pomiarów przy pomocy odpowiednich urządzeń, a zatem nie są to „złudzenia" czy „iluzje", podobne do dobrze znanych złudzeń zmysłowych w rodzaju efektu Müllera-Lyera. Z drugiej jednak strony, należy wystrzegać się naiwnej i z gruntu fałszywej interpretacji owych relatywistycznych efektów jako wynikających z przyczynowego wpływu ruchu na stan fizyczny przedmiotów. Spotyka się niekiedy sugestię, że przedmioty poruszające się z prędkością bliską prędkości światła będą poddane szczególnemu fizycznemu wpływowi wynikającemu z tego ruchu. Procesy fizyczne, chemiczne czy biologiczne ulegną spowolnieniu, a same przedmioty skurczą się w kierunku ruchu. Czasem dodaje się do tego, że efekty te nie będą mogły być zaobserwowane w poruszającym się układzie, gdyż urządzenia pomiarowe (zegary, pręty miernicze) same ulegną analogicznym deformacjom (jeśli np. zmierzymy zwolniony proces za pomocą wolno chodzącego zegara, rezultat pomiaru będzie niezmieniony). Efekty relatywistyczne można obserwować jedynie „z zewnątrz", gdyż wtedy urządzenia pomiarowe nie są zaburzone.

Jest to błędna interpretacja z prostego powodu – ruch nie jest własnością absolutną obiektów, a jedynie relatywną. Nie ma sensu mówić, że rakieta „naprawdę" się porusza, a Ziemia spoczywa – równie dobrze możemy stwierdzić, że rakieta stoi, a Ziemia pędzi w przeciwnym kierunku. W związku z tym nie możemy twierdzić, że ruch będzie obiektywnie wpływał na zachodzące procesy. Najlepiej zilustrować to na przykładzie efektu skrócenia Lorentza-Fitz-Geralda. Jak pamiętamy, hipoteza skrócenia pojawiła się przy okazji negatywnego rezultatu doświadczenia Michelsona-Morleya jako próba uratowania założenia o istnieniu eteru. Klu-

czowe dla tej hipotezy było twierdzenie, że przedmioty „naprawdę" i „obiektywnie" się poruszają. Prawdziwym ruchem jest ruch względem eteru i to on jest odpowiedzialny za fizyczne „spłaszczenie" ciał. Natomiast relatywistyczne skrócenie długości, omówione w niniejszym paragrafie, stosuje się do każdych dwóch układów pozostających we wzajemnym ruchu. Z punktu widzenia Ziemi skróceniu ulegają przedmioty znajdujące się w rakiecie, ale z punktu widzenia rakiety to ziemskie obiekty stają się krótsze. Żaden z tych efektów nie jest „bardziej realny" od drugiego. Być może bardziej adekwatnym opisem tego, co się dzieje w opisywanych przez nas sytuacjach, jest stwierdzenie, że to nie przedmioty i procesy fizyczne ulegają deformacjom, ale czas i przestrzeń, które zależą od przyjętego układu odniesienia.

Najbardziej spektakularnym przykładem relatywistycznych deformacji czasu i przestrzeni jest tzw. paradoks bliźniąt. Tutaj nie powinno być wątpliwości, że obserwowany efekt jest jak najbardziej realny. Oto mamy przed sobą parę bliźniaków, którzy urodzili się tego samego dnia, a teraz jeden z nich jest młodym mężczyzną, a drugi starcem. Trudno o bardziej obiektywny skutek relatywistycznych transformacji. Przeanalizujemy ten przypadek dokładniej w następnym paragrafie, kiedy wprowadzimy pojęcie interwału czasoprzestrzennego.

5.4. Geometria czasoprzestrzeni Minkowskiego

Szczególna teoria względności jest w pewnym sensie nową teorią czasu i przestrzeni czy też łącznie czasoprzestrzeni. Nie będzie zatem zaskoczeniem, że nowatorstwo tej teorii powinno przejawiać się w modyfikacji geometrycznych własności czasoprzestrzeni. Jak już wspominaliśmy przy okazji analizy własności czasoprzestrzeni Galileusza, geometria danej przestrzeni jest charakteryzowana przez kompletny zestaw jej „inwariantów", czyli własności metrycznych nieulegających zmianie przy zmianie układu współrzędnych. W wypadku geometrii Euklidesowej takim inwariantem, czy też niezmiennikiem, jest zwykła odległość przestrzenna między dwoma punktami, którą można zapisać dla przypadku trzech wymiarów w wersji „kwadratowej", jak następuje:

$$\Delta s^2 = \Delta x^2 + \Delta y^2 + \Delta z^2.$$

Geometria czasoprzestrzeni Galileusza jest nieco bardziej skomplikowana. Występują w niej dwa niezmienniki: odległość przestrzenna między punktami posiadającymi tę samą współrzędną czasową (równoczesnymi) oraz odległość czasowa między dowolnymi punktami. Teoria względności charakteryzuje się jeszcze inną geometrią, zwaną geometrią Minkowskiego. Operuje ona jednym inwariantem, którego postać jest do pewnego stopnia analogiczna do odległości Euklidesowej. Inwariant ten nosi nazwę interwału czasoprzestrzennego:

$$\Delta I^2 = (c\Delta t)^2 - \Delta x^2 - \Delta y^2 - \Delta z^2. \tag{5.4}$$

Można się przekonać, że wyrażenie powyższe pozostanie niezmienione pod wpływem transformacji Lorentza. Wystarczy w tym celu rozpisać wyrażenie $(ct')^2 - x'^2$, korzystając z transformacji Lorentza na x' i t', aby przekonać się, że rezultat będzie równy $(ct)^2 - x^2$, czyli taki sam jak interwał w układzie u. Uwaga: obliczenia te można znakomicie uprościć, przyjmując powszechnie stosowaną konwencję $c = 1$. Innymi słowy, stosujemy jednostki, w których prędkość światła wynosi 1 (czyli np. mierzymy odległość przestrzenną w latach

świetlnych). Wtedy transformacje Lorentza przyjmują dużo prostszą i bardziej „symetryczną" postać:

$$x' = \frac{x - vt}{\sqrt{1 - v^2}},$$
$$t' = \frac{t - vx}{\sqrt{1 - v^2}}.$$

Łatwo teraz policzyć, że interwał w układzie „primowanym" będzie dokładnie równy interwałowi w układzie „nieprimowanym":

$$t'^2 - x'^2 = \frac{1}{1 - v^2}[(t - vx)^2 - (x - vt)^2] = \frac{1}{1 - v^2}(t^2 - x^2)(1 - v^2) = t^2 - x^2 .$$

Jak widać z formuły (5.4), interwał czasoprzestrzenny inaczej „traktuje" współrzędne czasowe i przestrzenne: kwadrat interwału przestrzennego wchodzi do równania ze znakiem minus, a kwadrat interwału przestrzennego ze znakiem dodatnim. Rozważmy teraz trzy przypadki. W pierwszym zakładamy, że odległość czasowa między dwoma zdarzeniami „przeważa" nad odległością przestrzenną": $\Delta t^2 > \Delta s^2$ (pamiętajmy o umowie $c = 1$). Innymi słowy, kwadrat interwału czasoprzestrzennego ΔI^2 jest liczbą dodatnią. Fizyczna interpretacja takiej sytuacji jest prosta: dwa zdarzenia (czy też punkty czasoprzestrzenne) są oddzielone od siebie mniejszą odległością przestrzenną niż dystans, jaki w czasie dzielącym oba zdarzenia pokonuje promień świetlny. Innymi słowy, oba zdarzenia mogą zostać połączone sygnałem poruszającym się z prędkością podświetlną.[2] Oznacza to również, że będzie istniał układ odniesienia, poruszający się z prędkością mniejszą niż *c*, w którym dwa rozważane zdarzenia zachodzą w tym samym miejscu (ich odległość przestrzenna będzie równa zeru). Interwał czasoprzestrzenny, spełniający powyższe warunki, nazywamy interwałem czasopodobnym (*time-like interval*).

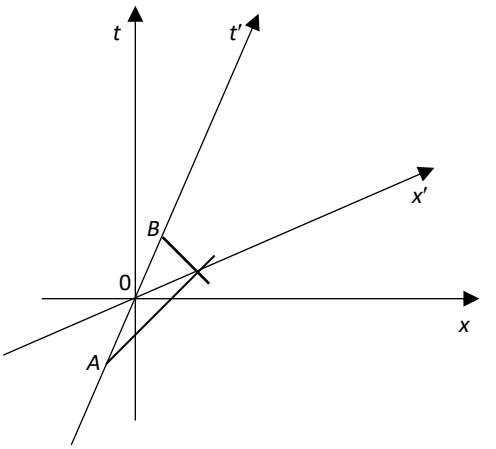
Drugi przypadek zachodzi, gdy $\Delta I^2 < 0$. Mamy wtedy do czynienia z interwałem przestrzennopodobnym (*space-like*). Zdarzenia rozdzielone takim interwałem są od siebie zbyt odległe przestrzennie, aby mogły być połączone sygnałem świetlnym. Natomiast można pokazać, że w takim wypadku istnieje układ odniesienia, w którym zdarzenia te będą równoczesne, czyli ich odległość czasowa Δt będzie równa zeru. Niekiedy nazywa się taką sytuację zachodzeniem relacji *quasi*-równoczesności. Chociaż relacja równoczesności nie jest niezmiennicza względem transformacji Lorentza, to relacja *quasi*-równoczesności już jest, jako zdefiniowana przy pomocy pojęcia interwału.

Trzecim, granicznym przypadkiem jest sytuacja, kiedy $\Delta I^2 = 0$. Dwa zdarzenia rozdzielone interwałem zerowym (zwanym także niekiedy światłopodobnym) mogą być połączone sygnałem biegnącym dokładnie z prędkością światła, gdyż ich odległość przestrzenna jest w każdym układzie równa drodze, jaką pokona promień świetlny w dzielącym je czasie. Interwały czasopodobne, przestrzennopodobne i zerowe wyznaczają pewien ważny podział czasoprzestrzeni względem danego zdarzenia. Żeby jednak omówić to dokładniej, musimy nauczyć się prostej, lecz użytecznej metody tworzenia diagramów czasoprzestrzennych.

[2] Fakt ten jest niezależny od wyboru układu odniesienia. Choć w różnych układach odniesienia wielkości interwałów przestrzennych i czasowych będą na ogół różne, to interwał czasoprzestrzenny jako inwariant pozostanie niezmieniony, a zatem i warunek $\Delta t > \Delta s$ będzie musiał być spełniony w każdym układzie.

Dzięki niej będziemy mogli łatwo i przejrzyście zobrazować wiele faktów i relacji między zdarzeniami, których opis w języku samej algebry może być mało czytelny.

Zacznijmy od graficznego przedstawienia osi dla danego układu odniesienia (rys. 5.7). Z oczywistych powodów będziemy mogli uwzględnić co najwyżej dwa wymiary przestrzenne, gdyż nie jesteśmy w stanie zobrazować na kartce papieru czterech wymiarów (czterech osi wzajemnie do siebie prostopadłych). Na razie uprośćmy sprawę jeszcze bardziej, rozważając tylko jeden wymiar przestrzenny x oprócz wymiaru czasowego t. Oś czasu wyznaczona będzie przez zbiór wszystkich punktów czasoprzestrzennych (zdarzeń), których współrzędna przestrzenna jest zerowa ($x = 0$). Analogicznie, oś przestrzenna jest dana równaniem $t = 0$. Każdy punkt w naszej dwuwymiarowej „czasoprzestrzeni" otrzyma swoje współrzędne (t, x) w zwykły sposób w układzie współrzędnych kartezjańskich.



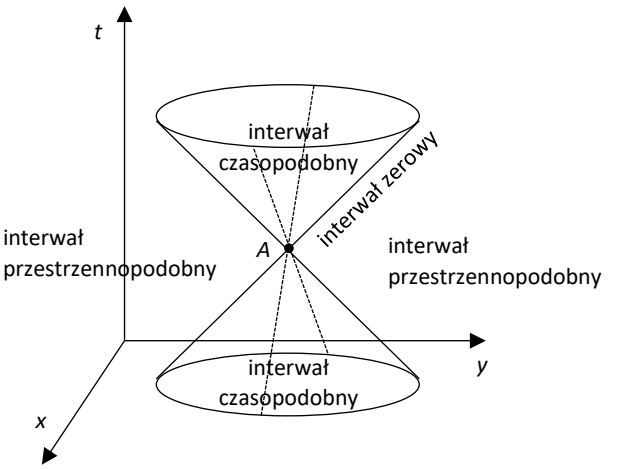
Rys. 5.7. Osie układów odniesienia poruszających się względem siebie

Postawmy teraz pytanie, jak na naszym diagramie przedstawić nowy układ odniesienia, który porusza się ze stałą prędkością w kierunku dodatnich wartości x („w prawo"). Z osią czasu t' nie będzie kłopotu – jest to linia wyznaczona jednostajnym ruchem punktu o zerowej współrzędnej w nowym układzie, a zatem będzie to linia prosta nachylona pod kątem do osi t. Na razie sytuacja wygląda tak samo jak w fizyce klasycznej. Różnica pojawi się, kiedy będziemy chcieli narysować oś x', czyli zbiór tych wszystkich zdarzeń, które mają współrzędną t' równą zeru. W czasoprzestrzeni Galileusza oś x' pokrywa się z osią x, gdyż czas nie ulega zmianie przy przejściu do innego układu odniesienia. Natomiast w szczególnej teorii względności musimy uwzględnić Einsteinowskie kryterium równoczesności. Zdarzenia na osi x' powinny być równoczesne ze zdarzeniem 0 na osi t'. Aby je wyznaczyć, zastosujmy alternatywny wariant kryterium Einsteina, równoważny z wersją przedstawioną we wcześniejszym paragrafie. Kryterium to jest następujące: wysyłamy promień świetlny w czasie t_A w kierunku lustra, promień odbija się od lustra i wraca do punktu wyjścia w czasie t_B. Uznajemy, że czas dojścia promienia do lustra jest równy dokładnie połowie czasu między t_A i t_B: $\frac{t_A+t_B}{2}$.

Wybierzmy teraz dowolny punkt na osi czasu t', wcześniejszy od punktu zero. Jak narysować tor promienia świetlnego na diagramie? To proste: jest to linia nachylona do osi x pod kątem 45 stopni (w przyjętych jednostkach światło biegnie z prędkością 1). Narysujmy zatem

taką linię z wybranego punktu *A*. Z kolei z punktu *B*, odległego od punktu 0 o ten sam interwał czasowy co punkt *A*, wyprowadźmy prostą („linię światła") w dół. Miejsce przecięcia obu prostych reprezentuje zdarzenie odbicia, które jest równoczesne w poruszającym się układzie z punktem zerowym. Zatem oś x' musi przechodzić przez ten punkt. Jak widzimy z rysunku, oś ta będzie nachylona w stosunku do osi *x*. Można dowieść na podstawie prostych rozważań geometrycznych, że kąt nachylenia osi x' w stosunku do osi *x* będzie równy kątowi między osiami *t* i t'. Transformacja Lorentza zmienia więc nachylenie obu osi: czasowej i przestrzennej.

Mogliśmy ten rezultat zgadnąć na podstawie znanego już faktu, że prędkość światła jest stała (równa 1) w każdym układzie odniesienia. Zatem linia światła musi dzielić kąt między osiami t' i x' na połowę, jak to pokazaliśmy wcześniej. Oś x', jak również wszystkie linie do niej równoległe, reprezentują „obszary" równoczesności w nowym układzie. Rozszerzając ten rezultat na przypadek dwóch wymiarów przestrzennych, możemy powiedzieć, że płaszczyzny równoczesności w układach poruszających się będą nachylone w stosunku do płaszczyzny wyjściowej, a ich nachylenie będzie tym większe, im większa jest prędkość układu. W wypadku trójwymiarowym mamy do czynienia nie z dwuwymiarową, a z trójwymiarową powierzchnią równoczesności (dlatego też zwana jest ona czasem „hiperpowierzchnią"). Jest niezmiernie trudno wyobrazić sobie dwie trójwymiarowe hiperpowierzchnie, nachylone do siebie pod kątem w czwartym wymiarze, ale tak właśnie wygląda wynikająca z teorii względności relacja między „równoczesnościami" w dwóch układach.

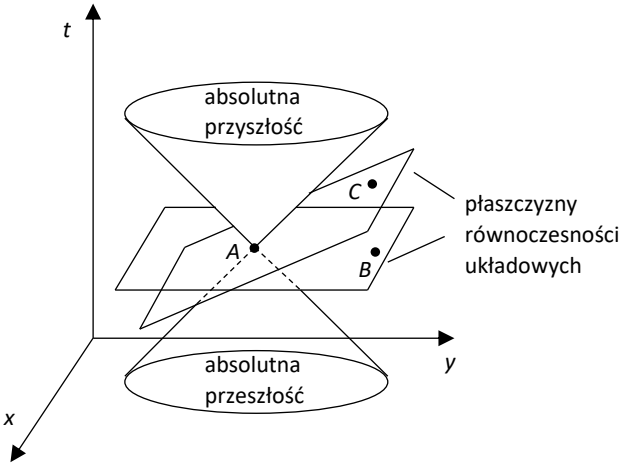


Rys. 5.8. Stożki świetlne w czasoprzestrzeni Minkowskiego

Wróćmy teraz do interwałów czasoprzestrzennych. Wybierzmy pewien punkt *A* w czasoprzestrzeni i narysujmy wybiegające z niego i zbiegające do niego linie światła (rys. 5.8). Dla przypadku dwóch wymiarów przestrzennych linie te utworzą dwa stożki, zwane stożkami świetlnymi. Rozważmy teraz obszary wewnątrz obu stożków. Łatwo zauważyć, że zdarzenia w nich zawarte są rozdzielone interwałem czasopodobnym od naszego wybranego zdarzenia *A*. Wynika to stąd, że linia prosta łącząca zdarzenie *A* z dowolnym zdarzeniem z tych obszarów będzie miała nachylenie względem osi czasu mniejsze niż linia światła (brzeg stożka), a zatem reprezentuje obiekt poruszający się wolniej niż światło. Można też dostrzec, że dla dowolnego zdarzenia wewnątrz stożka istnieje układ odniesienia, w którym to zdarzenie za-

chodzi dokładnie w tym samym miejscu co zdarzenie A. Jest to układ, którego oś t' łączy A z danym zdarzeniem. Zauważmy, że w tym układzie interwał czasoprzestrzenny redukuje się do czasu własnego, gdyż $\Delta s = 0$. Jest to ogólna prawidłowość: interwał czasoprzestrzenny dla zdarzeń zachodzących „na" danym ciele fizycznym jest po prostu równy czasowi własnemu tego ciała.

Zdarzenia, które znajdują się na powierzchni obu stożków, są oddzielone od zdarzenia A interwałem zerowym. Pozostała część czasoprzestrzeni, znajdująca się poza obydwoma stożkami, obejmuje wszystkie i tylko te zdarzenia, które są oddzielone od A interwałem przestrzennopodobnym. Połączenie tych zdarzeń ze zdarzeniem A sygnałem o prędkości nie większej niż prędkość światła jest niemożliwe. Jednocześnie możemy zauważyć, że dla dowolnego zdarzenia z obszaru poza stożkami istnieje układ odniesienia, w którym zachodzi ono równocześnie z A. Jest to np. układ, którego oś x' przechodzi dokładnie przez A i wybrane zdarzenie, a zatem, zgodnie z wcześniej omówionym geometrycznym obrazem transformacji Lorentza, oś czasowa t' takiego układu musi być nachylona do osi t pod takim samym kątem co kąt osi x' względem x.



Rys. 5.9. Podział czasoprzestrzeni na absolutną kauzalną przeszłość i przyszłość

Obszary rozdzielone stożkami świetlnymi mają interesujące i ważne interpretacje filozoficzne. Na wstępie przyjmijmy założenie, że żadne oddziaływanie przyczynowe nie może rozchodzić się z prędkością większą niż prędkość światła. Oddziaływania elektromagnetyczne na pewno spełniają ten warunek, a co do innych rodzajów nie znamy przypadków, żeby założenie to było złamane. Jedyna możliwość obejścia tej zasady, którą możemy nazwać zasadą lokalności relacji przyczynowej, pojawia się na gruncie mechaniki kwantowej – będziemy o tym mówić w kolejnych częściach książki. Obecnie jednak przyjmiemy, że jeśli dwa zdarzenia są powiązane relacją przyczynową, to musi istnieć możliwość ich połączenia sygnałem rozchodzącym się z prędkością co najwyżej równą prędkości światła. Wynika stąd, że wszystkie zdarzenia, które mogą wpływać na wybrane zdarzenie A, znajdują się bądź wewnątrz, bądź na powierzchni „dolnego" stożka świetlnego (rys. 5.9). Podobnie zdarzenia, na które może fizycznie wpłynąć zdarzenie A, są rozmieszczone w „górnym" stożku świetlnym. Z tego powodu dolny stożek nazywa się niekiedy „absolutną kauzalną przeszłością" zdarzenia A, a górny stożek jego „absolutną kauzalną przyszłością". Termin „absolutna" wskazuje,

że pojęcia te są niezależne od układu odniesienia, jako że stożki świetlne nie zmieniają się przy przejściu z jednego układu odniesienia do drugiego. Termin „kauzalna" oznacza z kolei, że interesuje nas tylko potencjalne oddziaływanie z wybranym zdarzeniem, a nie sam fakt czasowego poprzedzania lub następowania.

Jak się ma pojęcie absolutnej kauzalnej przeszłości/przyszłości do zwykłych pojęć przeszłości i przyszłości? Standardowe rozumienie np. przeszłości jest takie, że obejmuje ona wszystko, co wydarzyło się przed obecną chwilą. Na diagramie czasoprzestrzennym byłaby to zatem część czasoprzestrzeni o współrzędnej czasowej mniejszej niż współrzędna wybranego zdarzenia *A* (część poniżej płaszczyzny równoczesności przechodzącej przez *A*). Jednak w różnych układach odniesienia obszar ten będzie wyglądał różnie, gdyż sama płaszczyzna równoczesności ulega transformacji („nachyleniu"). Stożek świetlny obejmuje wszystkie te zdarzenia, które są wspólne wszystkim „układowym" przeszłościom. Można zatem alternatywnie scharakteryzować absolutną kauzalną przeszłość dla *A* jako zawierającą te zdarzenia, co do których wszyscy obserwatorzy zgodzą się, że zachodzą one wcześniej niż wyróżniony punkt *A*.

Czy zdarzenia zachodzące poza stożkami świetlnymi zbiegającymi w punkcie *A* nie mogą być po prostu traktowane jako równoczesne z *A*? Niestety to nie takie proste. Rozważmy dwa zdarzenia *B* i *C*, jak na rysunku 5.9. Choć oba są oddzielone przestrzennopodobnie od *A*, to jednak *C* znajduje się w absolutnej przyszłości od *B*. Jest niemożliwe, aby dwa zdarzenia równoczesne z trzecim nie były wzajemnie równoczesne. Zatem to nie równoczesność zachodzi między *B* i *A* oraz *C* i *A*. Relację tę można nazwać co najwyżej *quasi*-równoczesnością: jest to równoczesność zrelatywizowana do pewnego układu, przy czym dla par zdarzeń *B*, *A* i *C*, *A* są to dwa różne układy. *Quasi*-równoczesność nie posiada jednej ważnej własności równoczesności – nie jest mianowicie przechodnia.

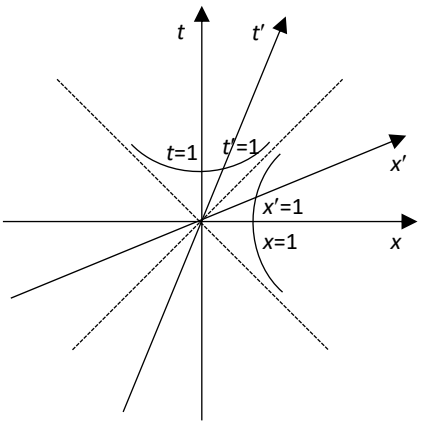
Istnieje ważny powód, dla którego nie powinniśmy lekko przyjmować istnienia oddziaływań (sygnałów) rozchodzących się z prędkością większą od prędkości światła. Występowanie takich sygnałów mogłoby otworzyć drogę do oddziaływania wstecznego w czasie, ze wszystkimi tego konsekwencjami (paradoks dziadka). Gdyby zdarzenie *C* z rys. 5.9 mogło oddziałać na przestrzennopodobnie oddzielone zdarzenie *A*, a z kolei zdarzenie *A* mogło wysłać sygnał do również przestrzennopodobnie z nim oddzielonego zdarzenia *B*, to *C* mogłoby oddziałać fizycznie na swoją absolutną przeszłość. Moglibyśmy wtedy skonstruować sytuację, w której *C* było wywołane przez *B*, a z drugiej strony *C* wysłałoby sygnał do *A* inicjujący wyeliminowanie zdarzenia *B*. Mielibyśmy zatem sprzeczność: zdarzenie *C* istnieje (bo z założenia jest wywołane przez wcześniejsze zdarzenie *B*), ale zarazem nie istnieje (bo *B* jest wyeliminowane przez ciąg przyczynowy zachodzący wstecz w czasie).

5.5. Efekty relatywistyczne na diagramach

Umiemy już przedstawić graficznie transformacje Lorentza, których rezultatem jest „nachylenie" zarówno osi czasowej, jak i przestrzennej. Pozostał jeszcze jeden ważny element geometrii czasoprzestrzeni, dotąd niewprowadzony. Chodzi mianowicie o to, jak w poruszającym się układzie odznaczyć na osiach odpowiednie jednostki długości dla czasu i prze-

strzeni. Pamiętajmy, że geometria czasoprzestrzeni nie jest geometrią Euklidesową, więc nie możemy się sugerować „zwykłą" odległością między punktami. W szczególności jeśli np. jeden centymetr na osi t reprezentuje interwał jednosekundowy, to na osi t' jedna sekunda na pewno nie będzie odpowiadała temu samemu centymetrowi. Podobną sytuację mamy nawet w klasycznej czasoprzestrzeni Galileusza. Jedna sekunda w poruszającym się układzie wyznaczona jest przez poziomą linię równoczesności wychodzącą z punktu o współrzędnej 1 na osi t, a zatem Euklidesowa długość odcinka jednosekundowego w układzie primowanym będzie większa niż w nieprimowanym. To oczywiście żaden efekt „relatywistyczny", tylko konsekwencja tego, że czasoprzestrzeń Galileusza nie ma struktury Euklidesowej.

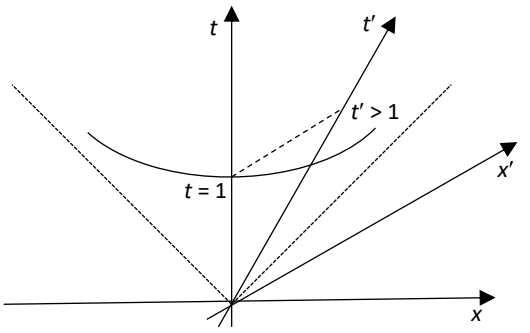
W czasoprzestrzeni Minkowskiego wyznaczenie jednostki czasu (i przestrzeni) jest trochę bardziej skomplikowane. Musimy w tym celu posłużyć się interwałem czasoprzestrzennym (5.4), który jest zawsze taki sam w każdym układzie odniesienia, a zatem daje nam możliwość porównania miar długości w różnych układach. Wybierzmy na osiach x i t wyjściowego układu punkty odpowiadające jednostkom $t = 1$ i $x = 1$. Zastanówmy się teraz, jak wyznaczyć wszystkie punkty, których interwał czasoprzestrzenny, liczony od początku układu (0, 0), jest taki sam jak interwał od wybranych jednostek $t = 1$ i $x = 1$. To nietrudne: np. interwał między punktem na osi x o współrzędnej 1 a początkiem układu wynosi –1, a zatem wszystkie punkty posiadające ten sam interwał muszą spełniać równanie $t^2 - x^2 = -1$. Podobnie w wypadku punktu $t = 1$, równanie będzie miało postać $t^2 - x^2 = 1$. Z geometrii wiemy, że równania te opisują hiperbole, zbiegające w nieskończoności do „przekątnych" (czyli linii światła), jak na rys. 5.10.



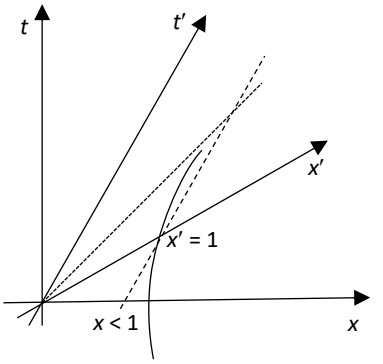
Rys. 5.10. Wyznaczenie jednostek w poruszającym się układzie

Weźmy teraz pod uwagę punkty przecięcia obu hiperbol stałego interwału odpowiednio z osiami t' i x'. Punkty te będą reprezentować poszukiwane jednostki czasu i przestrzeni w nowym układzie. Na przykład interwał punktu przecięcia hiperboli z osią t', obliczony w poruszającym się układzie, to po prostu kwadrat jego współrzędnej czasowej (współrzędna przestrzenna x' jest zerowa), ale ponieważ interwał ten w każdym układzie wynosi 1, współrzędna czasowa tego punktu jest również równa 1. Analogicznie w wypadku punktu na osi x', jego współrzędna przestrzenna musi równać się 1. Fakt, że wyznaczone w ten sposób punkty mają inną odległość Euklidesową od początku współrzędnych, nie ma większego zna-

czenia. Natomiast dzięki wprowadzonym jednostkom możemy łatwo zobrazować relatywistyczne efekty dylatacji czasu i skrócenia długości.



Rys. 5.11. Dylatacja czasu na diagramie

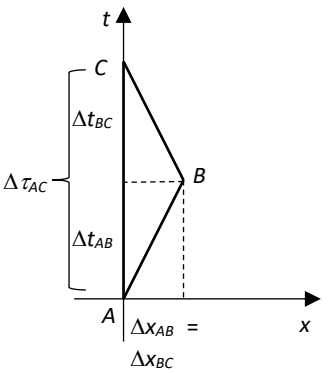


Rys. 5.12. Skrócenie długości na diagramie

Punkt na osi t o jednostkowej współrzędnej będzie miał współrzędną czasową, w poruszającym się układzie, wyznaczoną przez linię równoległą do osi x' (linię równoczesności w tym układzie), a zatem widać, że współrzędna ta będzie większa od 1 w primowanym układzie (rys. 5.11). Z kolei aby zilustrować skrócenie długości, należy rozważyć dwie równoległe linie, z których jedna przechodzi przez punkt zerowy (jest tożsama z osią czasu t'), a druga (przerywana) przez punkt o współrzędnej $x' = 1$. Linie te reprezentują ewolucję czasową początku i końca pręta o długości 1, spoczywającego w poruszającym się układzie. Długość pręta zmierzona w układzie „spoczywającym" będzie dana współrzędną punktu przecięcia drugiej linii z osią x (pamiętamy, że pomiar położenia obu końców pręta ma się odbyć jednocześnie w czasie $t = 0$), czyli znowu widzimy z rysunku, że będzie krótsza od jedności (rys. 5.12).

Przejdźmy na koniec do analizy słynnego paradoksu bliźniąt. Jak się za chwilę przekonamy, nie jest to żaden paradoks w sensie rozumowania prowadzącego do sprzeczności, a jedynie wyprowadzenie bardzo nieintuicyjnej konsekwencji z zasad relatywistyki. Załóżmy, że brat bliźniak został w domu, a jego siostra wyruszyła w podróż kosmiczną, z której

powróciła po latach. Jak się okazuje, czas trwania podróży mierzony czasem własnym siostry będzie krótszy od czasu własnego brata, czyli siostra stanie się młodsza od swojego brata bliźniaka. Narysujmy diagram podróży siostry z perspektywy układu odniesienia brata. Punkt *A* oznacza zdarzenie wylotu, punkt *B* osiągnięcie celu podróży kosmicznej i zawrócenie, a punkt *C* szczęśliwy powrót na Ziemię (rys. 5.13).



Rys. 5.13. Czasoprzestrzenny diagram paradoksu bliźniąt

Obliczmy interwały czasoprzestrzenne między punktami *AB*, *BC* i *AC* w układzie brata. Interwał między *A* i *C* będzie oczywiście równy interwałowi czasowemu $\Delta I_{AC} = \Delta\tau_{AC}$, gdyż odległość przestrzenna tych zdarzeń jest równa zeru. (Pamiętajmy, że interwał czasoprzestrzenny między dowolnymi zdarzeniami rozdzielonymi czasopodobnie jest zawsze równy czasowi własnemu, czyli mierzonemu w układzie, w którym zdarzenia te mają tę samą lokalizację przestrzenną.) Pozostałe dwa interwały są dane następująco:

$$\Delta I_{AB}^2 = \Delta t_{AB}^2 - \Delta x_{AB}^2 ,$$

$$\Delta I_{BC}^2 = \Delta t_{BC}^2 - \Delta x_{BC}^2$$

i oczywiście mamy

$$\Delta\tau_{AC} = \Delta t_{AB} + \Delta t_{BC} .$$

Skoro interwał czasoprzestrzenny dla danego obiektu jest po prostu równy jego czasowi własnemu, to możemy napisać:

$$\Delta\tau_{AB}^2 = \Delta t_{AB}^2 - \Delta x_{AB}^2 ,$$

$$\Delta\tau_{BC}^2 = \Delta t_{BC}^2 - \Delta x_{BC}^2 ,$$

gdzie $\Delta\tau_{AB}$ i $\Delta\tau_{BC}$ są czasami podróży od *A* do *B* i od *B* do *C* mierzonymi przez siostrę. Z powyższych równań dostajemy natychmiast nierówności:

$$\Delta\tau_{AB} < \Delta t_{AB},$$

$$\Delta\tau_{BC} < \Delta t_{BC},$$

a zatem

$$\Delta\tau_{AC} > \Delta\tau_{AB} + \Delta\tau_{BC} .$$

Wniosek: czas podróży mierzony ziemskimi zegarami będzie dłuższy od czasu zarejestrowanego w rakiecie kosmicznej. Siostra będzie więc po powrocie młodsza od swojego brata bliźniaka. Wydaje się to paradoksalne, ale tylko dlatego, że nie mieliśmy nigdy do czynienia z dalekimi podróżami kosmicznymi. Nie ma nic sprzecznego w założeniu, że czas może „płynąć" inaczej w różnych układach odniesienia. Nie powinniśmy więc upierać się przy naszych intuicjach nabytych w doświadczeniu z wolno poruszającymi się obiektami w niewielkiej od nas odległości (w porównaniu ze skalą kosmiczną). Natomiast warto może wyjaśnić źródło „niesymetryczności" pomiędzy bratem a siostrą. Co odróżnia sytuację podróżniczki od stanu brata-domatora, skoro, jak wiemy, ruch jest względny i równie dobrze możemy powiedzieć (z perspektywy lecącej rakiety), że to brat razem z całą Ziemią i układem słonecznym wybrał się w podróż kosmiczną?

Odpowiedź jest widoczna gołym okiem na diagramie czasoprzestrzennym: trajektoria siostry jest opisana linią łamaną, a trajektoria brata linią prostą, niezależnie od tego, w jakim (inercjalnym) układzie odniesienia opisujemy całą sytuację. W pewnym momencie siostra musi zmienić układy odniesienia – z poruszającego się ze stałą prędkością w kierunku od Ziemi, do poruszającego się w przeciwnym kierunku. Natomiast „leniwy" brat pozostaje cały czas w tym samym układzie. Z geometrii czasoprzestrzeni Minkowskiego wynika, że długość drogi po linii łamanej (mierzona oczywiście interwałem czasoprzestrzennym) jest krótsza od długości po linii prostej. Wydaje się to absurdalne, ale tylko dlatego, że nasze intuicje wyrosły na gruncie geometrii Euklidesa, w której obowiązuje tzw. nierówność trójkąta. Podkreślmy, że efekt postarzenia brata jest niezależny od wybranego układu odniesienia. Na przykład możemy opisać całą sytuację z miejsca pomiędzy Ziemią a rakietą, z którego perspektywy zarówno brat, jak i siostra się poruszają – najpierw w przeciwnych kierunkach, a potem siostra zaczyna doganiać brata. Inna możliwość to przetransformowanie się do układu odniesienia związanego z rakietą w pierwszej fazie podróży. Wtedy na początku siostra jest stacjonarna, a brat pędzi w przeciwnym kierunku, po czym siostra zaczyna lecieć w kierunku brata z dwa razy większą prędkością. Niezależnie od tego, jak opiszemy całą sytuację, pozostaje faktem, że brat nie zmienia układów odniesienia, podczas gdy siostra je zmienia.[3]

5.6. Niektóre filozoficzne konsekwencje szczególnej teorii względności

Szczególna teoria względności jest jedną z najintensywniej analizowanych przez filozofów teorii fizycznych (obok być może mechaniki kwantowej). Nie sposób w niniejszym podręcznikowym opracowaniu uwzględnić bogactwa wszystkich ważnych dla filozofa aspektów tej teorii. Należy jednak zwrócić uwagę na najczęściej przytaczane filozoficzne konsekwen-

[3] Oczywiście wszystko to zakłada, że mówimy cały czas o układach inercjalnych. Gdybyśmy dopuścili układy nieinercjalne, których prędkość może ulegać zmianie, to moglibyśmy wybrać układ odniesienia związany „na stałe" z rakietą siostry i wtedy kinematyczne zachowanie brata wyglądałoby dokładnie tak samo jak siostry w układzie Ziemi: brat poruszałby się w jedną stronę, a potem by zawrócił. Jednak układ rakiety nie jest fizycznie „równoważny" układowi związanemu z Ziemią, gdyż rakieta podlega przyspieszeniu podczas zawracania. Równoważność może być przywrócona, dopiero kiedy uwzględnimy grawitację – układy przyspieszające są równoważne układom „stacjonarnym" w polu grawitacyjnym. Jak się jednak okazuje, pole grawitacyjne ma dokładnie taki sam zaburzający wpływ na czas co przebywanie w nieinercjalnym układzie odniesienia – będziemy o tym mówić w rozdziale poświęconym ogólnej teorii względności.

cje relatywistyki. Przede wszystkim skoncentrujmy uwagę na problemie, który pojawił się już przy okazji analizy mechaniki newtonowskiej, a mianowicie na kwestii ontologicznego statusu czasu i przestrzeni. Nie podlega dyskusji, że teoria względności definitywnie obaliła stanowisko absolutyzmu w kwestii statusu czasu i przestrzeni traktowanych jako oddzielne byty. Jak się przekonaliśmy, czasowe i przestrzenne relacje między zdarzeniami nie są absolutne w sensie fizycznym, tj. są różne w różnych układach odniesienia. Dwa zdarzenia równoczesne w jednym układzie na ogół nie będą równoczesne w układzie poruszającym się względem tego pierwszego. Podobnie odległość przestrzenna między zdarzeniami zmienia się przy przejściu od układu do układu, nawet jeśli zdarzenia te są równoczesne w jednym z nich. Zatem i czas, i przestrzeń muszą zmieniać się przy zmianie układu odniesienia. Lokalizacje czasowe dwóch zdarzeń, które uznajemy za identyczne w jednym układzie, mogą nie być identyczne w innym (to samo dotyczy lokalizacji przestrzennych, co zresztą było prawdą nawet w fizyce newtonowskiej). Innymi słowy, każdy układ odniesienia posiada swój własny czas i swoją własną przestrzeń.

Wynika stąd, że w ontologicznej wersji sporu o naturę czasu i przestrzeni szala przechyla się na korzyść relacjonizmu. Czas nie może być substancją niezależną od istnienia przedmiotów materialnych, skoro w każdym układzie odniesienia, wyznaczonym przecież za pomocą materialnych urządzeń pomiarowych, przyjmuje inną postać. Podobnie rzecz się ma z przestrzenią. Czy zatem możemy ogłosić Leibniza zwycięzcą w jego sporze z Newtonem? Sprawa nie jest oczywista, gdyż problem może być przeformułowany tak, aby pytanie dotyczyło nie statusu czasu i przestrzeni, ale nierozerwalnej całości, jaką jest czasoprzestrzeń. Fundamentalna relacja czasoprzestrzenna, czyli relacja koincydencji (zachodzenie w tym samym czasie i tym samym miejscu) jest jawnie niezależna od układu odniesienia. Zatem można twierdzić, że struktura złożona z punktów czasoprzestrzennych nie jest ontycznie uzależniona od zajmujących te punkty obiektów materialnych. Szczególna teoria względności mówi nam jedynie, że „rozbicie" czasoprzestrzeni na niezależny czas i dopełniającą go przestrzeń nie jest zdeterminowane w jednoznaczny, obiektywny sposób. Natomiast status czasoprzestrzeni jako całości pozostaje nadal kwestią otwartą.

Substancjalizm czasoprzestrzenny pozostaje natomiast narażony na zarzut Leibniza ze złamania zasady tożsamości przedmiotów nieodróżnialnych (lub też zasady racji dostatecznej, jak to omówiliśmy w rozdziale 2.). Substancjalna interpretacja czasoprzestrzeni dopuszcza istnienie alternatywnych wszechświatów, różniących się jedynie tożsamością punktów i obszarów czasoprzestrzennych zajmowanych przez dane obiekty materialne, nie naruszając relacji czasoprzestrzennych między tymi obiektami, a zatem także nieodróżnialnych empirycznie. Z kolei argument Newtona z wiadrem, kontrujący Leibniza, nie ma tutaj bezpośredniego zastosowania, jako że istnienie sił pozornych w układach nieinercjalnych (np. obracających się) nie może być prosto wytłumaczone istnieniem absolutnej czasoprzestrzeni – jak pamiętamy, potrzebne jest w tym wypadku założenie istnienia absolutnej *przestrzeni*, względem której dokonuje się ruch obrotowy. Jednakże absolutna przestrzeń jest, jak już wspomnieliśmy, jawnie niezgodna z zasadami relatywistyki.

Zatem chociaż teoretycznie możliwa jest obrona absolutyzmu i substancjalizmu w stosunku do czasoprzestrzeni na gruncie szczególnej teorii względności, to jednak relacjonizm wydaje się być w znacznie lepszej sytuacji. Sytuacja ta zmieni się radykalnie przy przejściu do ogólnej teorii względności, gdzie sformułowane zostaną nowe, mocne argumenty za substancjalizmem. Co ciekawe, pojawi się także nowy argument za relacjonizmem, będący w pewnym sensie unowocześnioną i matematycznie zaawansowaną wersją argumentu Leib-

niza z przesunięcia. Spór o ontologiczny status czasoprzestrzeni jest – jak widać – daleki od rozstrzygnięcia, choć niewątpliwie nastąpił tutaj ogromny teoretyczny postęp związany z rozwojem teorii fizycznych.

Szczególna teoria względności ma również ważne konsekwencje dotyczące jednego z bardziej kontrowersyjnych problemów w filozofii czasu – problemu upływu czasu i związanego z nim podziału na przeszłość, teraźniejszość i przyszłość. Doświadczenie upływu czasu jest jednym z bardziej fundamentalnych doświadczeń ludzkiego życia. Mimo tej fundamentalności doprecyzowanie tego pojęcia sprawia ogromną trudność. Upływ czasu jest pewnego rodzaju zmianą, która zasadniczo różni się od zmian takich jak np. ruch w przestrzeni. Ruch zachodzi w czasie, natomiast upływ czasu sam jest zmianą czasu z przyszłego na teraźniejszy i przeszły. Niektórzy uważają, że do opisu zmian czasu potrzebny jest czas „drugiego rzędu", co prowadzi do regresu (czas drugiego rzędu wymaga czasu trzeciego rzędu do opisu jego zmian i tak dalej). Krytycy pojęcia upływu czasu zwracają również uwagę, że przy dosłownym rozumieniu tego pojęcia można postawić pytanie, jak szybko płynie czas. Jednakże prędkość upływu czasu nie ma dobrze określonego sensu fizycznego (w jakich jednostkach byłaby mierzona ta prędkość: sekundy na sekundę?).[4]

Szczególna teoria względności dostarcza nowego argumentu w sporze o ontologiczny status upływu czasu. Kluczowym założeniem zwolenników jego obiektywności jest teza o istnieniu niezależnego od obserwatora podziału na zdarzenia teraźniejsze, przeszłe i przyszłe. Teraźniejszość obejmuje wszystkie zdarzenia, które są równoczesne z aktem percepcji (świadomości). Jednakże wiemy, że relacja równoczesności zależy od stanu ruchu obserwatora. Dwoje obserwatorów poruszających się względem siebie uzna za równoczesne inne zdarzenia, a tym samym „teraźniejszości" dla tych obserwatorów będą różne. Istnieje nieskończenie wiele płaszczyzn równoczesności przechodzących przez dane zdarzenie punktowe. Która z tych płaszczyzn wyznacza „prawdziwą" teraźniejszość? Relatywizacja równoczesności do układu odniesienia mocno sugeruje, że obiektywna teraźniejszość nie ma sensu fizycznego.

Zwolennicy dynamicznego charakteru czasu mają oczywiście możliwość obrony swojego stanowiska. Po pierwsze, można utrzymywać, że chociaż empirycznie nie jesteśmy w stanie wybrać jednej spośród wielu płaszczyzn równoczesności, to jednak obiektywnie istnieje jedna taka wyróżniona płaszczyzna, definiująca zdarzenia, które dzieją się naprawdę „teraz". Taka strategia prowadzi do odseparowania empirii od ontologii, czyli do uznania, że są pewne obiektywne fakty, które nigdy nie będą mogły być przez nas poznane empirycznie. Niektórzy twierdzą, że możliwe jest wyróżnienie pewnego uprzywilejowanego układu odniesienia przez odwołanie się do faktów spoza samej teorii względności *sensu stricto* – np. do faktów kosmologicznych, dotyczących rozkładu materii w globalnej skali we wszechświecie. Wreszcie istnieje strategia oparta na zastąpieniu pojęcia globalnej teraźniejszości pojęciem teraźniejszości lokalnej, ograniczonej do danego ciała fizycznego i jego historii. Każde ciało, a zatem także i każdy obserwator, posiadałoby własną przeszłość, przyszłość i teraźniejszość. Rozwiązanie takie można by nazwać „solipsyzmem temporalnym". Jakkolwiek

[4] Zarzut „sekund na sekundę" bywa jednak odpierany – np. filozof fizyki Tim Maudlin podaje jako przykład sytuację wymiany walut. Możemy mówić o współczynniku zamiany o wymiarze „złotówka na dolar", ale również możemy rozważać współczynnik zamiany złotówek na złotówki, o stałej wartości równej jeden. Dodajmy, że w teorii względności istnieje pojęcie „czterowektora prędkości" (w skrócie „czteroprędkości"), którego długość w każdym układzie wynosi jeden (szczegóły będą podane w następnym paragrafie). Można argumentować, że pojęcie to reprezentuje prędkość upływu czasu.

debata ontologiczna na temat upływu czasu pozostaje nierozstrzygnięta, to jednak należy podkreślić, że teoria względności znacznie ogranicza dopuszczalne interpretacje tego zjawiska.

5.7. Relatywistyczna dynamika

Dla większości filozofów szczególna teoria względności „kończy się" na problemie czasu i przestrzeni. Natomiast dla fizyka nowa teoria czasoprzestrzeni to dopiero początek. Zasadniczym celem każdej teorii fizycznej jest przewidywanie przyszłego zachowania układów fizycznych. Cel ten realizowany jest przez sformułowanie praw dynamiki, które umożliwiają obliczenie przyszłego stanu układu na podstawie znajomości stanu teraźniejszego. Jak pamiętamy, fundamentem mechaniki klasycznej jest prawo dynamiki Newtona $F = ma$, które daje nam teoretyczną możliwość obliczenia trajektorii ciał fizycznych poddanych działaniu odpowiednich sił. Niestety, prawo Newtona pozostaje w oczywistym konflikcie z zasadami relatywistyki. Łatwo to pokazać: wystarczy wyobrazić sobie ciało, na które działa stała siła przez odpowiednio długi czas. W takim wypadku ciało to mogłoby osiągnąć dowolnie dużą prędkość, co jest niezgodne z zasadą nieprzekraczalności prędkości światła. Relatywistyczne prawo dynamiki musi ulec modyfikacji.

Najprostszym sposobem wprowadzenia do nowych zasad dynamiki jest po prostu napisanie odpowiednich równań i pokazanie, że unikają one powyższej trudności. Na przykład zastosowanie stałej siły zgodnie z nowym równaniem będzie prowadzić do tego, że tempo przyrostu prędkości ciała będzie spadać dokładnie w taki sposób, aby graniczną, lecz niemożliwą do osiągnięcia prędkością była prędkość światła. Większość popularnych wstępów do szczególnej teorii względności zadowala się wprowadzeniem i opisaniem nowych równań dynamiki. My jednak spróbujmy dłuższej, ale może ciekawszej drogi, zaczynając od wprowadzenia nowych pojęć matematycznych. Podstawowym pojęciem wykorzystanym w niniejszym paragrafie i w późniejszych rozdziałach będzie pojęcie *czterowektora*.

Dobrze znanym rodzajem wektorów, wykorzystywanym w fizyce, są wektory w trójwymiarowej przestrzeni Euklidesa. Jak wiadomo z elementarnej geometrii, wybierając dany układ współrzędnych, możemy przedstawić każdy wektor za pomocą trójki liczb, reprezentujących długości składowych tego wektora wzdłuż osi naszego układu. Nie jest zatem zaskoczeniem, że w czterowymiarowej czasoprzestrzeni Minkowskiego wektory będą miały cztery, a nie trzy składowe: jedną czasową i trzy współrzędne. Dla odróżnienia od „zwykłych" wektorów będziemy nazywać je czterowektorami. Ponieważ geometria czterowymiarowej czasoprzestrzeni różni się od geometrii Euklidesowej, czterowektory będą także odpowiednio różne od trójwektorów w przestrzeni Euklidesowej. Różnica ta ujawnia się w sposobie transformacji. Przechodząc z jednego układu współrzędnych do innego, musimy odpowiednio zmodyfikować liczby reprezentujące dany wektor. Natomiast zmianie nie powinna ulec długość danego wektora – długość liczona oczywiście odpowiednim „inwariantem" danej geometrii.

Rozważmy najprostszy przykład czterowektora – czterowektor przesunięcia, łączący początek układu współrzędnych z dowolnym punktem w czasoprzestrzeni. Do tej pory symbolizowaliśmy współrzędne punktów jako (t, x, y, z). Obecnie zastosujemy oznaczenia za pomocą indeksów od 0 do 3: (x^0, x^1, x^2, x^3). Współrzędna zerowa będzie współrzędną czasową, a współrzędne od 1 do 3 – przestrzenne. Czterowektor przesunięcia będziemy ogólnie symbolizować jako X^μ, gdzie $\mu = 0, 1, 2, 3$. Korzystając ze znanych wzorów na transformacje

Lorentza (5.3), możemy napisać, jak będą wyglądały składowe tego samego wektora w nowym układzie odniesienia, poruszającym się z prędkością v w kierunku osi x^1 (cały czas przy założeniu $c = 1$):

$$(X')^0 = \frac{X^0 - vX^1}{\sqrt{1 - v^2}},$$

$$(X')^1 = \frac{X^1 - vX^0}{\sqrt{1 - v^2}},$$

$$(X')^2 = X^2,$$

$$(X')^3 = X^3.$$

Przyjmiemy ogólnie, że każdy czterowektor transformuje swoje składowe w powyższy sposób.[5] Z kolei „długość" czterowektora będzie dana znaną formułą opartą na interwale czasoprzestrzennym:

$$(X^0)^2 - (X^1)^2 - (X^2)^2 - (X^3)^2.$$

Długość czterowektora nie ulega zmianie przy zmianie układu współrzędnych (jest inwariantem przyjętych transformacji czasoprzestrzeni).[6]

Wprowadźmy teraz nowy, ważny rodzaj czterowektora: tzw. czteroprędkość. Standardowa definicja prędkości jest dobrze znana: jest to wektor, którego składowe są dane przez pochodne po czasie składowych wektora przesunięcia: $v_x = \frac{dx}{dt}$, $v_y = \frac{dy}{dt}$, $v_z = \frac{dz}{dt}$. Analogicznie, składowe czterowektora prędkości będą dane przez zróżniczkowanie składowych czterowektora przesunięcia X^μ po czasie. Jednakże w szczególnej teorii względności istnieje wiele „rodzajów" czasu (skoro w każdym układzie współrzędnych czas jest inny). Aby czterowektor prędkości transformował się we właściwy sposób, przyjęty czas musi być niezmienniczy, a zatem mamy tylko jeden dopuszczalny wybór – musi to być czas własny τ, czyli czas mierzony w układzie, w którym dane ciało spoczywa. Definicja czteroprędkości będzie więc następująca:

$$U^\mu = \left(\frac{dX^0}{d\tau}, \frac{dX^1}{d\tau}, \frac{dX^2}{d\tau}, \frac{dX^3}{d\tau}\right).$$

Współrzędne U^1, U^2, U^3 (które możemy skrótowo oznaczyć jako U^i, pamiętając o konwencji, że łacińskie indeksy przebiegają liczby 1, 2 i 3) są „zwykłymi" składowymi prędkości, lecz obliczonymi przy pomocy czasu własnego. Aby obliczyć czteroprędkość danego ciała w wybranym układzie, należy zamienić czas własny na czas t mierzony w tym właśnie układzie. Formuła „zamiany" jest nam znana z wcześniejszych rozważań: $\tau = t\sqrt{1 - v^2}$.

[5] Powyższe równania opisują jedynie transformację czterowektorów ze względu na tzw. pchnięcie Lorentzowskie (*Lorentz boost*), czyli przejście do układu poruszającego się z pewną prędkością. Ogólnie do zasad transformacji należy dołączyć jeszcze Euklidesowe transformacje trójwymiarowej przestrzeni, takie jak translacje i obroty. Grupa wszystkich przekształceń zachowujących równania szczególnej teorii względności nazywa się grupą Poincarégo.

[6] Inną ważną wielkością zachowaną podczas przejścia między układami współrzędnych jest iloczyn skalarny czterowektorów. Niech A^μ i B^ν będą dwoma czterowektorami. Iloczyn $A^\mu B^\mu$ definiujemy jako $A^0B^0 - A^1B^1 - A^2B^2 - A^3B^3$. Rozpisując transformację składowych obu wektorów, jak w powyższej formule, możemy policzyć, że wynik iloczynu będzie taki sam w obu układach.

Wstawiając to wyrażenie do powyższej definicji, otrzymamy następującą formułę na składowe przestrzenne czterowektora prędkości:

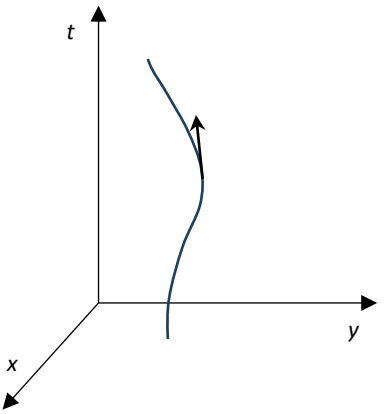
$$U^i = \left(\frac{dX^1}{dt} \frac{1}{\sqrt{1-v^2}}, \frac{dX^2}{dt} \frac{1}{\sqrt{1-v^2}}, \frac{dX^3}{dt} \frac{1}{\sqrt{1-v^2}} \right)$$

lub też po prostu (v^i jest *i*-tą składową zwykłej prędkości ciała względem danego układu):

$$U^i = \frac{v^i}{\sqrt{1-v^2}}.$$

Jak natomiast zinterpretować „zerową" składową czteroprędkości? Wstawiając w miejsce czasu własnego τ czas układowy *t*, i pamiętając, że X^0 to nic innego, jak właśnie czas układowy, otrzymamy proste równanie:

$$U^0 = \frac{1}{\sqrt{1-v^2}}.$$



Rys. 5.14. Czterowektor prędkości styczny do linii świata

Składowa zerowa czteroprędkości w danym układzie odniesienia jest w pewnym sensie współczynnikiem transformacji między czasem własnym a czasem mierzonym w tym układzie. Ogólnie czterowektor prędkości można przedstawić jako jednostkowy wektor styczny do „linii świata" danego ciała (trajektorii tego ciała w czterowymiarowej przestrzeni – por. rys. 5.14).[7] W układzie, w którym dane ciało spoczywa, trzy przestrzenne składowe czteroprędkości będą równe zeru, ale składowa czasowa U^0 pozostaje różna od zera (jest równa 1). Oczywiste, że nawet jeśli linia świata ciała jest w danym układzie prostopadła do płaszczyzny równoczesności, wektor styczny do tej linii nie przestanie istnieć, choć w tym układzie będzie miał tylko jedną składową – czasową.

[7] Proszę, sprawdźcie sami, że „długość" czterowektora prędkości obliczona przy pomocy interwału czasoprzestrzennego jest zawsze równa 1. Wskazówka: $v^2 = \left(\frac{dX^1}{dt}\right)^2 + \left(\frac{dX^2}{dt}\right)^2 + \left(\frac{dX^3}{dt}\right)^2$.

Możemy teraz przejść do fundamentalnego dla zasad dynamiki pojęcia pędu. Pamiętamy, że pęd łączy się z prędkością prostą formułą: $p = mv$. Ta sama zależność obowiązuje w teorii względności, zatem możemy zapisać:

$$P^{\mu} = mU^{\mu}.$$

Wynikają stąd następujące zależności między składowymi przestrzennymi czteropędu a „zwykłą" prędkością:

$$P^{i} = \frac{mv^{i}}{\sqrt{1 - v^{2}}},$$

gdzie $v^{i} = \frac{dX^{i}}{dt}$. Zauważmy, że wraz ze wzrostem prędkości v pęd rośnie w tempie dużo większym niż wynikałoby to z prostej proporcjonalności. W szczególności kiedy v zbliża się do prędkości światła (czyli 1), pęd rośnie do nieskończoności. Ma to fundamentalne znaczenie dla zasad relatywistycznej dynamiki. Pamiętamy, że drugą zasadę dynamiki Newtona można przedstawić za pomocą równania:

$$F = \frac{dp}{dt},$$

które oznacza, że tempo przyrostu pędu jest równe działającej sile. Stosując tę zasadę do pędu relatywistycznego, otrzymujemy konsekwencję, iż zastosowanie stałej siły nie spowoduje wzrostu prędkości do nieskończoności. Pęd co prawda będzie rósł w stałym tempie opisanym powyższym równaniem, ale skorelowana z nim prędkość v będzie w coraz wolniejszym tempie zbliżała się do prędkości granicznej $c = 1$, nigdy jednak jej nie osiągając (gdyż pęd w takiej sytuacji musiałby być nieskończony). Zatem wymagana modyfikacja zasad dynamiki dokonała się za pomocą modyfikacji pojęcia pędu.

Pozostała nam jeszcze kwestia interpretacji zerowej składowej czterowektora pędu. Jak łatwo zauważyć, wynosi ona:

$$P^{0} = \frac{m}{\sqrt{1 - v^{2}}}.$$

Jaki jest sens fizyczny tej formuły? Spróbujmy go odgadnąć, rozważając przybliżenie powyższej formuły dla małych prędkości. W tym celu skorzystamy z następującego matematycznego przybliżenia. Dla niewielkich wartości v ułamek z pierwiastkiem w mianowniku można z dobrą dokładnością przedstawić następująco:

$$\frac{1}{\sqrt{1 - v^{2}}} \approx 1 + \frac{v^{2}}{2}.$$

Stosując powyższe przybliżenie, otrzymamy następująca formułę:

$$P^{0} = m + \frac{mv^{2}}{2}.$$

Drugi składnik sumy po prawej stronie równania rozpoznajemy jako energię kinetyczną ciała poruszającego się z prędkością v. Co jednak z pierwszym składnikiem? Aby zachować spójność jednostek, przywróćmy może prędkość światła c, którą dla wygody pomijaliśmy jako równą jedności. W takim zapisie prawa strona równania będzie wyglądać następująco:

$$mc^2 + \frac{mv^2}{2}.$$

Jestem pewien, że każdy czytelnik rozpozna teraz pierwszy składnik nawet szybciej niż ten drugi, „klasyczny". Oczywiście jest to słynna Einsteinowska formuła na tzw. relatywistyczną energię spoczynkową, $E = mc^2$. Łącznie z drugim członem całość wyrażenia opisuje energię całkowitą. Zatem zerowa składowa czterowektora pędu okazała się niczym innym jak całkowitą energią ciała.[8]

Musimy jednak pamiętać, że dokonaliśmy tutaj dużego skoku myślowego. Poszukując interpretacji dla składowej P^0, przeprowadziliśmy najpierw graniczne przejście dla niewielkich prędkości. Zwykle takie przejście odtwarza nam dobrze znane równania fizyki klasycznej. Na przykład pamiętamy, że przypadek graniczny dla niewielkich prędkości we wzorach na transformację Lorentza daje nam klasyczną transformację Galileusza. Podobnie efekty relatywistyczne (dylatacja czasu, skrócenie długości) ulegają zanikowi przy rozważaniu prędkości dużo mniejszych od c. Natomiast w powyższym wypadku ewidentnie wyrażenie na P^0 nie przechodzi w dobrze znany wzór na energię kinetyczną. Niezależnie od tego, jak mała jest prędkość v, wyrażenie określające P^0 będzie zawierało człon, który nie pojawia się w fizyce klasycznej (dodajmy: człon o ogromnej wartości liczbowej, proporcjonalnej do kwadratu prędkości światła). Interpretując wyrażenie mc^2 jako pewną nową, nieznaną wcześniej „formę" energii, przyznajemy tym samym, że rozbieżności między fizyką klasyczną a relatywistyczną pojawiają się także w znanym nam z doświadczenia codziennego obszarze niewielkich prędkości i małych odległości. Dlaczego zatem nie zauważyliśmy wcześniej tej różnicy?

Dokładna odpowiedź wymagałaby zagłębienia się w samo pojęcie energii i jego związek z danymi obserwacyjnymi, na co nie mamy tutaj miejsca (choć dotknęliśmy już tego problemu we wcześniejszych rozdziałach przy okazji omawiania zasady zachowania energii i równoważności energii mechanicznej i cieplnej). Możemy jednak zauważyć, że energia spoczynkowa ciała sama przez się nie jest obserwowalna. Jej obecność może ujawnić się dopiero wtedy, gdy mamy do czynienia z procesem, w którym występuje różnica między całkowitą masą na początku i na końcu procesu. W tym wypadku ubytek masy Δm musi być zrekompensowany energią kinetyczną o wartości Δmc^2. Takich procesów w świecie makroskopowym nie obserwujemy, natomiast świat subatomowy dostarcza nam mnóstwa przykładów na zamianę energii spoczynkowej w energię kinetyczną. Najbardziej spektakularnym przykładem są reakcje jądrowe, w tym rozszczepienie ciężkich jąder atomowych (lub też fuzja jąder lekkich). Masa produktów rozszczepienia jest odrobinę mniejsza od masy jądra wyjściowego, a różnica ta znajduje ujście w energii kinetycznej – bądź to gwałtownie w formie eksplozji nuklearnej, bądź też „łagodnie" w formie wydzielonego ciepła w reaktorze jądrowym.

W poniższej ramce dyskutuję szczególny przypadek fotonów, który wymaga specjalnego potraktowania. Jeśli nie jesteście specjalnie zainteresowani tym przypadkiem, na tym kończy się wprowadzenie do szczególnej teorii względności dla dociekliwych filozofów. Następne dwa paragrafy są przeznaczone dla osób bardziej oswojonych z wyższą matematyką – dotyczą one pewnych pojęciowych i formalnych kwestii związanych z relacją między klasyczną teorią elektromagnetyzmu a teorią względności.

[8] Z tego powodu czterowektor pędu nazywa się również czterowektorem energii-pędu.

Nietrudno zauważyć, że wiele równań szczególnej teorii względności „załamuje się" w przypadku obiektów poruszających się z prędkością światła. (Według obecnego stanu wiedzy jedynymi takimi obiektami są fotony, czyli kwanty światła. Przez pewien czas sądzono, że również neutrina należą do tej samej kategorii, ale przypuszczenie to zostało obalone doświadczalnie.) Okazuje się, że do fotonów nie można zastosować np. pojęcia czteroprędkości. Wynika to stąd, że czas własny fotonów jest równy zeru. Pamiętamy, że interwał czasoprzestrzenny między dwoma punktami połączonymi linią światła jest zerowy, a czas własny to nic innego jak interwał czasoprzestrzenny. Zatem różniczkowanie po zmiennej τ, dla której $d\tau = 0$, nie ma matematycznego sensu. Można ten problem wyrazić jeszcze inaczej: nie istnieje jednostkowy wektor styczny do linii świata promienia świetlnego, z prostego powodu: wszystkie wektory styczne do niej mają długość (liczoną Lorentzowsko) równą zeru. Jednakże czterowektor prędkości musi mieć długość jednostkową, więc fotony nie mogą mieć zdefiniowanej czteroprędkości.

Natomiast czteropęd nie musi być jednostkowy – jego długość jest, jak łatwo policzyć, równa m. Zatem matematycznie jest możliwe przypisanie fotonom czteropędu stycznego do linii światła, z tym że musimy wtedy założyć, że $m = 0$. Stąd mamy prosty wiosek, że fotony są bezmasowe. Czterowektor pędu dla fotonu ma postać (E, p^1, p^2, p^3), gdzie $E = p$ (lub też, wracając do zwykłych jednostek, $E = cp$; p to oczywiście całkowity pęd). Energia pojedynczego fotonu jest dana wzorem $h\nu$, gdzie ν jest częstotliwością, a zatem pęd bezmasowego fotonu to $\frac{h\nu}{c}$. Dodajmy jeszcze na koniec, że w ogólnym przypadku relatywistyczna relacja między energią a pędem wynika z faktu, że kwadrat długości czteropędu jest równy m^2, czyli $E^2 - p^2 = m^2$, a zatem $E = \sqrt{m^2 + p^2}$ lub też w dowolnych jednostkach prędkości światła c, $E = \sqrt{m^2c^4 + c^2p^2}$. Dla $m = 0$ wzór ten przechodzi w równość $E = cp$, obowiązującą dla fotonów.

5.8.* Relatywistyczna teoria elektromagnetyzmu: siła Lorentza

W niniejszym paragrafie zajmiemy się dokładniej kwestią sformułowania relatywistycznie poprawnej (tj. Lorentzowsko niezmienniczej) wersji teorii elektromagnetyzmu Maxwella. W pewnym sensie zadanie to zostało już wykonane, jako że teoria Maxwella nawet w swojej pierwotnej postaci jest zasadniczo niezmiennicza względem transformacji Lorentza (to było zresztą powodem jej problemów w zestawieniu z zasadą względności Galileusza). Dobrze jest jednak nadać teorii Maxwella postać jawnie relatywistycznie niezmienniczą przez wyprowadzenie jej z fundamentalnych zasad relatywistyki. Poza tym nie wszystkie prawa elektromagnetyzmu są rzeczywiście zgodne z zasadami szczególnej teorii względności. Na przykład wzór na siłę Lorentza, mimo swojej nazwy, nie jest Lorentzowsko niezmienniczy i musi być zastąpiony relatywistycznie poprawną formułą.

Nasze podejście będzie dość abstrakcyjne – wyprowadzimy prawa elektromagnetyzmu z zasady najmniejszego działania, którą wprowadziliśmy w rozdziale 2. przy okazji omówienia alternatywnych podejść do mechaniki newtonowskiej. Abstrakcyjność tego podejścia polega na tym, że będziemy musieli zgadnąć właściwą formę działania, tak aby wyprowadzone z niego prawa zgadzały się z empirycznie potwierdzonymi prawami teorii Maxwella. Jest natomiast kluczowe, aby przyjęte działanie (a w konsekwencji także lagrangian) miało jawnie

relatywistycznie niezmienniczą postać, co zagwarantuje nam także niezmienniczość całej teorii. Zacznijmy od zagadnienia, jak ogólnie opisać w formalizmie Lagrange'owskim zachowanie ciała materialnego obdarzonego ładunkiem w polu elektromagnetycznym. Jak pamiętamy z rozdziału 2. (paragraf 2.8), faktyczną trajektorię ciała materialnego poddanego działaniu odpowiednich sił możemy wyznaczyć, obliczając wielkość zwaną działaniem dla każdej możliwej trajektorii od punktu wyjścia do punktu końcowego i wybierając tę, dla której działanie przyjmuje wartość ekstremalną (zwykle minimalną). Działanie przedstawiamy jako całkę od punktu początkowego do końcowego z wielkości zwanej lagrangianem, która w standardowych zastosowaniach mechaniki klasycznej jest różnicą między energią kinetyczną a energią potencjalną ciała. Z kolei warunek minimalizacji działania jest matematycznie równoważny pewnemu równaniu zwanemu równaniem Eulera-Lagrange'a (2.2). Ostatecznie trajektorię cząstki wyznaczamy, rozwiązując to równanie dla danego lagrangianu.

Obecnie naszym celem będzie nie tyle obliczenie konkretnej trajektorii cząstki, co znalezienie odpowiedniego równania ruchu, które musi być spełnione przez cząstkę w polu elektromagnetycznym – czyli relatywistycznego wariantu wzoru na siłę Lorentza. Jednakże na razie nie dysponujemy odpowiednim lagrangianem – będziemy chcieli skonstruować go przez przyjęcie odpowiedniej formy działania. Załóżmy, że działanie opisujące zachowanie cząstki jest sumą dwóch składników: działania swobodnego, opisującego cząstkę, na którą nie działa żadna siła, oraz działania opisującego oddziaływanie cząstki z polem. Oba składniki działania muszą być relatywistycznie niezmiennicze – jak już wiemy z poprzedniego paragrafu, znaczy to, że muszą one być dane w postaci skalarów, niezależnych od przyjętego układu współrzędnych. Działanie swobodne najlepiej przedstawić za pomocą czasu własnego τ, który, jak mówiliśmy, jest niezmienniczy względem transformacji Lorentza (pamiętamy z poprzednich paragrafów konwencję $c = 1$):

$$\mathcal{A}_{free} = -m\int_a^b d\tau = -m\int_a^b \sqrt{1-v^2}dt.$$

Z kolei część działania opisująca oddziaływanie musi zawierać informację na temat pola. Przyjmijmy na próbę, że pole elektromagnetyczne opisane jest, w zgodzie z formalizmem teorii względności, za pomocą pewnego czterowektora $A_\mu(t,x)$, który jest oczywiście funkcją położenia w czasoprzestrzeni. Iloczyn skalarny takiego wektora z wektorem infinitezymalnego przesunięcia dx^μ jest skalarem, a zatem pozostaje niezmienniczy względem wyboru układu odniesienia. Spróbujmy więc przyjąć następującą postać działania opisującego oddziaływanie cząstki z polem, stosując konwencję sumacyjną Einsteina, nakazującą sumowanie po powtarzających się indeksach (e jest ładunkiem cząstki):

$$\mathcal{A}_{int} = e\int_a^b A_\mu(t,x)dx^\mu.$$

Spodziewam się, że wielu czytelników w tym momencie da upust swojej frustracji – skąd wiadomo, że tak powinno wyglądać odpowiednie działanie? Na jakiej podstawie możemy przyjąć taką a nie inną formę matematycznych wyrażeń opisujących dane zjawisko? Szczególnie dla filozofa przyjmowanie daleko idących tez bez odpowiedniego uzasadnienia – empirycznego czy też matematycznego – może wydawać się podejrzane. Filozofowie przyzwyczajeni są do kwestionowania nawet najbardziej oczywistych twierdzeń w rodzaju „stół,

który widzę przed sobą, naprawdę istnieje". Moja odpowiedź na te jak najbardziej uzasadnione obiekcje jest taka, że nie powinniśmy traktować powyższych konstrukcji jako logicznych dedukcji z niebudzących żadnych wątpliwości aksjomatów, a raczej jako zastosowanie metody prób i błędów w celu odtworzenia pewnych znanych już rezultatów. W języku angielskim taką metodę nazywa się nieco bardziej uczenie *reverse engineering* – to dobieranie pewnych pojęć i zasad tak, aby uzyskać oczekiwany rezultat. Naszym celem jest uzyskanie znanych praw elektromagnetyzmu – lub taż praw zbliżonych do znanej formy praw elektromagnetyzmu – korzystając z fundamentalnych zasad mechaniki (w tym z zasady najmniejszego działania) oraz zachowując niezmienniczość relatywistyczną. Jeśli za pierwszym podejściem nam się nie uda, będziemy próbować innego podejścia. Oczywiście tak naprawdę wiadomo, że musi się udać – w końcu nie wymyśliłem tej metody sam, ale nauczyłem się jej od innych, którzy ją dobrze sprawdzili.

Z nadzieją, że uwagi te rozwiały choć część wątpliwości, powrócę teraz do zadania wyprowadzenia prawa rządzącego oddziaływaniem cząstek z polem. Sprawa jest raczej prosta. W wyrażeniu na działanie $\mathcal{A}_{int}$ dla interakcji dokonamy formalnej zamiany zmiennej całkowania dx^μ na zmienną czasową dt przez podstawienie $dx^\mu = \frac{dx^\mu}{dt}dt$. W rezultacie otrzymamy postać lagrangianu opisującego oddziaływanie z polem A_μ (jest to wyrażenie całkowane po zmiennej czasowej w celu obliczenia działania):

$$\mathcal{L}_{int} = eA_\mu \frac{dx^\mu}{dt} = eA_\mu \dot{x}^\mu.$$

Pamiętając, że $x^0 = t$ oraz $\frac{dt}{dt} = 1$, możemy napisać wyrażenie na całkowity lagrangian (przypominam, że indeksy łacińskie przebiegają trzy współrzędne przestrzenne 1, 2, 3):

$$\mathcal{L} = -m\sqrt{1+\dot{x}^2} + eA_0(t,x) + e\dot{x}^n A_n(t,x). \tag{5.5}$$

Mam nadzieję, że drobne przekształcenia formalne nie przyprawiły Czytelnika o zawrót głowy. Teraz musimy sobie przypomnieć „straszne" równanie Eulera-Lagrange'a (2.2):

$$\frac{d}{dt}\left(\frac{\partial \mathcal{L}}{\partial \dot{x}^p}\right) = \frac{\partial \mathcal{L}}{\partial x^p}.$$

Jedyne, co nam pozostało, to zróżniczkować wyrażenie (5.5) na lagrangian względem zmiennych $\dot{x}^p$ i x^p, pamiętając, że A_μ zależy od x^p, ale nie od $\dot{x}^p$. Liczę na to, że bardziej wprawni matematycznie Czytelnicy będą w stanie zrobić to sami, ale jeśli macie z tym kłopoty, przyjmijcie rezultat „na wiarę":

$$\frac{\partial \mathcal{L}}{\partial \dot{x}^p} = \frac{m\dot{x}^p}{\sqrt{1-\dot{x}^2}} + eA_p,$$

$$\frac{\partial \mathcal{L}}{\partial x^p} = e\frac{\partial A_0}{\partial x^p} + e\dot{x}^n \frac{\partial A_n}{\partial x^p}.$$

Wiem, robi się trochę gorąco. Jeszcze nieco cierpliwości i zobaczymy światełko w tunelu. Na razie wstawmy powyższe do równania Eulera-Lagrange'a:

$$\frac{d}{dt}\left(\frac{m\dot{x}^p}{\sqrt{1-\dot{x}^2}}\right) + e\frac{dA_p}{dt} = e\frac{\partial A_0}{\partial x^p} + e\dot{x}^n \frac{\partial A_n}{\partial x^p}.$$

Jeszcze musimy się zająć pochodną $\frac{dA_p}{dt}$. Jest to pochodna całkowita, a więc uwzględnia ona zarówno jawną zależność $A_p(t,x)$ od t, jak i zależność „ukrytą", ze względu na zależność współrzędnych przestrzennych x od t. Pamiętajmy, że metoda minimalizacji działania oparta jest na poszukiwaniu trajektorii ciała, a trajektoria to nic innego jak uzależnienie współrzędnych przestrzennych od czasu w formie zależności funkcyjnej $x^p(t)$. Z rachunku różniczkowego wiemy, że pochodna całkowita względem danej zmiennej x jest równa sumie pochodnych cząstkowych po każdej zmiennej razy jej pochodna po zmiennej x. Oszczędzę Czytelnikowi kilku żmudnych przekształceń i napiszę wynik końcowy (zachęcam jednak odważniejszych do wypróbowania swoich umiejętności matematycznych – naprawdę sprawia to dużo przyjemności, kiedy uzyska się właściwy wynik).

$$\frac{d}{dt}\left(\frac{m\dot{x}^p}{\sqrt{1-\dot{x}^2}}\right) = e\left(\frac{\partial A_0}{\partial x^p} - \frac{\partial A_p}{\partial t}\right) + e\dot{x}^n\left(\frac{\partial A_n}{\partial x^p} - \frac{\partial A_p}{\partial x^n}\right). \tag{5.6}$$

Teraz możemy się przyjrzeć nieco dokładniej uzyskanemu rezultatowi. Po lewej stronie równości powinniśmy rozpoznać wyrażenie znane z poprzedniego paragrafu – jest to relatywistyczna formuła na tempo zmiany pędu (ściślej, jego składowej p), czyli w skrócie relatywistyczna siła działająca na cząstkę. Prawa strona formuły zawiera dwa człony – jeden niezależny od prędkości, a drugi zależny. Spróbujmy odświeżyć nieco pamięć i napisać standardową formułę na całkowitą siłę Lorentza (elektryczną i magnetyczną):

$$\boldsymbol{F} = e\boldsymbol{E} + e\boldsymbol{v} \times \boldsymbol{B}.$$

Porównując oba wyrażenia, możemy dojść do wniosku, że pierwszy nawias po prawej stronie równości (5.6) reprezentuje natężenie pola elektrycznego:

$$E_p = \frac{\partial A_0}{\partial x^p} - \frac{\partial A_p}{\partial t}. \tag{5.7}$$

Z drugim wyrażeniem jest pewien kłopot, gdyż mamy tam dwa indeksy – jeden indeks p zgodny z indeksem po lewej stronie, a drugi „uwikłany" z indeksem składowych prędkości $\dot{x}^n$. Proponuję, żeby Czytelnik rozpisał sobie całą sumę, używając „zwykłych" oznaczeń na współrzędne przestrzenne x, y i z zamiast x^1, x^2, i x^3. Na przykład niech x^p będzie współrzędną z. Dostaniemy wtedy następujące wyrażenie

$$\dot{x}\left(\frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x}\right) + \dot{y}\left(\frac{\partial A_y}{\partial z} - \frac{\partial A_z}{\partial y}\right).$$

Wyrażenie to jest podobne do znanej nam formuły (4.10) na składowe iloczynu wektorowego dwóch wektorów. Iloczyn $\boldsymbol{A} \times \boldsymbol{B}$ jest wektorem, którego składowa, np. w kierunku z, jest dana odpowiednią kombinacją składowych wektorów $\boldsymbol{A}$ i $\boldsymbol{B}$:

$$A_xB_y - B_xA_y$$

i podobnie dla pozostałych składowych. Aby uzyskać analogiczną formę dla powyższego wyrażenia, musimy wprowadzić wektor $\boldsymbol{B}$ o następujących składowych x i y:

$$B_x = \frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}, \tag{5.8}$$

$$B_y = \frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x},$$

a także dla składowej w kierunku z (pomijam odpowiednie wyprowadzenie, ale działa ono podobnie do powyższego):

$$B_z = \frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}. \tag{5.9}$$

Zatem drugie wyrażenie we wzorze na relatywistyczną siłę Lorentza można przedstawić jako składową p iloczynu wektorowego prędkości i nowo wprowadzonego wektora $\boldsymbol{B}$: $\boldsymbol{v} \times \boldsymbol{B}$. Jest to dokładnie tak samo jak w oryginalnym, nierelatywistycznym wyrażeniu na siłę pochodzącą od pola magnetycznego. Czyli wektor $\boldsymbol{B}$ należy zinterpretować jako natężenie pola magnetycznego. Jego związek z czterowektorem A_μ wynika z powyższych równań (5.8) i (5.9): $\boldsymbol{B}$ jest iloczynem wektorowym wektora nabla ∇ i trójwektora A_n, zdefiniowanego przez trzy współrzędne przestrzenne czterowektora A_μ: $\boldsymbol{B} = \nabla \times \boldsymbol{A}$. Inaczej mówiąc, $\boldsymbol{B}$ jest rotacją wektorowego potencjału magnetycznego $\boldsymbol{A}$**.**

Tyle wysiłku, aby wyprowadzić wzór, który praktycznie znaliśmy już na początku spotkania z teorią elektromagnetyzmu! Jednakże pewien postęp się dokonał. Po pierwsze, upewniliśmy się, że wzór na siłę Lorentza, po niezbędnych poprawkach, jest w pełni zgodny z zasadami relatywistyki. Po drugie, uzyskaliśmy szczegóły matematycznej relacji pomiędzy wektorami pola elektrycznego i magnetycznego a czterowektorem A_μ. Ten czterowektor to nic innego jak czteropotencjał elektromagnetyczny, którego zerowa składowa reprezentuje potencjał elektryczny (skalar), a pozostałe trzy składowe wektorowy potencjał magnetyczny. Potencjał elektryczny A_0 (ze znakiem minus) przechodzi w dobrze znany potencjał elektrostatyczny V (którego gradient jest równy polu $\boldsymbol{E}$), kiedy potencjał magnetyczny jest stały. Chociaż wzór na siłę Lorentza można zapisać wyłącznie przy pomocy potencjału A_μ, bez potrzeby wprowadzania wektorów $\boldsymbol{E}$ i $\boldsymbol{B}$, to jednak należy pamiętać, że potencjał ten dany jest tylko z dokładnością do transformacji cechowania, a więc nie odzwierciedla w pełni realności fizycznej. Można udowodnić (nie będziemy tego robić), że dodanie do składowych czterowektora A_μ pochodnych cząstkowych z dowolnego skalara S o postaci $\frac{\partial S}{\partial x^\mu}$ (czyli innymi słowy gradientu S) nie zmienia wartości działania $\mathcal{L}$, a zatem nie ma żadnego wpływu na fizyczną sytuację.

Pozostała nam jeszcze jedna kwestia – mianowicie wprowadzenie tensora pola elektromagnetycznego, który unifikuje pozornie odrębne pola elektryczne i magnetyczne. Pojawia się on w naturalny sposób, kiedy przepiszemy wzór na siłę Lorentza (5.6) w pełnej postaci czterowektorowej. Znów zachęcam Czytelnika do wyprowadzenia tego wzoru z wersji dla trójwektorów (przy czym równanie dla współrzędnej zerowej jest po prostu założone, jako że wynika ono z równań dla pozostałych składowych):

$$m\frac{d^2x^\mu}{d\tau^2} = e\left(\frac{\partial A_\nu}{\partial x^\mu} - \frac{\partial A_\mu}{\partial x^\nu}\right)u^\nu,$$

gdzie u^ν jest czteroprędkością cząstki (por. poprzedni paragraf). Wyrażenie w nawiasie z dwoma indeksami μ i ν intepretujemy jako nowy tensor – tensor pola elektromagnetycznego:

$$F_{\mu\nu} = \frac{\partial A_\nu}{\partial x^\mu} - \frac{\partial A_\mu}{\partial x^\nu}. \tag{5.10}$$

Zauważmy, że składowe tego tensora będą odpowiednimi składowymi pola elektrycznego i magnetycznego. Na przykład weźmy składową F_{10}. Biorąc pod uwagę wzór (5.7) na zależność między polem E_p a potencjałem A_μ, otrzymujemy $F_{10} = E_x$. Z kolei np. $F_{21} = -B_z$ (na podstawie równań (5.8) i (5.9), definiujących składowe wektora pola magnetycznego). Zwróćmy ponadto uwagę, że tensor $F_{\mu\nu}$ jest antysymetryczny: $F_{\mu\nu} = -F_{\nu\mu}$, z czego od razu wynika, że $F_{\mu\mu} = 0$ (elementy „diagonalne" tensora są równe 0). Kompletna postać tensora $F_{\mu\nu}$ w formie macierzowej podana jest poniżej.

$$\begin{pmatrix} 0 & -E_x & -E_y & -E_z \\ E_x & 0 & B_z & -B_y \\ E_y & -B_z & 0 & B_x \\ E_z & B_y & -B_x & 0 \end{pmatrix}. \tag{5.11}$$

Na koniec pokażmy, że tensor $F_{\mu\nu}$ nie wyznacza jednoznacznie czterowektora potencjału elektromagnetycznego A_μ – mamy tu do czynienia ze swobodą cechowania co do wyboru odpowiedniego potencjału. Niech będzie dany potencjał A_μ. Rozważmy nowy, „przeskalowany" potencjał:

$$A'_\mu = A_\mu + \frac{\partial S}{\partial x^\mu},$$

gdzie S jest dowolnym skalarem. Obliczając formę „nowego" tensora elektromagnetycznego, dostaniemy:

$$F'_{\mu\nu} = \frac{\partial A'_\nu}{\partial x^\mu} - \frac{\partial A'_\mu}{\partial x^\nu} = \frac{\partial A_\nu}{\partial x^\mu} - \frac{\partial A_\mu}{\partial x^\nu} - \frac{\partial S}{\partial x^\mu \partial x^\nu} + \frac{\partial S}{\partial x^\nu \partial x^\mu},$$

a ponieważ dwukrotne różniczkowanie po zmiennych x^μ i x^ν jest przemienne, otrzymamy z powrotem „stary" tensor: $F'_{\mu\nu} = F_{\mu\nu}$. Zatem dodanie gradientu dowolnego skalara do A_μ nie ma żadnych konsekwencji fizycznych – jest to transformacja cechowania.

5.9.* Wyprowadzenie równań Maxwella

Filozofowie „racjonaliści" mają skłonność do wyprowadzania nawet najbardziej oczywistych tez i twierdzeń z podstawowych zasad, co do których nie można mieć absolutnie żadnych wątpliwości. Taką metodę postępowania obrał np. Kartezjusz, który usiłował uzasadnić całą naszą empiryczną wiedzę (i dzięki temu odeprzeć narzucające mu się wątpliwości sceptyczne) na podstawie pierwszej zasady *Cogito ergo sum*. Podobne zadanie postawił sobie Spinoza i wielu innych. Chociaż fizycy na ogół nie podzielają radykalnego sceptycyzmu co do wiarogodności doświadczenia zmysłowego, to jednak w niektórych wypadkach podążają w podobnym kierunku, wyprowadzając aksjomatycznie pewne prawa przyrody z pierwszych zasad zamiast z uogólnień doświadczalnych. Jest to zadanie interesujące z wielu względów – możliwość takich dedukcji daje nam potencjalnie wgląd w najbardziej uniwersalne prawidłowości świata fizycznego o charakterze niemal metafizycznym. Żaden filozof zainteresowany ogólną refleksją nad rzeczywistością nie powinien przejść obojętnie wobec takich prób.

Obecnie pokażemy, w jaki sposób można z pewnych pierwszych zasad wyprowadzić cztery podstawowe prawa Maxwella, opisujące zjawiska elektromagnetyczne. W istocie rzeczy w poprzednim paragrafie dokonaliśmy już zasadniczego kroku w tym kierunku. Na podstawie zasady najmniejszego działania oraz zasady niezmienniczości względem transformacji Lorentza uzasadniliśmy regułę oddziaływania ładunków elektrycznych z polem elektromagnetycznym. Co więcej, wyprowadziliśmy również wzajemne zależności między polem elektrycznym i magnetycznym a potencjałem elektromagnetycznym. Okazuje się, że na podstawie samych tych zależności można matematycznie udowodnić dwa z czterech równań Maxwella – te, w których nie występują ładunki elektryczne ani prąd (równania te nazywamy jednorodnymi). Z kolei dwa pozostałe równania – niejednorodne – mogą być wyprowadzone przy założeniu dodatkowej przesłanki w formie zasady zachowania ładunku elektrycznego. Pokażmy najpierw szczegóły tych dedukcji, a następnie zastanowimy się nad ich metafizyczną interpretacją.

Zacznijmy od syntetycznego ujęcia relacji między wektorami $\boldsymbol{E}$ i $\boldsymbol{B}$ a czterowektorem potencjału A_μ, które wcześniej rozpisaliśmy na współrzędne (formuły 5.7 – 5.9). Łatwo sprawdzić (proszę, zróbcie to samodzielnie), że relacje te dadzą się wyrazić pojedynczymi wzorami obejmującymi całe wektory, a nie tylko ich składowe:

$$\begin{aligned} \boldsymbol{E} &= -\left(\frac{\partial \boldsymbol{A}}{\partial t} - \boldsymbol{\nabla} A_0\right), \\ \boldsymbol{B} &= \boldsymbol{\nabla} \times \boldsymbol{A}\,. \end{aligned} \tag{5.12}$$

Skorzystamy teraz z dwóch równości matematycznych, które można prosto udowodnić, korzystając z definicji iloczynu skalarnego i wektorowego (pierwsza z nich wynika natychmiast z faktu, że iloczyn wektorowy dwóch wektorów jest prostopadły do każdego z nich):

$$\begin{aligned} \boldsymbol{\nabla} \cdot (\boldsymbol{\nabla} \times \boldsymbol{A}) &= 0, \\ \boldsymbol{\nabla} \times (\boldsymbol{\nabla} S) &= 0. \end{aligned} \tag{5.13}$$

Stosując pierwszą równość w (5.13) do wyrażenia na wektor pola magnetycznego $\boldsymbol{B}$ z (5.12), dostajemy natychmiast jedno z równań Maxwella (charakteryzujące pole magnetyczne jako bezźródłowe):

$$\boldsymbol{\nabla} \cdot \boldsymbol{B} = 0\,.$$

Obliczając rotację z pola elektrycznego wyrażonego w pierwszym z równań (5.12), uzyskamy:

$$\boldsymbol{\nabla} \times \boldsymbol{E} = -\left(\frac{\partial \boldsymbol{\nabla} \times \boldsymbol{A}}{\partial t} - \boldsymbol{\nabla} \times (\boldsymbol{\nabla} A_0)\right).$$

Rotacja z gradientu A_0 znika, zgodnie z drugim równaniem (5.13), a zatem po prawej stronie równości zostaje tylko jeden człon, który rozpoznajemy jako pochodną z wektora natężenia pola magnetycznego. W ten sposób otrzymaliśmy znane nam prawo Faradaya:

$$\boldsymbol{\nabla} \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}.$$

Wynik ten jest jednak zaskakujący. Pamiętamy, że celem, jaki nam przyświecał, było wyprowadzenie praw Maxwella z pewnych fundamentalnych zasad opisujących świat fizycz-

nych. Powyższe rozumowanie pokazuje, że dwa z czterech równań Maxwella można wyprowadzić z równości (5.12) oraz pewnych twierdzeń matematycznych. Czym jednak są owe równości? Przypomnijmy, że wprowadziliśmy je przy okazji analizy wzoru na relatywistyczną siłę Lorentza jako pewnego rodzaju skróty dla odpowiednich wyrażeń, które się tam pojawiały. Zatem wydaje się, że formuły (5.12) to nic innego, jak definicje wektorów $\boldsymbol{E}$ i $\boldsymbol{B}$. Po prostu umawiamy się, aby odpowiednio nazwać pewne matematyczne kombinacje czterowektora A_μ. Lecz jeśli tak, to jednorodne prawa Maxwella okazują się matematycznymi konsekwencjami konwencji językowych, a więc mają charakter zdań analitycznych *a priori*, prawdziwych na mocy znaczenia, które właściwie nic nie mówią na temat świata. Mamy znów do czynienia z problemem, który pojawił się przy okazji analizy praw mechaniki Newtona – fundamentalne prawa przyrody przybierają status definicji.

Pamiętamy jednak, że przy bliższym zbadaniu sprawa analityczności praw dynamiki newtonowskiej okazała się bardziej złożona. Pokazaliśmy, że elementy konwencjonalne łączą się w nich z niewątpliwymi elementami empirycznymi, które mogłyby okazać się fałszywe. Można argumentować, że z podobną sytuacją mamy do czynienia obecnie. Kluczem do rozwiązania zagadki jest fakt, że wiele definicji zawiera element faktualny w postaci twierdzenia o istnieniu odpowiednio definiowanego obiektu (czyli, innymi słowy, są to definicje twórcze: por. wpis w ramce na s. 38). Jeśli na przykład zdecydujemy się na zdefiniowanie pewnej liczby rzeczywistej a przy pomocy równania $a^2 = -1$, to warunek istnienia nie będzie spełniony, a zatem definicja jest niepoprawna. W wypadku formuł (5.12) nie chodzi nam o istnienie matematyczne, ale raczej fizyczne. Dokładniej, możemy argumentować, że równania te zakładają już fizyczne istnienie odpowiedniej wielkości A_μ – potencjału elektromagnetycznego. Ujmując to jeszcze inaczej: równania (5.12) implikują, że dla eksperymentalnie zdefiniowanych wielkości $\boldsymbol{E}$ i $\boldsymbol{B}$ istnieje odpowiedni fizyczny obiekt A_μ, który spełni odpowiednie matematyczne równania. Tak jednak być nie musi! Na przykład jest kwestią empiryczną, że wektor pola magnetycznego da się zawsze przedstawić w postaci rotacji innego wektora. Gdyby istniały ładunki magnetyczne, byłaby możliwość istnienia pola magnetycznego o niezerowej dywergencji, które nie mogłoby być przedstawione jako rotacja. Analogiczny argument można zastosować do równania charakteryzującego wektor pola elektrycznego.

Co zatem z naszym racjonalistycznym ideałem wiedzy? Jakie dodatkowe nieanalityczne (tj. syntetyczne) zasady musimy dołączyć do praw matematyki i analitycznych definicji pojęć, aby uzyskać potrzebne nam założenie o istnieniu odpowiedniego potencjału elektromagnetycznego? Sprawa nie jest prosta. Niektórzy fizycy uważają, że takim brakującym aksjomatem jest przytaczana już wcześniej zasada najmniejszego działania. Można ją ująć w formie reguły metodologicznej: dla każdego typu zjawiska fizycznego należy poszukiwać odpowiednio sformułowanego działania w formie całki po trajektoriach z pewnego wyrażenia zwanego lagrangianem, tak aby minimalizowanie tego działania odtworzyło nam faktyczne zachowanie rozważanego układu fizycznego. Taka reguła jednak nie mówi nam, jak konkretnie ma wyglądać owo działanie ani stowarzyszony z nim lagrangian. Pamiętamy z poprzedniego paragrafu, że kluczowym elementem naszego rozumowania, prowadzącego do relatywistycznego wzoru na siłę Lorentza (a w konsekwencji do jednorodnych równań Maxwella), było przyjęcie odpowiedniej formy działania $\mathcal{A}_{int}$ opisującego interakcję. Ten wybór nie był jednakże uzasadniony żadną ogólną przesłanką o charakterze aksjomatycznym. Był raczej podyktowany tym, aby „się zgadzało”, tj. aby uzyskany wzór na siłę Lorentza miał postać

przypominającą znaną wcześniej formułę. Poprzez „odwrotną inżynierię" zgadliśmy odpowiednie wyrażenie na lagrangian, co pozwoliło nam w konsekwencji na wprowadzenie pól elektrycznych i magnetycznych jako matematycznych funkcji potencjału. Istotnym elementem tej strategii było mimo wszystko uprzednie założenie przynajmniej przybliżonej prawdziwości zasad elektromagnetyzmu znanych z doświadczenia.

Przykro mi, że rozczaruję zwolenników filozoficznego racjonalizmu, ale wydaje się, że bez uprzednio zdobytej empirycznej wiedzy na temat oddziaływania ładunków elektrycznych z polem elektrycznym i magnetycznym nie bylibyśmy w stanie dokonać odpowiedniego wyprowadzenia praw elektromagnetyzmu z „pierwszych zasad". Mimo to uzyskaliśmy ciekawy rezultat. Połączenie zasad relatywistyki ze znaną wcześniej prawidłowością dotyczącą działania pola elektromagnetycznego na ładunki elektryczne pozwoliło nam na wyprowadzenie dwóch praw opisujących zachowanie samych tych pól. Oddziaływanie pól na ładunki i pewne relacje między polami, w tym fakt nieistnienia ładunków magnetycznych, okazały się powiązane. Jest to zaskakujące, choć może nie tak bardzo jak ewentualna możliwość wyprowadzenia całej teorii elektromagnetyzmu z pierwszych zasad. Pomyślcie tylko: zjawisko indukcji elektromagnetycznej, opisane w prawie Faradaya, może być teoretycznie wyprowadzone z relatywistycznego uogólnienia wzoru na siłę Lorentza.[9] Czy to znaczy, że eksperymentalne badania Faradaya i innych były bezwartościowe? Pozostawiam to pytanie Czytelnikowi.

Zostały nam jeszcze dwa równania Maxwella, zawierające odniesienie do ładunków i prądu (równania niejednorodne). Co z nimi? Tutaj okazuje się, że same relacje między polami $\boldsymbol{E}$, $\boldsymbol{B}$ a czterowektorem potencjału są niewystarczające. Musimy znowu wrócić do podstaw i zacząć od zasady najmniejszego działania. Tym razem jednak zastosujemy tę zasadę nie do opisu ruchu cząstki w polu elektromagnetycznym, ale do wyprowadzenia, jak ewoluuje samo pole. Zmieniamy zatem perspektywę – pole elektromagnetyczne staje się samo podmiotem zmian fizycznych, a nie prostym dostarczycielem sił działających na obiekty fizyczne. Jak jednak opisać ewolucję pola w języku trajektorii? W tym celu trzeba zmodyfikować pojęcie działania. W wypadku pojedynczej cząstki działanie opisane zostało jako całka, której parametrem całkowania był czas, a granice całkowania wyznaczone były przez stan początkowy i stan końcowy. W wypadku pola sytuacja przedstawia się inaczej. Pole przypisuje każdemu punktowi w czasoprzestrzeni pewną wielkość. „Trajektorią" pola będzie po prostu przypisanie wartości pola każdemu punktowi w pewnym zamkniętym obszarze czasoprzestrzeni. Problemem, który chcemy rozwiązać, jest pytanie, jaka będzie wartość pola wewnątrz danego obszaru przy założeniu, że znane jest pole na brzegu tego obszaru. W tym celu musimy policzyć odpowiednie działanie i zminimalizować jego wartość. Obliczmy zatem całkę z lagrangianu w całej objętości rozważanego obszaru, czyli zmiennymi całkowania będą wszystkie cztery współrzędne: jedna czasowa i trzy przestrzenne:

$$\mathcal{A} = \int \mathcal{L}\, dtdxdydz.$$

[9] Ten wynik nie będzie tak zaskakujący, jeśli przypomnimy sobie dyskusję z paragrafu 4.6, gdzie pokazaliśmy, że efekt wzbudzania siły elektromotorycznej w obwodzie da się wytłumaczyć samą siłą Lorentza przy przejściu do innego układu odniesienia. Podobnie fakt, iż pole magnetyczne może zmienić jedynie kierunek ruchu naładowanej cząstki, a nie jej prędkość, zawiera w sobie ukrytą informację o nieistnieniu ładunków magnetycznych.

Lagrangian $\mathcal{L}$ będzie w ogólności funkcją zarówno wartości pola φ, jak i pochodnych pola po każdej współrzędnej: $\frac{\partial\varphi}{\partial t}, \frac{\partial\varphi}{\partial x}, \frac{\partial\varphi}{\partial y}, \frac{\partial\varphi}{\partial z}$. Równanie Eulera-Lagrange'a w wypadku pól musi zatem uwzględnić te wszystkie parametry lagrangianu:

$$\sum_{\mu=0}^{3} \frac{\partial}{\partial x^{\mu}} \frac{\partial \mathcal{L}}{\partial \frac{\partial\varphi}{\partial x^{\mu}}} = \frac{\partial \mathcal{L}}{\partial \varphi}.$$

Wygląda to może trochę przerażająco, ale struktura tego równania jest w zasadzie zbliżona do struktury oryginalnego równania Eulera-Lagrange'a dla pojedynczej cząstki. Zapiszmy to oryginalne równanie dla współrzędnej *x* cząstki (analogiczne równania mamy dla pozostałych współrzędnych *y* i *z*):

$$\frac{d}{dt}\left(\frac{\partial \mathcal{L}}{\partial \dot{x}}\right) = \frac{\partial \mathcal{L}}{\partial x}.$$

Dla pola odpowiednikiem współrzędnej *x* (zwanej również stopniem swobody cząstki) jest wartość pola φ. Zatem wydawać by się mogło, że po lewej stronie równania powinniśmy mieć pochodną cząstkową z lagrangianu po $\dot{\varphi}$ (czyli pochodnej φ po czasie: $\frac{d\varphi}{dt}$). Ponieważ jednak obecnie pole φ zależy od czterech zmiennych x^{μ}, a nie od jednego czasu, jak to było w wypadku położenia *x* cząstki, nic dziwnego, że musimy obliczyć pochodne cząstkowe z funkcji $\mathcal{L}$ po wszystkich czterech szybkościach zmian pola względem każdej ze współrzędnych oraz ich szybkość zmian po tej współrzędnej i zsumować rezultaty.

Pozostaje nam jedynie znaleźć odpowiedni lagrangian, aby wyprowadzić brakujące równania Maxwella. Bagatela! Niestety, znów wyciągamy królika z kapelusza. Musimy rozważyć dwa przypadki: sytuację, kiedy nie ma ładunków i prądów, i przypadek z ładunkami i prądami. W pierwszej sytuacji przyjęty lagrangian będzie miał następującą postać:

$$\mathcal{L} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} \tag{5.14}$$

Cóż mogę powiedzieć? Zapomnijmy o jakichś próbach apriorycznego uzasadnienia tej formuły i spróbujmy po prostu ją przeanalizować. Jak się domyślacie, funkcja $\mathcal{L}$ została tak dobrana, aby wynik był odpowiedni. Przypomnijmy jeszcze raz konwencję sumacyjną Einsteina: skrótowy zapis $F_{\mu\nu}F^{\mu\nu}$ oznacza, że mamy tutaj sumę odpowiedniej liczby składników tensora pola elektromagnetycznego: $F_{00}F^{00} + F_{01}F^{01} + \cdots$. Tensor $F^{\mu\nu}$ ze wskaźnikami górnymi („kontrawariantny") różni się tylko tym od tensora ze wskaźnikami dolnym („kowariantnego"), że wszystkie składowe z wskaźnikami zerowymi mają przeciwne znaki ($F^{0\nu} = -F_{0\nu}$ i $F^{\nu 0} = -F_{\nu 0}$).[10] Możecie policzyć, biorąc macierzową postać tensorów $F_{\mu\nu}$ i $F^{\mu\nu}$ (5.11), że w rezultacie tej długiej sumy dostaniemy całkiem prosty wynik:

$$\mathcal{L} = \frac{1}{2}(E^2 - B^2).$$

Jest to wyrażenie niezmiennicze względem transformacji Lorentza (jako skalar). Jednakże obecnie najważniejsze staje się zidentyfikowanie zmiennej pola φ lagrangianu oraz jej po-

[10] Dokładną definicję tensorów kontra- i kowariantnych znajdziecie w rozdziale poświęconym ogólnej teorii względności.

chodnych cząstkowych, tak abyśmy mogli zastosować uogólnione równanie Eulera-Lagrange'a. Co jest polem w naszym wypadku? Wiadomo, że ani pole elektryczne, ani magnetyczne nie są odpowiednie, jako że nie są one niezmiennicze względem transformacji Lorentza. Musimy zatem znów wykorzystać czterowektor potencjału elektromagnetycznego A_μ. Ponieważ tensor $F_{\mu\nu}$ został wprowadzony właśnie jako funkcja potencjału (5.10), bez trudności możemy przedstawić powyższy lagrangian (5.14) w zależności od A_μ:

$$\mathcal{L} = -\frac{1}{4}\left(\frac{\partial A_\nu}{\partial x^\mu} - \frac{\partial A_\mu}{\partial x^\nu}\right)\left(\frac{\partial A^\nu}{\partial x^\mu} - \frac{\partial A^\mu}{\partial x^\nu}\right).$$

Zauważmy, że $\mathcal{L}$ nie zależy od samego potencjału A_μ a tylko od jego pochodnych cząstkowych. Oznacza to, że w równaniu Eulera-Lagrange'a prawa strona będzie równa zeru. Lewą stronę natomiast trzeba wyliczyć, różniczkując powyższe wyrażenie po odpowiednich pochodnych. Oszczędzę Czytelnikowi żmudnych obliczeń (choć nie są one wcale tak bardzo skomplikowane) i napiszę efekt końcowy w syntetycznej postaci równania tensorowego:

$$\frac{\partial F^{\mu\nu}}{\partial x^\nu} = 0.$$

Trochę trudno uwierzyć, ale to równanie zawiera w sobie dwa niejednorodne równania Maxwella przy założeniu, że nie ma ładunków i prądu elektrycznego. Na przykład dla $\mu = 0$ równanie powyższe przechodzi w następującą równość:

$$\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} = 0,$$

co łatwo rozpoznajemy jako dywergencję wektora $\boldsymbol{E}$ równą zeru. Z kolei dla $\mu = 1, 2, 3$ powyższe równanie da nam odpowiednie składowe wektorowego równania wyrażającego prawo Ampère'a-Maxwella dla przypadku, kiedy nie ma prądu (gęstość prądu $\boldsymbol{j}$ wynosi zero).

Odsapnijmy chwilkę przed ostatnią prostą. Będziemy teraz chcieli wyprowadzić ogólne niejednorodne równania Maxwella w sytuacji, kiedy mamy zarówno niezerowe ładunki elektryczne, jak i niezerowy prąd. Dodatkowym założeniem, potrzebnym do wyprowadzenia tych równań, będzie zasada zachowania ładunku elektrycznego, która jest jednym z fundamentalnych praw przyrody. Jak dotąd nie zaobserwowano żadnego zjawiska łamiącego tę zasadę. Jak jednak wyrazić ją matematycznie? Przede wszystkim musimy doprecyzować, co dokładnie oznacza zachowanie ładunku. Jedna z możliwych interpretacji („globalna") byłaby taka, że całkowity ładunek elektryczny we wszechświecie nie ulega zmianie. Jednak taka zasada nie wyklucza sytuacji, w której ładunek elektryczny w pewnej objętości (np. na Ziemi) nagle by zniknął, i natychmiast pojawiłby się w innej objętości (np. na Księżycu). Sumaryczny ładunek byłby zachowany, ale jednak w laboratorium na Ziemi zaobserwowalibyśmy niewytłumaczalne zniknięcie ładunku.[11] Takie zjawisko złamałoby zasadę zachowania ładunku w wersji lokalnej, którą można wyrazić w twierdzeniu, że wszelka zmiana ładunku w danej

[11] Można się zastanowić, czy takie zjawisko nie łamie zasady zachowania ładunku w wersji globalnej. Pamiętamy bowiem, że w szczególnej teorii względności wybór płaszczyzny równoczesności jest sprawą do pewnego stopnia umowną. W jednym układzie odniesienia zdarzenie „zniknięcia" ładunku na Ziemi i jego „pojawienia" się na Księżycu są równoczesne, a więc globalnie ładunek nie ulega zmianie. Rozważmy jednak inny układ odniesienia, w którym pojawienie się ładunku na Księżycu zachodzi później niż jego zniknięcie na Ziemi. W takim układzie przez pewną chwilę będziemy mieli brakujący ładunek – już go nie ma na Ziemi, a jeszcze się nie pojawił na Księżycu.

objętości musi być „skompensowana" odpowiednim przepływem prądu przez granice tej objętości.

Lokalną wersję zasady zachowania ładunku przedstawimy matematycznie w postaci równania, które zrównuje szybkość zmiany gęstości ładunku w danym punkcie (czy też infinitezymalnym obszarze) z „wypływem" prądu z tego punktu (obszaru). Jak pamiętamy z paragrafu (4.9), wypływ danego pola jest matematycznie wyrażany jego dywergencją. Zatem równanie, które czasami nazywane jest równaniem ciągłości, będzie następujące:

$$\frac{\partial \rho}{\partial t} = -\nabla \cdot \boldsymbol{j}.$$

Może być ono przepisane w formie czterowektorowej, typowej dla teorii względności. Wprowadźmy czterowektor gęstości prądu J^{μ} w następujący sposób:

$$J^{\mu} = (\rho, j_x, j_y, j_z).$$

Zatem składowa zerowa czterowektora J^{μ} to gęstość ładunku, a pozostałe składowe są tożsame ze składowymi gęstości prądu. Równanie ciągłości można teraz napisać w eleganckiej formie (pamiętajmy, że $t = x^0$):

$$\frac{\partial J^{\mu}}{\partial x^{\mu}} = 0.$$

Ostatnim krokiem naszej żmudnej drogi do wyprowadzenia niejednorodnych równań Maxwella będzie oczywiście zgadnięcie formy nowego lagrangianu. Do wyrażenia (5.14) określającego lagrangian dla sytuacji bezźródłowej dodamy człon reprezentujący oddziaływanie ze źródłami pól elektrycznych i magnetycznych, czyli zawierający gęstość ładunku i prądu. Będzie on miał postać iloczynu skalarnego dwóch czterowektorów (co zapewnia niezmienniczość Lorentzowską):

$$\mathcal{L}_{źr} = J^{\mu} A_{\mu}.$$

Zauważmy, że dodanie tego członu do lagrangianu (5.14) nie zmienia wartości pochodnych względem tempa zmian czterowektora A_{μ}(czyli względem $\frac{\partial A_{\mu}}{\partial x^{\nu}}$), gdyż w nowym członie nie występują pochodne A_{μ}. Zatem lewa strona równania Eulera-Lagrange'a nie ulegnie zmianie. Natomiast prawa strona równania już nie będzie wynosić zero, bo lagrangian zależy od samego A_{μ}. Jego pochodna cząstkowa względem A_{μ} to oczywiście J^{μ}. Czyli ostatecznie nasze równanie Eulera-Lagrange'a daje następujący wynik:

$$\frac{\partial F^{\mu\nu}}{\partial x^{\nu}} = J^{\mu}.$$

Jest to pełna postać niejednorodnych równań Maxwella (proszę, rozpiszcie sobie to równanie na składowe, żeby się przekonać, że dla $\mu = 0$ daje nam ono różniczkowe prawo Gaussa, a dla pozostałych wartości prawo Ampère'a-Maxwella w pełnej postaci z prądem).

Chwileczkę, ale czy nie mówiliśmy wcześniej, że do wyprowadzenia pełnych równań Maxwella będzie nam potrzebna zasada zachowania ładunku (równanie ciągłości)? W którym momencie wykorzystaliśmy to równanie? Okazuje się, że równanie ciągłości pozwala argumentować, że działanie wyznaczone nowym lagrangianem jest niezmiennicze względem transformacji cechowania, czyli nie zmienia się przy zamianie czterowektora potencjału A_{μ}

na „równie dobry" potencjał $A_\mu + \frac{\partial S}{\partial x^\mu}$. Nie będziemy tego dowodzić – zostawmy tę kwestię bardziej sprawnym matematycznie czytelnikom do samodzielnej weryfikacji.

Podsumujmy wyniki uzyskane w tym paragrafie w celu uwypuklenia pewnych ogólnych prawidłowości. Okazuje się, że dwa typy równań Maxwella – jednorodne i niejednorodne – różnią się znacznie co do ich umocowania w pewnych ogólnych zasadach. Równania jednorodne (to wyrażające brak ładunków magnetycznych oraz prawo Faradaya) dają się wyprowadzić z lagrangianu opisującego oddziaływanie naładowanych cząstek z polem magnetycznym, przy założeniu istnienia odpowiednich zależności między polami elektrycznym i magnetycznym a czterowektorem potencjału. Natomiast równania niejednorodne (prawo Gaussa i prawo Ampère'a-Maxwella) wymagają zastosowania formalizmu Lagrange'owskiego do opisu ewolucji samego pola. Odpowiedni lagrangian dla sytuacji braku źródeł zapisujemy w postaci funkcji czterowektora potencjału, podczas gdy w przypadku ogólnym dodajemy do tego lagrangianu człon zawierający informację o gęstości ładunku i prądu (czterowektor gęstości prądu). Zarówno w wypadku równań jednorodnych, jak i niejednorodnych formułujemy odpowiednie lagrangiany w wersji niezmienniczej względem transformacji Lorentza oraz względem transformacji cechowania. Niezmienniczość względem transformacji cechowania w wypadku równań jednorodnych jest gwarantowana matematyczną postacią lagrangianu, podczas gdy w wypadku równań niejednorodnych wymaga ona przyjęcia założenia o lokalnym zachowaniu ładunku elektrycznego (równanie ciągłości).

Pytania i problemy

1. Jaki wpływ na hipotezę eteru miały doświadczenia Fizeau i Michelsona-Morley'a?

2. Omów szczegółowo strukturę doświadczenia Michelsona-Morley'a. Przedstaw możliwe wyjaśnienia jego negatywnego rezultatu. Dlaczego zostały one odrzucone na korzyść nowej teorii czasu i przestrzeni?

3. W jaki sposób można zdefiniować równoczesność dla zdarzeń odległych przestrzennie? Podaj dwie wersje takiej definicji. Pokaż, że definicja sygnałowa prowadzi do wniosku, że równoczesność zdarzeń zależy od przyjętego układu odniesienia.

4. Porównaj transformacje Lorentza z transformacjami Galileusza. Przy jakim założeniu te pierwsze mogą być zredukowane do drugich?

5. Pokaż, w jaki sposób transformacje Lorentza uzasadniają istnienie dwóch efektów relatywistycznych: dylatacji czasu i skrócenia Lorentza. Co to jest czas własny danego obiektu?

6. Porównaj relatywistyczne skrócenie długości wynikające z transformacji Lorentza z hipotezą Lorentza-FitzGeralda mającą wyjaśnić negatywny rezultat doświadczenia Michelsona-Morley'a. Dlaczego te dwa efekty są różne? Czy zgodnie z teorią względności przedmioty w ruchu ulegają fizycznemu „spłaszczeniu"?

7. Czy w relatywistyce jest prawdą, że jeśli ciało porusza się z pewną prędkością względem jednego układu, a ten układ porusza się z inną prędkością względem nas, to prędkość ciała względem nas będzie sumą (w ogólności wektorową) tych dwóch prędkości? Jak wygląda ta sytuacja, jeśli poruszające się ciało jest fotonem?

8. Wyjaśnij pojęcie interwału czasoprzestrzennego oraz podział na interwały czasopodobne, przestrzennopodobne i zerowe. Dla każdego z tych pojęć podaj jego dwie definicje – jedną opartą na możliwości połączenia zdarzeń sygnałem świetlnym, a drugą na istnieniu odpowiednich układów odniesienia, w których zachodzi równoczesność lub kolokacja.

9. Omów pojęcie stożka świetlnego oraz oparte na nim pojęcia absolutnej (kauzalnej) przyszłości i przeszłości. Jak mają się one do pojęć relatywnej (układowej) przeszłości, teraźniejszości i przyszłości?

10. Co to jest *quasi*-równoczesność? Czy dwa zdarzenia *quasi*-równoczesne z trzecim zdarzeniem są zawsze *quasi*-równoczesne ze sobą?

11. Niech A będzie częścią wspólną wszystkich obszarów czasoprzestrzeni reprezentujących relatywną (układową) przyszłość pewnego zdarzenia. Jakim obszarem będzie A? Rozważ analogiczne pytanie dla przypadku relatywnych przeszłości.

12. Narysuj na diagramie czasoprzestrzennym oś czasową i przestrzenną dla układu poruszającego się względem danego układu w kierunku rosnących wartości współrzędnej x oraz w kierunku malejących wartości x. Jak wyznaczyć na tych osiach punkty odpowiadające jednostkowej wartości współrzędnej czasowej i przestrzennej?

13. Przedstaw rozumowanie prowadzące do „paradoksu" bliźniąt, korzystając z warunku niezmienniczości interwału czasoprzestrzennego oraz z pojęcia czasu własnego. Co fizycznie odróżnia bliźniaka astronautę od bliźniaka pozostającego na Ziemi?

14. Dwóch braci bliźniaków wyleciało z Ziemi w przeciwnych kierunkach w rakietach kosmicznych, po czym rakieta jednego z braci zawróciła i zaczęła lecieć z dwukrotnie większą prędkością w kierunku drugiego z braci. Który z braci będzie starszy w momencie ich spotkania?

15. Jakie są konsekwencje szczególnej teorii względności dla filozoficznego sporu o status czasu i przestrzeni między absolutyzmem a relacjonizmem? Czy konsekwencje te dotyczą także sporu o status czasoprzestrzeni?

16. Czy teoria względności obala przekonanie o istnieniu obiektywnego upływu czasu? Rozważ możliwe próby obrony tego przekonania.

Literatura uzupełniająca

Bardzo dobre wprowadzenie do szczególnej teorii względności znajdziecie w pierwszym rozdziale znanego podręcznika: B.F. Schutz, *Wstęp do ogólnej teorii względności*, PWN, Warszawa 1995.

Niektóre historyczne aspekty rozwoju szczególnej teorii względności w kontekście klasycznej teorii elektromagnetyzmu omówione są w cytowanej już książce: J. T. Cushing, *Philosophical Concepts in Physics*, Cambridge University Press, Cambridge 1998.

Niniejszy rozdział zawdzięcza wiele eleganckiemu wprowadzeniu do matematycznych podstaw teorii względności w drugim tomie z cyklu „Co musisz wiedzieć, żeby zacząć zajmować się fizyką": L Susskind, A. Friedman, *Szczególna teoria względności i klasyczna teoria pola. Teoretyczne minimum*, Prószyński i S-ka, Warszawa 2021.

ROZDZIAŁ 6. OGÓLNA TEORIA WZGLĘDNOŚCI

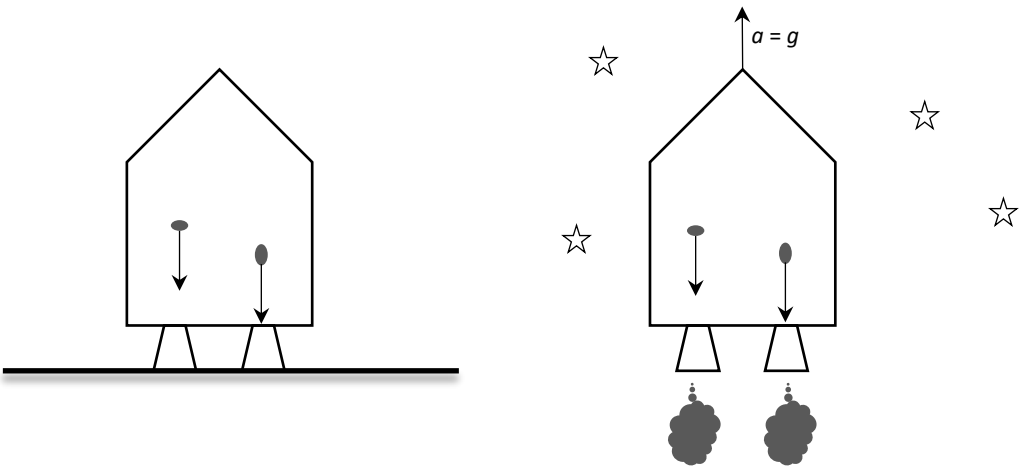
Co jest szczególnego w szczególnej teorii względności (STW)? Jak sama nazwa sugeruje, teoria ta dotyczy jedynie pewnego specjalnego przypadku, a zatem wymaga odpowiedniego uogólnienia. Szczególna teoria względności opisuje zjawiska mechaniczne w wyróżnionych układach odniesienia, mianowicie w układach inercjalnych. Pamiętamy, że transformacje Lorentza, które są fundamentem STW, łączą współrzędne czasowe i przestrzenne mierzone w jednym układzie ze współrzędnymi w innym układzie, poruszającym się ze stałą prędkością względem tego pierwszego. Co jednak, jeśli drugi układ zacznie przyspieszać? Szczególna teoria względności nie odpowiada na pytanie, jak w takim układzie będą wyglądały pomiary czasu i przestrzeni. Do tego potrzebna jest nowa teoria, zwana ogólną teorią względności (OTW). Jej fundamentem jest postulat, aby wszystkie układy odniesienia traktować jednakowo, niezależnie od tego, w jakim stanie ruchu się znajdują ani nawet czy przyjęte współrzędne są prostoliniowe, czy krzywoliniowe. Jednym z celów, jakie przyświecały Einsteinowi przy próbie sformułowania nowej uogólnionej teorii czasu i przestrzeni, było, aby równania tej teorii „wyglądały" tak samo w każdym możliwym do wyobrażenia układzie odniesienia. Przez układ odniesienia rozumiemy dowolne przyporządkowanie czwórki liczb (współrzędnych) każdemu zdarzeniu w czasoprzestrzeni. Jedynym warunkiem nałożonym na takie przyporządkowanie (zwane w języku matematyki mapą) jest wymóg, aby respektowało ono relację „bliskości" między punktami (matematycznie relację bliskości określa się mianem „topologii"). Chodzi o to, aby mapa narzucona na rozmaitość złożoną z punktów czasoprzestrzennych nie dokonywała „skoków" przy przejściu od punktu do sąsiadującego z nim punktu, a zatem była zgodna z topologią czasoprzestrzeni.

Jak jednak w praktyce dokonać takiego uogólnienia pojęcia układu współrzędnych? Okazuje się, że kluczem do rozwiązania tego problemu jest grawitacja. Wydawać się to może zaskakujące, ale jak zobaczymy wkrótce, istnieje zdumiewająco bliska analogia między polem grawitacyjnym a przyspieszającymi układami odniesienia. Ogólna teoria względności realizuje równocześnie dwa na pozór odmienne cele. Dokonuje zrównania statusu wszystkich układów odniesienia oraz formułuje zupełnie nową koncepcję oddziaływań grawitacyjnych. Grawitacja jest tutaj czymś w rodzaju zaburzenia czasu i przestrzeni, a nie zwykłym oddziaływaniem między ciałami obdarzonymi masą. Aby sformułować nową teorię grawitacji, potrzebne będzie wykorzystanie subtelnych pojęć geometrycznych opisujących krzywiznę czasu i przestrzeni. Pojęcia te umożliwią nam wyprecyzowanie jeszcze jednej ważnej różnicy

między szczególną a ogólną teorią względności. Szczególna teoria opisuje czasoprzestrzeń, która jest „płaska" – nie w sensie bycia dwuwymiarową powierzchnią, ale w sensie braku zakrzywienia. Linie proste w czasoprzestrzeni Minkowskiego są „naprawdę" proste. Natomiast geometria ogólnej teorii względności jest zakrzywiona, podobnie do geometrii na powierzchni kuli. W takiej geometrii nawet możliwie najprostsze linie mają swoją wewnętrzną krzywiznę. Krzywizna w teorii względności to rezultat obecności ciał obdarzonych masą, a jej obserwowalnym skutkiem jest zachowanie ciał w polu grawitacyjnym, które poruszają się tak, jakby działała na nie siła. Jednak to nie siła powoduje odpowiednie zakrzywienie torów ciał w polu grawitacyjnym, a sama geometria czasu i przestrzeni.

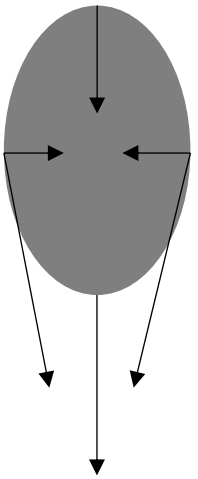
6.1. Zasada równoważności Einsteina

Zacznijmy jednakże stopniowo od omówienia zasady równoważności Einsteina. Pamiętamy z wcześniejszych rozdziałów, że Galileusz argumentował za zasadą względności – żadne zjawiska fizyczne nie powinny wyróżniać układów poruszających się względem siebie ze stałą prędkością. Słynny argument Galileusza ze statkiem pokazuje, że nie jesteśmy w stanie odróżnić za pomocą „wewnętrznych" obserwacji, czy znajdujemy się pod pokładem statku poruszającego się ze stałą prędkością, czy też statku spoczywającego. Natomiast jeśli statek zacznie przyspieszać – np. kołysząc się na wodzie – wywoła to obserwowalne efekty w postaci przesuwania się przedmiotów pod wpływem sił pozornych. Einstein zauważył, że argument Galileusza z nieodróżnialności obserwacyjnej układów inercjalnych można uogólnić na przypadek układów przyspieszających, jeśli uwzględni się grawitację. W tym celu rozważmy dwie sytuacje – w jednej rakieta kosmiczna porusza się z dala od źródeł grawitacji ze stałym przyspieszeniem równym przyspieszeniu ziemskiemu $a = g$, a w drugiej rakieta stoi nieruchomo na powierzchni Ziemi (rys. 6.1). W obu przypadkach zachowanie przedmiotów wewnątrz rakiety będzie takie samo – wszystkie ciała będą spadać ze stałym przyspieszeniem równym g. Zatem astronauta zamknięty w rakiecie nie będzie w stanie na podstawie lokalnych obserwacji stwierdzić, która z sytuacji ma miejsce.



Rys. 6.1. Zasada równoważności Einsteina

Zwróćmy uwagę na dwa ważne aspekty zasady równoważności Einsteina. Po pierwsze, zasada ta wynika z empirycznego faktu równości między masą bezwładnościową a masą grawitacyjną. Jak pamiętamy z rozdziału poświęconego mechanice klasycznej, masa pełni w niej dwojaką rolę: jako współczynnik proporcjonalności w drugiej zasadzie dynamiki $F = ma$ oraz jako miara „ładunku" grawitacyjnego dla danego ciała, występująca w prawie powszechnego ciążenia Newtona. Tylko przy założeniu, że te dwie masy są identyczne, możemy wyprowadzić wniosek, że wszystkie ciała w polu grawitacyjnym będą spadać z takim samym przyspieszeniem, a zatem ich zachowanie będzie nieodróżnialne od zachowania w układzie przyspieszającym. Po drugie, tezę o równoważności układu przyspieszającego z układem spoczywającym w polu grawitacyjnym należy opatrzeć ograniczeniem „lokalnie". W istocie istnieje obserwacyjny sposób na odróżnienie obu sytuacji, jeśli uwzględnimy ciała znacznie od siebie oddalone. W wypadku układu przyspieszającego kierunek działania siły pozornej jest taki, sam niezależnie od położenia rozważanych ciał (jest on skierowany przeciwnie do kierunku przyspieszenia całego układu). Natomiast w centralnym polu grawitacyjnym, pochodzącym od ciała takiego jak Ziemia, zarówno kierunek jak i wartość siły grawitacyjnej (przyspieszenia grawitacyjnego) zmieniają się w zależności od położenia ciała. Ciała rozdzielone „w poziomie" będą się *de facto* nieco zbliżać do siebie, a ciała rozdzielone w pionie oddalać (efekt ten jest często opisywany jako wywołany działaniem tzw. „sił pływowych", które skutkują odpowiednią deformacją rozciągłych ciał w polu grawitacyjnym – por. rys. 6.2). Zatem obserwacyjna równoważność zachodzi tylko dla małych obszarów, dla których możemy pominąć efekty pływowe. Pod tym względem zasada równoważności Einsteina różni się od zasady względności Galileusza – ta ostatnia stosuje się globalnie do wszystkich zjawisk w danym układzie, niezależnie od ich wzajemnego przestrzennego oddalenia.[1]

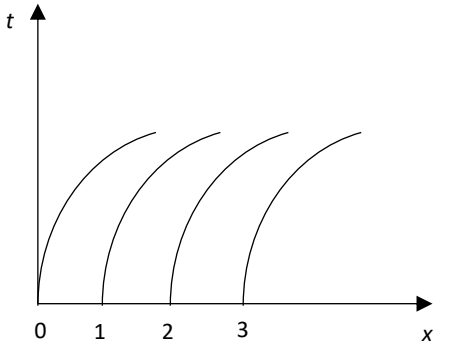


Rys. 6.2. Siły pływowe w centralnym polu grawitacyjnym

[1] Powstaje pytanie: co z siłami inercjalnymi powstającymi w układach obracających się, takimi jak siła odśrodkowa czy Coriolisa? Czy da się je „wyeliminować" w analogiczny do powyższego sposób, przechodząc do układu spoczywającego w polu grawitacyjnym? Odpowiedź jest twierdząca, ale wymaga to przejścia z klasycznej teorii grawitacji do teorii grawitacji jako zakrzywienia czasoprzestrzeni. W klasycznej teorii grawitacji nie ma miejsca np. na siły zależne od prędkości ciał, jak wypadku sił Coriolisa.

Innym przykładem lokalnej empirycznej nieodróżnialności jest przypadek swobodnego spadku w polu grawitacyjnym. Osoby w spadającej windzie znajdują się w stanie nieważkości ze względu na równoważenie się siły grawitacyjnej z siłą pozorną (to samo zresztą dotyczy mniej dramatycznej sytuacji orbitowania wokół Ziemi na stacji kosmicznej). Zatem lokalnie taka sytuacja jest nie do odróżnienia od pozostawania w układzie inercjalnym z dala od wszelkich źródeł grawitacji. Fakt ten zdaje się sugerować następujące uogólnienie pojęcia układu inercjalnego w przypadku występowania pola grawitacyjnego: układem inercjalnym możemy nazwać każdy układ znajdujący się w spadku swobodnym w polu grawitacyjnym. Innymi słowy, spadek swobodny będzie odpowiednikiem prostoliniowego i jednostajnego ruchu ciała pozbawionego działania wszelkich sił, opisanego pierwszym prawem dynamiki Newtona. Pozostaje oczywiście kwestia, że trajektoria czasoprzestrzenna ciała spadającego swobodnie w polu grawitacyjnym nie jest linią prostą (jest zakrzywiona w przestrzeni lub czasoprzestrzeni, ze względu na występujące przyspieszenie). Tę różnicę między klasycznym ruchem swobodnym a spadkiem swobodnym w polu grawitacyjnym będzie można jednak „zminimalizować" przez wprowadzenie pojęcia wewnętrznej krzywizny geometrii i geodezyjnych.

Możliwy związek między grawitacją a krzywizną czasoprzestrzeni można zauważyć, korzystając z zasady równoważności. Z jednej strony wiemy, że pole grawitacyjne jest empirycznie równoważne przyspieszeniu odpowiedniego układu nieinercjalnego. Z drugiej strony, układy przyspieszające charakteryzują się osiami współrzędnych, które są liniami zakrzywionymi z perspektywy układu nieprzyspieszającego. Jak widzimy na rys. 6.3, punkty czasoprzestrzenne, których współrzędne przestrzenne w układzie jednostajnie przyspieszającym są takie same (np. $x = 0$, $x = 1$ i tak dalej), tworzą parabole, a nie linie proste. Można więc sugerować, że grawitacja ma coś wspólnego z zakrzywieniem. Jednakże tę luźną myśl należy odpowiednio wyprecyzować pojęciowo i matematycznie. W następnym paragrafie spróbujemy przybliżyć pojęcie wewnętrznej krzywizny (czaso)przestrzeni i jego związek z kinematycznym zachowaniem ciał.



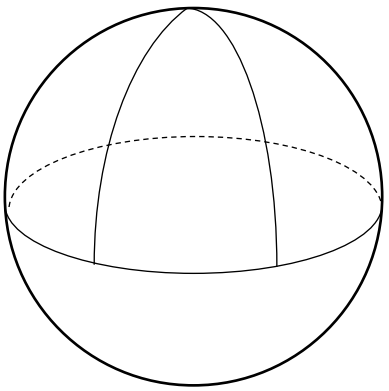
Rys. 6.3. Punkty o tych samych współrzędnych w układzie przyspieszającym

6.2. Krzywizna (czaso)przestrzeni

Popularne ujęcia ogólnej teorii względności często odwołują się do przykładu geometrii na powierzchni kuli (na sferze). Jest to przypadek bardzo użyteczny do celów dydaktycznych,

jednak obciążony istotnymi ograniczeniami. Po pierwsze, geometria sfery jest dwuwymiarowa, podczas gdy przestrzeń fizyczna ma trzy wymiary. Po drugie – i ważniejsze – sfera jest obiektem przestrzennym, natomiast teoria względności bierze pod uwagę geometryczne własności całej czterowymiarowej czasoprzestrzeni. Dopiero krzywizna czasoprzestrzeni z niepomijalnym aspektem czasowym ujawnia istotne własności zachowania ciał w polu grawitacyjnym. Mając w pamięci to ostrzeżenie, posłużmy się mimo wszystko przykładem sfery jako poglądowym, aczkolwiek uproszczonym narzędziem.

Wybierzmy dowolne dwa punkty na sferze. Jak zdefiniować odległość między tymi punktami, nie opuszczając przy tym dwuwymiarowej powierzchni? Naturalnym podejściem, stosowanym zresztą na co dzień w geografii i transporcie lotniczym, jest minimalizacja długości dróg między tymi punktami. Odległością będzie długość najkrótszej z możliwych dróg dojścia od jednego do drugiego punktu. Z oczywistych względów najkrótszą drogą nie będzie linia prosta – można się przekonać, że najmniejszą odległość osiągniemy poruszając się po fragmentach wielkich kół, tj. okręgów powstałych przez przecięcie dwóch wybranych punktów na sferze i środka sfery jedną płaszczyzną. Wielkie koła na sferze reprezentują linie „proste" (czy też najprostsze z możliwych) w danej geometrii. Takie linie nazywamy „geodezyjnymi". Rozważmy teraz dwa punkty na równiku i wyprowadźmy z nich dwie geodezyjne pod kątem prostym do równika (rys. 6.4). Wyobraźmy sobie, że po obu liniach poruszają się np. dwie mrówki. Choć na początku wędrówki ich drogi można było uznać za równoległe, to jednak w momencie dojścia do bieguna drogi się zbiegają. Z punktu widzenia mrówek wygląda to tak, jakby istniała między nimi siła przyciągająca, która doprowadziła w końcu do spotkania. Jednakże w rzeczywistości nie ma żadnej siły, a jedynie krzywizna geometrii, która powoduje, że „proste" wyglądające pierwotnie na równoległe mogą się zetknąć w odległym punkcie.



Rys. 6.4. Geodezyjne prostopadłe do równika zbiegają się w biegunie

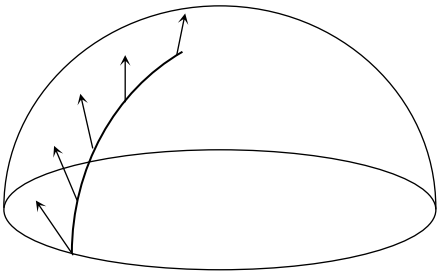
Warto wspomnieć jeden z fundamentów geometrii Euklidesowej, jakim jest tzw. piąty postulat (aksjomat równoległych). Geometria, jakiej uczymy się w szkole, opiera się na intuicyjnym założeniu, że przez jeden punkt można przeprowadzić dokładnie jedną prostą równoległą do danej prostej. Okazuje się, że na sferze zasada ta jest złamana, jeśli tylko zinterpretujemy pojęcie linii prostych jako odnoszące się do wielkich okręgów. Dla danej

„prostej" (geodezyjnej) na sferze nie istnieje żadna prosta do niej równoległa. Wynika to z faktu, że każde dwa wielkie koła na sferze muszą się przeciąć. Geometria sfery jest nieeuklidesowa – należy ona do kategorii tzw. geometrii Riemanna o dodatniej krzywiźnie. Są ponadto geometrie nieeuklidesowe o ujemnej krzywiźnie (tzw. geometria Łobaczewskiego), dla których mamy więcej niż jedną prostą równoległą do danej przechodzącą przez wybrany punkt. Dwuwymiarowa geometria Łobaczewskiego może być zilustrowana przykładem „siodła" (przełęczy górskiej).

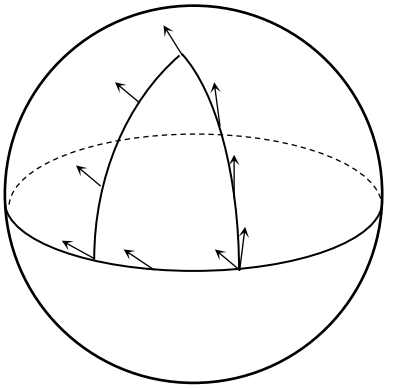
Wracając do naszego przykładu z mrówkami na sferze, zauważmy, że dostarcza on nam poglądowego argumentu za tym, że oddziaływania grawitacyjne można próbować wyjaśnić czysto geometrycznie przy pomocy pojęcia krzywizny. Trzeba przy tym podkreślić, że geometryczna interpretacja możliwa jest tylko w wypadku grawitacji, a nie innych oddziaływań (np. elektromagnetycznych), ponieważ tylko grawitacja ma charakter uniwersalny, tj. działa na wszystkie ciała jednakowo. Dzięki temu faktowi możemy zinterpretować grawitację jako własność (czaso)przestrzeni niezależną od tego, jakie ciało znajduje się w danym obszarze. Podstawowa idea tego podejścia jest taka, że swobodne ciała w polu grawitacyjnym poruszają się po liniach „prostych" (czyli geodezyjnych), tak jakby nie były poddane działaniu żadnej siły. Jednakże same linie geodezyjne charakteryzują się wewnętrznym zakrzywieniem z powodu odstępstw od geometrii Euklidesowej. Jak się przekonamy, zakrzywienie to jest uzależnione od występujących w danym obszarze mas.

Przyjrzyjmy się bliżej pojęciu krzywizny, które na razie traktowaliśmy dość swobodnie, opierając się na intuicjach. Jak rozpoznać, że dana dwuwymiarowa powierzchnia jest zakrzywiona? Okazuje się, że nie jest to banalne zadanie. Dla ilustracji problemu porównajmy dwa przypadki: rozważaną wyżej powierzchnię kuli oraz powierzchnię walca. Intuicyjnie obie powierzchnie wyglądają na zakrzywione. Jednakże matematycznie tylko pierwsza z nich spełnia formalne warunki krzywizny. Geometria powierzchni walca jest całkowicie Euklidesowa: dwie proste prostopadłe do trzeciej nie przetną się ze sobą. Ujmując sprawę obrazowo: powierzchnię walca można „rozprostować", tworząc zwykłą dwuwymiarową płaszczyznę (tak jak można rozprostować gazetę zwiniętą w rulon). Natomiast analogiczne rozprostowanie w wypadku sfery jest niemożliwe bez zmiany wewnętrznej geometrii. Te intuicyjne rozważania mogą być ujęte nieco precyzyjniej przy pomocy ważnego pojęcia „przeniesienia równoległego" (*paralel transport*). Rozważmy dowolny punkt na sferze i wyprowadźmy z niego wektor styczny do sfery w pewnym kierunku. Matematycznie można zdefiniować następującą operację na wybranym wektorze: łączymy dwa punkty dowolną linią, a następnie „przesuwamy" wektor z punktu początkowego wzdłuż tej linii do punktu końcowego, zachowując jego kierunek (zdefiniowany przy pomocy tzw. infinitezymalnego pojęcia równoległości). Innymi słowy, dzielimy daną krzywą na bardzo małe („nieskończenie małe") odcinki i na każdym z nich staramy się zachować ten sam kierunek danego wektora (rys. 6.5).[2] Oczywiście z perspektywy zewnętrznej wektor styczny do powierzchni sfery będzie zmieniał kierunek, ale względem samej sfery kierunek pozostaje niezmienny dzięki temu, że w każdym niewielkim kroku będziemy zachowywać równoległość.

[2] Formalnie to zachowanie kierunku (czy też infinitezymalną równoległość) wyraża się operacją zwaną pochodną kowariantną lub koneksją. Wspomnimy o tym pojęciu w paragrafach oznaczonych „gwiazdką".



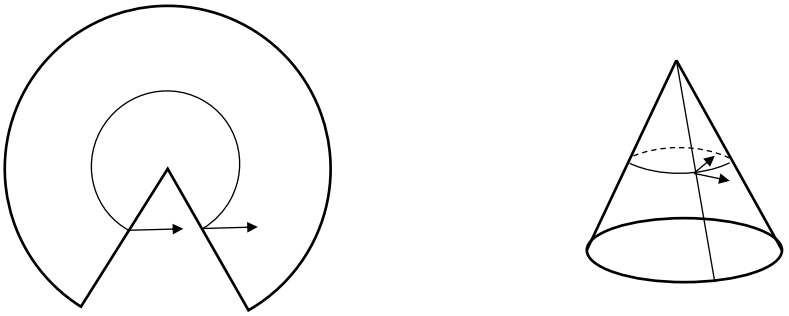
Rys. 6.5. Przeniesienie równoległe wektora na sferze



Rys. 6.6. Przeniesienie równoległe wektora po krzywej zamkniętej na sferze. Wektor końcowy nie pokrywa się z początkowym

Przy pomocy pojęcia przeniesienia równoległego możemy zdefiniować matematyczne pojęcie krzywizny. W tym celu musimy rozważyć linię zamkniętą na danej powierzchni. Dokonując przesunięcia równoległego danego wektora po tej linii, porównujemy ze sobą wektor początkowy i końcowy w tym samym punkcie. Jeśli oba wektory się pokryją, mamy do czynienia z powierzchnią płaską. Jeśli natomiast wektor końcowy będzie skierowany w inną stronę niż wektor początkowy, oznacza to, że rozważana powierzchnia charakteryzuje się krzywizną, której wielkość jest zależna od wielkości kąta między obydwoma wektorami (im większa jest rozbieżność kątowa między nimi, tym większa krzywizna). Zastosujmy to pojęcie do przypadku sfery. Rozważmy łamaną linię zamkniętą w kształcie trójkąta, którego bokami są odpowiednio: odcinek równika i dwie linie łączące się w pewnym punkcie. Wychodząc z punktu na równiku i przenosząc dany wektor równolegle po pierwszej z linii, a następnie wracając wzdłuż innej linii i równika, otrzymamy wektor różniący się kierunkiem od oryginału (rys. 6.6). Natomiast analogiczna operacja na powierzchni walca nie przyniesie zmiany kierunku (najlepiej to zauważyć „rozwijając" powierzchnię walca – od razu widać, że powierzchnia musi być płaska).

Ciekawym przypadkiem krzywizny jest powierzchnia stożka. Ponieważ stożek można rozwinąć bez deformacji do postaci płaskiej (wyobraźmy sobie rozcięcie stożka linią wychodzącą z wierzchołka, tzw. tworzącą stożka, a następnie rozprostowanie go – dostaniemy wtedy powierzchnię płaską z „wyciętym" trójkątem), generalnie jego powierzchnia charakteryzuje się zerową krzywizną. Jednakże jest wyjątek, mianowicie wierzchołek stożka. Aby przekonać się, że mamy tu do czynienia z niezerową krzywizną, rozważmy linię zamkniętą wokół wierzchołka (np. okrąg). Na rozprostowanej powierzchni stożka łatwo zauważymy, że wybierając w punkcie początkowym dany wektor i przemieszczając go wzdłuż naszej krzywej, otrzymamy na końcu drogi inny wektor (pamiętajmy o tym, że punkt końcowy drogi na rozcięciu jest tożsamy z punktem początkowym – por. rys. 6.7) Można pokazać, że kąt między wektorem początkowym i końcowym jest równy kątowi rozwarcia naszego stożka. Zmiana kierunku transportowanego wektora zachodzi tylko w wypadku linii zamkniętych obejmujących wierzchołek – dla wszystkich innych linii zamkniętych przeniesienie nie zmieni kierunku wektora.



Rys. 6.7. Przeniesienie równoległe wektora po krzywej zawierającej wierzchołek stożka. Rysunek po lewej powstał w wyniku rozcięcia stożka wzdłuż jego tworzącej i rozłożeniu na płaszczyźnie. Ponieważ wektory początkowy i końcowy tworzą inne kąty z linią rozcięcia, po ponownym „sklejeniu" wektory te nie pokryją się ze sobą (rys. po prawej), co wskazuje na istnienie krzywizny w wierzchołku stożka

Matematycznie ścisła definicja pojęcia krzywizny określa ją w danym punkcie. Należy więc zmodyfikować nasze kryterium oparte na transporcie równoległym wektorów po krzywej zamkniętej – trzeba mianowicie przejść w granicy od coraz mniejszych krzywych zamkniętych do danego punktu. Oczywiście w takim wypadku sama wielkość rozbieżności kątowej jest bezużyteczna jako miara krzywizny w punkcie, ponieważ różnica między wektorem początkowym a końcowym podczas transportu równoległego będzie zbiegać do zera dla coraz mniejszych „pętli". Rozwiązaniem jest przyjęcie miary krzywizny R jako stosunku rozbieżności kątowej do pola powierzchni zamkniętej daną krzywą.[3] W takim wypadku krzywi-

[3] Warto dodać, że pełna definicja krzywizny obejmuje również znak, który może być dodatni lub ujemny. Aby to uzyskać, porównujemy kierunek „obrotu" wybranego wektora podczas transportu równoległego z kierunkiem obiegu krzywej. Jeśli oba kierunki są takie same, krzywizna jest dodatnia (jak na powierzchni kuli). Jeśli natomiast przeniesiony wektor obróci się w kierunku przeciwnym do

zna będzie stosunkiem dwóch liczb zbiegających do zera, który może okazać się dowolną liczbą. Nietrudno zauważyć, że dla stożka i linii zamkniętych obejmujących jego wierzchołek (por. wpis w ramce) stosunek ten będzie zbiegał do nieskończoności, gdyż rozwartość kątowa między wektorem początkowym a końcowym jest stała niezależnie od danej krzywej, a pole zamknięte krzywą oczywiście zbiega do zera dla coraz to mniejszych krzywych. Zatem krzywizna w punkcie wierzchołkowym stożka jest nieskończona.

Przypadek dwuwymiarowych powierzchni zakrzywionych można uogólnić na dowolną liczbę wymiarów, tak aby zastosować pojęcie krzywizny np. do czterowymiarowej czasoprzestrzeni. Ogólnie jednak krzywizna będzie wtedy reprezentowana nie przez liczbę *R* (skalar), ale przez obiekt matematyczny zwany tensorem. Z tensorami zetknęliśmy się już w rozdziałach poświęconych elektromagnetyzmowi i szczególnej teorii względności – w uproszczeniu są to obiekty reprezentowane w każdym układzie współrzędnych przez *n*-tki liczb (składowe tensora). Każdy tensor posiada pewną liczbę indeksów, które określają jego *typ*. Tensor o jednym indeksie jest po prostu wektorem, a liczba jego składowych jest równa wymiarowi danej przestrzeni. Tensor o liczbie indeksów *k* będzie miał $k \cdot d$ składowych, gdzie *d* jest wymiarem przestrzeni. Okazuje się, że tensor krzywizny (zwany również tensorem Riemanna) w ogólności musi mieć cztery indeksy: $R_{\alpha\beta\gamma\delta}$, a zatem charakteryzuje się on szesnastoma składowymi w czterowymiarowej czasoprzestrzeni. Nie wszystkie jednak składowe tensora Riemanna są różne.

Tensor Riemanna jest powiązany z innym tensorem, który odgrywa fundamentalną rolę w geometrii czasoprzestrzeni. Jest to tensor metryczny, który określa „infinitezymalną" (nieskończenie małą) odległość w danym punkcie przestrzeni. Przy jego pomocy można precyzyjnie określić odległość między dowolnymi dwoma punktami w przestrzeni (w języku matematycznym odległość nazywa się często „metryką", stąd nazwa tensor metryczny). Warto przyjrzeć się mu dokładniej. Zacznijmy od znanej formuły na odległość między dwoma punktami w geometrii Euklidesa, gdy znana jest różnica między współrzędnymi tych punktów. Przyjmijmy, że nasza przestrzeń ma trzy wymiary, a współrzędne dowolnego punktu dane są w postaci trzech liczb *x*, *y* i *z*. Różnice współrzędnych między dwoma nieskończenie bliskimi punktami (określa się je mianem różniczek) oznaczamy jako *dx*, *dy* i *dz*. Zgodnie z twierdzeniem Pitagorasa, odległość między takimi punktami jest dana wzorem $ds^2 = dx^2 + dy^2 + dz^2$. Ogólnie jednak wyrażenie na infinitezymalną odległość (przesunięcie) może zawierać dowolne kombinacje różniczek współrzędnych: *dxdy*, *dzdy* itd. Wzór opisujący wszystkie możliwe warianty postaci infinitezymalnego przesunięcia można napisać jako:

$$ds^2 = \sum_{i,j=1}^{3} g_{ij} dx^i dx^j.$$

Zamiast pisać *x*, *y*, *z*, stosujemy teraz bardziej uniwersalną formę współrzędnych x^1, x^2, x^3. Współczynniki g_{ij} są właśnie składowymi tensora metrycznego, który jak widać, ma typ równy 2. W wypadku geometrii Euklidesowej wyrażonej we współrzędnych kartezjańskich tensor metryczny będzie miał bardzo proste składowe: $g_{11} = g_{22} = g_{33} = 1$, a wszystkie pozostałe składowe są równe 0. Wyraża się to czasami przy pomocy użytecznego symbolu zwanego deltą Kroneckera δ_{ij}; przyjmuje ona wartość jeden, gdy $i = j$, a zero, gdy $i \neq j$. Należy pamiętać, że tensor metryczny będzie miał różną postać w różnych krzywoliniowych ukła-

kierunku obiegu, mamy do czynienia z krzywizną ujemną, jak na powierzchni siodła.

dach współrzędnych. Jednakże o ile rozważana przestrzeń ma zerową krzywiznę, zawsze można znaleźć układ współrzędnych, w którym tensor metryczny przyjmuje formę delty Kroneckera. Natomiast jeśli dana przestrzeń charakteryzuje się niezerową krzywizną, tensor metryczny nie ma formy δ_{ij} w żadnym układzie współrzędnych. Dla ilustracji podajmy formę tensora metrycznego określającego geometrię na sferze. Standardowe współrzędne na sferze to tzw. współrzędne sferyczne, określone przy pomocy dwóch kątów θ (szerokość geograficzna) i φ (długość geograficzna). Kąt θ przebiega wartości od $-\frac{\pi}{2}$ do $\frac{\pi}{2}$, a kąt φ od 0 do 2π. Można policzyć, że różniczkowa odległość ds^2 na sferze o promieniu r dana jest następującym wzorem w zależności od kątów φ i θ:

$$ds^2 = r^2 d\theta^2 + r^2 \sin^2\theta \, d\varphi^2.$$

Zatem tensor metryczny przyjmuje następujące współrzędne: $g_{11} = r^2$, $g_{22} = r^2 \sin^2\theta$, a pozostałe współrzędne $g_{12} = g_{21} = 0$. Jest on różny od delty Kroneckera δ_{ij} i nie da się przedstawić za jej pomocą w żadnym układzie współrzędnych.

Wspomnijmy jeszcze, że pojęcie przeniesienia równoległego może posłużyć nam do alternatywnego zdefiniowania geodezyjnych, które uprzednio scharakteryzowaliśmy jako najkrótsze linie łączące dane punkty. Rozważmy dowolną linię i wyprowadźmy w każdym jej punkcie wektor do niej styczny. Jeśli teraz okaże się, że wektory styczne do danej linii są do siebie równoległe, linię tą nazwiemy geodezyjną. Warunek równoległości stycznych można zapisać w postaci równania, które zrównuje pochodną kowariantną wektora stycznego z zerem. Można to zapisać formalnie jako

$$\nabla_\mu V^\mu = 0,$$

gdzie ∇_μ oznacza pochodną kowariantną (zwaną również koneksją afiniczną), a V^μ – wektor styczny do danej krzywej. Ponieważ pochodna kowariantna danego wektora reprezentuje jego zmienność wzdłuż danej krzywej, równanie geodezyjnych oznacza, że wektory styczne do linii geodezyjnych nie zmieniają swojego kierunku (pozostają równoległe).

Przejdźmy teraz z rozważań czysto geometrycznych do fizyki. W rozdziale poświęconym szczególnej teorii względności mówiliśmy, że ruch cząstki opisujemy za pomocą trajektorii w czterowymiarowej czasoprzestrzeni (jest to linia świata tej cząstki). Wprowadziliśmy również pojęcie czterowektora prędkości, który jest jednostkowym wektorem stycznym do linii świata cząstki, a jego składowe przestrzenne są składowymi zwykłej prędkości względem czasu własnego. Z powyższych rozważań wiemy, że jeśli trajektoria cząstki jest linią geodezyjną, wektory styczne do trajektorii nie zmieniają swojego kierunku, czyli ich pochodna jest równa zeru. Zatem dla cząstki poruszającej się po geodezyjnej jej czteroprędkość U^μ nie ulega zmianie przy przesuwaniu się wzdłuż trajektorii:

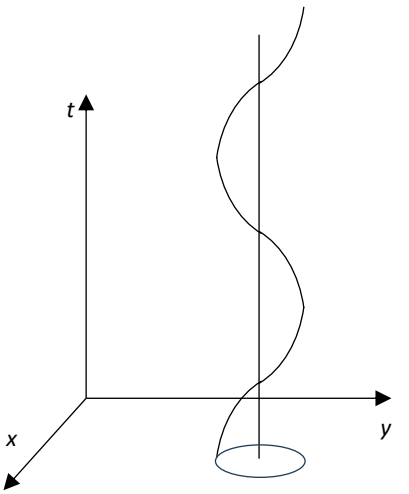
$$\nabla_\mu U^\mu = 0. \tag{6.1}$$

Można interpretować powyższe równanie jako uogólnienie pierwszego prawa dynamiki Newtona. Oznacza to, że cząstki swobodne, czyli niepoddane działaniu żadnych sił, poruszają się po geodezyjnych. Obserwacja ta stanowi klucz do opisu zachowania obiektów w geometrii z krzywizną. Zgodnie z zasadą bezwładności mechaniki klasycznej, ciała swobodne poruszają się po trajektoriach będących liniami prostymi.[4] W geometrii charakteryzu-

[4] Pamiętajmy, że chodzi tu o trajektorie w czasoprzestrzeni! Prostoliniowość trajektorii czasoprzestrzennej oznacza, że dane ciało nie będzie przyspieszać.

jącej się krzywizną zasada ta pozostaje zachowana, z tym że rolę linii prostych odgrywają linie geodezyjne. Grawitacja przejawia się jedynie w zakrzywieniu geometrii, a nie w formie działającej siły.

Ilustracją naszych rozważań może być opis obiegu planety dookoła Słońca. Na rysunku 6.8 została przedstawiona trajektoria czasoprzestrzenna planety. Przyjmuje ona kształt rozciągniętej spirali (z trzech wymiarów przestrzennych wybraliśmy dwa, które są wystarczające do opisu ruchu planety, jako że zachodzi on na płaszczyźnie). Zakrzywienie linii świata planety interpretujemy czysto geometrycznie, gdyż traktujemy trajektorię jako linię geodezyjną w geometrii o niezerowej krzywiźnie. Mówiąc nieco metaforycznie, to nie Słońce powoduje bezpośrednio zmienność toru planety, a sama czasoprzestrzeń, która „popycha" planetę w danym kierunku. Zwróćmy również uwagę na to, że tor planety w czterech wymiarach nie odbiega w sposób znaczący od „prawdziwej" linii prostej (pamiętajmy, że przy założeniu $c = 1$ jedna jednostka czasu będzie odpowiadać ogromnej odległości przestrzennej, pokonywanej przez światło w tym czasie). Skoro prędkość planety w ruchu obiegowym jest nieporównywalnie mniejsza od prędkości światła, jej linia świata będzie niemal pionowa i lokalnie będzie w bardzo niewielkim stopniu różnić się od linii prostej. Oznacza to, że krzywizna czasoprzestrzeni w obszarach odległych od Słońca jest niewielka, a geometria lokalnie odbiega w praktycznie niezauważalny sposób od Euklidesowej.



Rys. 6.8. Trajektorie czasoprzestrzenne Słońca i okrążającej go planety. Po odpowiednim dobraniu jednostek spiralna trajektoria planety „rozciągnie się" do postaci niemal linii prostej

6.3. Równanie Einsteina

Żadne popularne wprowadzenie do ogólnej teorii względności nie może się obejść bez omówienia podstawowego równania tej teorii, czyli równania pola Einsteina. Jego ścisłe sformułowanie wymaga dość zaawansowanej matematyki – wprowadzimy ją w paragrafach oznaczonych gwiazdką. Natomiast podstawowy sens fizyczny tego równania jest raczej prosty. W klasycznym newtonowskim ujęciu grawitację wytwarzają obiekty obdarzone masą, czyli materia. Ponieważ obecnie zastępujemy siły grawitacyjne „zaburzeniem" czasoprzestrzeni, a dokładniej jej krzywizną, naturalne wydaje się, że materia będzie wywoływać krzy-

wiznę czasoprzestrzeni. Równanie Einsteina precyzuje tę myśl. W najogólniejszym zarysie sens tego równania można wyrazić następująco:

geometria = materia.

Musimy jeszcze ustalić, jaka dokładnie własność materii odpowiada za jaką własność geometrii. Niestety sytuacja jest bardziej skomplikowana niż w teorii Newtona, w której jedyne źródło grawitacji stanowi masa. W ogólnej teorii względności to nadal (częściowa) prawda, lecz oprócz masy rolę źródeł pola grawitacyjnego pełni również energia, jako że w istocie jest to jedna i ta sama wielkość. Do tego na krzywiznę czasoprzestrzeni ma wpływ nie tylko sama energia, ale także to, jak zmienia się ona w czasie. Dokładniej w równaniu Einsteina występują wielkości określające *przepływ* energii w różnych kierunkach. Na przykład strumień pyłu kosmicznego niesie pewną energię, dla której możemy policzyć, jaka jej ilość przepływa przez jednostkę powierzchni w jednostce czasu (jest to analogiczne do pojęć strumieni pól elektrycznych i magnetycznych, omówionych w rozdziale poświęconym elektromagnetyzmowi). Dodatkowo wkład w zakrzywianie czasoprzestrzeni wnosi także pęd obiektów wraz z informacją, jak wiele takiego pędu przepływa przez daną powierzchnię. Chociaż w typowych sytuacjach wkład pochodzący od pędu jest nieporównywalnie mniejszy od wkładu energii (która jest wyrażona wzorem mc^2, a zatem ma ogromną wartość), to jednak w ścisłym równaniu należy uwzględnić także i ten element.

Syntetyczne ujęcie wszystkich wielkości reprezentujących materię i jej wkład w krzywiznę czasoprzestrzeni ma postać tzw. tensora energii-pędu, oznaczanego przez $T^{\mu\nu}$. Jest to tensor drugiego rzędu, a zatem posiada szesnaście składowych, które możemy przedstawić w postaci tabeli z czterema rzędami i kolumnami. Poszczególne rzędy zawierają wielkości charakteryzujące, odpowiednio, gęstość i strumień energii oraz gęstości i strumienie poszczególnych składowych pędu. Natomiast równanie Einsteina można przepisać w następującej formie:

$$G^{\mu\nu} = kT^{\mu\nu},$$

gdzie k jest pewną stałą. Wyrażenie po lewej stronie oznacza pewną wielkość charakteryzującą geometrię czasoprzestrzeni. Einstein długo rozważał różne możliwości zdefiniowania tajemniczego tensora $G^{\mu\nu}$. Kierował się przy tym dwiema wskazówkami. Po pierwsze, powyższe równanie powinno odtwarzać znane prawo grawitacji Newtona dla przypadku niewielkich mas oraz małych prędkości. Chodzi o to, że teoria grawitacji Newtona jest dobrze potwierdzona empirycznie przez dostępne nam dane, a zatem nowa teoria powinna również mieć takie potwierdzenie. Pozostaje jedynie problem, jak wykonać przejście graniczne od OTW do teorii Newtona, skoro posługują się one zupełnie odmiennymi pojęciowo metodami opisu grawitacji – w postaci sił czy też potencjałów u Newtona i w postaci własności metrycznych czasoprzestrzeni w OTW. Potrzebne jest tutaj pewne „prawo pomostowe", łączące stare ujęcie z nowym. Einstein pokazał, że w przypadku małych mas i prędkości rolę potencjału grawitacyjnego teorii newtonowskiej może pełnić czasowa składowa tensora metrycznego g_{00}. Daje to nam sugestię, że w sytuacji granicznej tensor $G^{\mu\nu}$ powinien przejść w odpowiednią funkcję składowej g_{00} (dokładniej w kombinację drugich pochodnych z g_{00}, zwaną laplasjanem – szczegóły znajdziecie w paragrafie z gwiazdką).

W tym miejscu warto wspomnieć bardzo ważną zasadę stosowaną w fizyce i omawianą przez filozofów, zwaną zasadą korespondencji (oryginalnie sformułował ją duński

fizyk Niels Bohr, o którym będziemy jeszcze mówić w następnym rozdziale). Zasada korespondencji głosi, że każda nowa teoria, która zastępuje starą, powinna w przybliżeniu „redukować" się do starej teorii w obszarze, w którym ta była dobrze potwierdzona empirycznie. Przykładem zastosowania tej zasady jest przejście graniczne od transformacji Lorentza w STW do transformacji Galileusza przy założeniu, że prędkości ciał są dużo mniejsze of prędkości światła *c* (oraz dodatkowo, że rozważane odległości przestrzenne są niewielkie). Obecnie dyskutujemy przykład innego zastosowania zasady korespondencji – równania OTW powinny odtworzyć równania STW dla przypadku niewielkich mas, a także przejść w równania klasycznej teorii grawitacji Newtona przy dodatkowym założeniu niewielkich prędkości. Jeszcze inny przykład zasady korespondencji w działaniu dotyczy mechaniki kwantowej.

Druga wskazówka uwzględnia zasadę zachowania energii i pędu. Ze względu na to, że pęd i energia nie mogą „zniknąć" w danym obszarze czasoprzestrzeni, tensor $T^{\mu\nu}$ musi spełniać równanie zwane równaniem ciągłości. Einstein zasugerował, że tensor geometryczny $G^{\mu\nu}$ powinien spełniać to samo równanie na mocy praw matematyki. Takie założenie miałoby automatycznie zapewnić spełnienie zasady zachowania energii i pędu. Ściślej rzecz ujmując, równanie Einsteina, które oczywiście nie jest dowodliwe matematycznie, implikowałoby prawo zachowania energii i pędu. Einstein pokazał, że oba narzucone warunki są spełnione przez pewną kombinację trzech wielkości: tensora krzywizny Ricciego $R^{\mu\nu}$ (powiązanego matematycznie z tensorem Riemanna), tensora metrycznego oraz skalara krzywizny R (kombinację tę nazywamy oczywiście tensorem Einsteina). Zatem równanie Einsteina „w pełnej krasie" wygląda następująco:[5]

$$R^{\mu\nu} - \frac{1}{2} g^{\mu\nu} R = kT^{\mu\nu}. \tag{6.2}$$

Mamy już dwie fundamentalne zasady, na których wspiera się ogólna teoria względności: równanie Einsteina, określające, jak materia wpływa na geometrię, oraz zasadę dynamiki głoszącą, że ciała poruszają się po liniach geodezyjnych (ujętą skrótowo w postaci równania $\nabla_\mu U^\mu = 0$). Hermann Weyl wyraził kwintesencję OTW w zgrabnym sformułowaniu „materia mówi czasoprzestrzeni, jak się ma zakrzywiać, czasoprzestrzeń mówi materii, jak się ma poruszać". Może być jednak zaskoczeniem, że związek grawitacji i ruchu z własnościami czasoprzestrzennymi nie jest wcale wyjątkową cechą ogólnej teorii względności, odróżniającą ją od klasycznej teorii Newtona. Okazuje się, że teoria Newtona daje się również „zgeometryzować" na wzór OTW. Zapisując klasyczne równanie ruchu dla cząstki w polu grawitacyjnym w języku geometrii różniczkowej, można potencjał grawitacyjny włączyć do pochodnej kowariantnej, uzyskując w ten sposób równanie geodezyjnej o takiej samej postaci co równanie OTW. Jednakże zakrzywiona geometria teorii Newtonowskiej będzie inna niż geometria OTW. Zasadnicza różnica polega na tym, że OTW jest uogólnieniem szczególnej teorii względności z jej strukturą geometrii Minkowskiego, która jest odmienna od struktury czasoprzestrzeni Newtona-Galileusza. Zatem ogólna teoria względności opiera się nie na dwóch, lecz trzech filarach. Są to: równanie Einsteina (6.2) łączące krzywiznę geometrii z rozkładem materii, równanie ruchu ciał w zakrzywionej czasoprzestrzeni (6.1) oraz lokalna geometria Minkowskiego w granicy słabych pól.

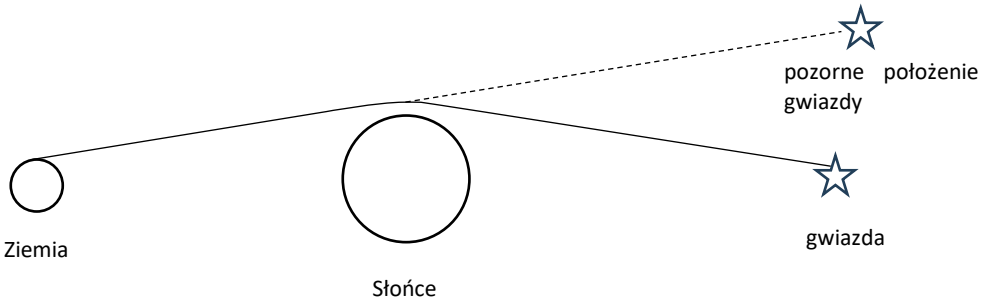
[5] Równanie to zostało niezależnie wyprowadzone przez niemieckiego matematyka Davida Hilberta. Niektórzy uważają, że powinno ono nosić nazwę równania Einsteina-Hilberta.

6.4. Empiryczne konsekwencje OTW

Ogólna teoria względności w niezmiernie elegancki sposób wyjaśnia obserwowalne fakty dotyczące oddziaływań grawitacyjnych – przede wszystkim ich unikalną cechę, jaką jest uniwersalność (nadawanie takiego samego przyspieszenia grawitacyjnego wszystkim ciałom). Realizuje ona również ideał niezależności fundamentalnych faktów fizycznych od dowolnie przyjętego układu odniesienia, w którym te fakty opisujemy. Jednakże trzeba przyznać, że OTW jest teorią bardzo skomplikowaną formalnie. Jej podstawowe równania wymagają bardzo zaawansowanych technik matematycznych, daleko wykraczających poza newtonowski i leibnizjański rachunek różniczkowy. Rozwiązanie równań Einsteina nawet dla najprostszych sytuacji jest zadaniem niezwykle trudnym i w wielu wypadkach praktycznie niewykonalnym ze względu na stopień matematycznej złożoności. Powstaje zatem pytanie, czy zastąpienie klasycznej teorii grawitacji Newtona teorią Einsteina jest uzasadnione empirycznymi faktami. Czy istnieją dane empiryczne, które jednoznacznie wskazują, że ogólna teoria względności prawidłowo opisuje rzeczywistość? Okazuje się, że istnieje szereg spektakularnych obserwacji i eksperymentów, które wzmacniają przekonanie, że teoria grawitacji Einsteina jest bliższa prawdy niż jej klasyczna poprzedniczka.

Ważna grupa testów empirycznych ogólnej teorii względności dotyczy zachowania promieni świetlnych w pobliżu obiektów o znacznej masie. W klasycznej teorii grawitacji światło nie oddziałuje z polem grawitacyjnym. Jak pokazaliśmy w rozdziale poświęconym szczególnej teorii względności, fotony nie posiadają masy spoczynkowej, a zatem zgodnie z prawem ciążenia Newtona siła oddziaływania grawitacyjnego na foton wynosi zero. W rezultacie promienie świetlne powinny poruszać się zawsze po liniach prostych, bez względu na otaczające je pole grawitacyjne. Inaczej natomiast sprawa wygląda na gruncie ogólnej teorii względności. Zgodnie z obowiązującymi w OTW zasadami dynamiki, wszystkie obiekty, włączając fotony, poruszają się po liniach geodezyjnych w czasoprzestrzeni. Jak wiemy, linie geodezyjne w pobliżu skupisk mas ulegają zakrzywieniu zgodnie z równaniem Einsteina. Zatem promień świetlny w pobliżu takich mas powinien również ulec widocznemu zakrzywieniu.

Obiektem o największej masie w naszym układzie jest oczywiście Słońce. Nasunęło to myśl, że największy efekt zakrzywienia powinien być obserwowany przy przejściu promienia świetlnego w pobliżu Słońca. Angielski astronom Arthur Eddington, zafascynowany teorią Einsteina, wpadł na pomysł empirycznego sprawdzenia tego efektu. Wybrał dogodny moment całkowitego zaćmienia Słońca (które akurat wypadało w Afryce, zatem Eddington musiał zorganizować wyprawę na ten kontynent). W czasie zaćmienia, które blokuje jasność słoneczną, Eddington zarejestrował na kliszy fotograficznej pozorne położenie gwiazd w pobliżu tarczy słonecznej. Porównał je następnie z „normalnym" położeniem z dala od Słońca (rys. 6.9). Okazało się, że gwiazdy widoczne w pobliżu tarczy słonecznej ulegają niewielkiemu przesunięciu, które odpowiada ugięciu promienia świetlnego przechodzącego obok Słońca (jest to efekt podobny do ugięcia promienia przebiegającego przez rozgrzane powietrze). Zatem przewidywanie teorii względności zostało potwierdzone (mamy tu do czynienia z klasycznym przykładem *experimentum crucis*, opisanym w pierwszym rozdziale).



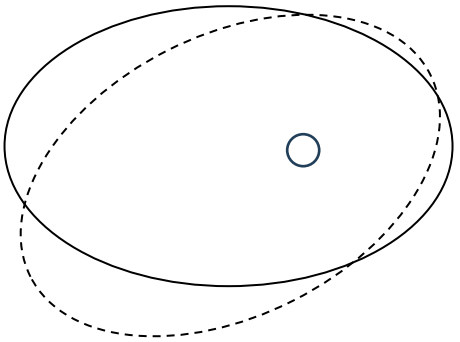
Rys. 6.9. Efekt ugięcia promienia świetlnego w pobliżu Słońca

Innego przykładu „anomalnego" (z punktu widzenia teorii klasycznej) zachowania światła w polu grawitacyjnym dostarcza eksperyment wykonany w latach sześćdziesiątych ubiegłego stulecia na powierzchni Ziemi (doświadczenie Pounda-Rebki-Snidera). Ogólna teoria względności przewiduje, że fala świetlna poruszająca się w polu grawitacyjnym w kierunku od źródła grawitacji (w kierunku malejącego pola) będzie zmniejszać swoją częstotliwość (wydłużać długość fali w kierunku ku czerwieni). Z kolei częstotliwość fali rozchodzącej się w kierunku danej masy będzie odpowiednio rosła (przesunięcie ku kolorowi niebieskiemu). Można to uzasadnić *quasi*-klasycznie przez wskazanie, że energia pojedynczego fotonu zależy od jego częstotliwości ν ($E = h\nu$), a ponieważ foton będzie tracił część swojej energii na pokonanie pola grawitacyjnego (lub też jej nabierał), jego częstotliwość musi ulec odpowiednio zmniejszeniu (lub zwiększeniu). Jednakże bardziej poprawne wytłumaczenie tego efektu odwołuje się do „wpływu" ciał obdarzonych masą na czasoprzestrzeń, a zatem także na czas. W zależności od odległości od danej masy, czas będzie płynął nieco inaczej (wolniej w pobliżu dużej masy), a ponieważ częstotliwość jest liczbą drgań na jednostkę czasu, będzie to miało wpływ na zmianę częstotliwości promieniowania.

Opisany powyżej efekt w pobliżu Ziemi jest niezwykle mały. Jednakże istnieje metoda eksperymentalna pozwalająca na jego detekcję. Opiera się ona na zjawisku emisji i absorpcji promieniowania gamma o ściśle określonej częstotliwości przez jądra atomowe. Dane jądro atomowe może wypromieniować foton wskutek przejścia jego nukleonu (protonu lub neutronu) z jednego poziomu energetycznego na drugi. To samo jądro może z powrotem zaabsorbować foton, ale musi się on charakteryzować dokładnie taką samą energią (musi dokładnie „trafić" w przerwę energetyczną pomiędzy stanem podstawowym a wzbudzonym jądra). Jeśli foton będzie miał nieco inną energię, absorpcja nie zajdzie. Doświadczenie wykorzystujące ten efekt ma bardzo prosty schemat: umieszczamy dwa jądra atomowe tego samego rodzaju na różnych wysokościach w polu grawitacyjnym Ziemi. Jądro znajdujące się np. wyżej wypromieniowuje foton, który następnie powinien być zaabsorbowany przez drugie jądro. W normalnych warunkach taka absorpcja powinna zachodzić bez przeszkód, gdyż foton ma dokładnie potrzebną energię. Jednak w polu grawitacyjnym energia ta ulegnie zmianie, a więc foton nie „trafi" w cel. Wynik doświadczenia pokazuje, że istotnie fotony zmieniają energię i częstotliwość podczas pokonywania pola grawitacyjnego.

Najbardziej spektakularna konsekwencja ogólnej teorii względności dotycząca wpływu materii na zachowanie światła to możliwość istnienia obiektów zwanych czarnymi dziurami. Przywołajmy jeszcze raz sytuację ugięcia promienia świetlnego w pobliżu Słońca. Co się stanie, jeśli rozważymy ciało o rozmiarach Słońca lecz o dużo większej masie? Jest dość oczywiste, że zakrzywienie promienia świetlnego będzie wtedy odpowiednio większe. Zwiększając masę ciała, teoretycznie powinniśmy osiągnąć sytuację, w której tor promienia zamknie się wokół tego ciała. Oznacza to, że światło nie będzie w stanie opuścić rejonu w pobliżu takiego supermasywnego obiektu – będzie „krążyć" wokół jego centrum w nieskończoność. Taki obiekt to oczywiście czarna dziura. Dla danego rozmiaru gwiazdy można policzyć, jaka musi być jej masa, aby stała się ona czarną dziurą. Przez długi czas czarne dziury uważane były za konstrukty czysto teoretyczne, głównie ze względu na niemożliwość ich bezpośredniego zaobserwowania (skoro, jak założyliśmy, światło nie jest w stanie się z takiego obiektu „wyrwać"). Jednakże można przekonać się o istnieniu czarnych dziur na drodze pośredniej. Na przykład w tzw. układzie podwójnym, w którym dwie gwiazdy obiegają się wzajemnie, jedna z gwiazd może być czarną dziurą. W takiej sytuacji zauważymy świecącą gwiazdę, obiegającą po ciasnej orbicie pewien punkt, w którym nie widać żadnego obiektu. Inny obserwowalny efekt istnienia czarnej dziury powstaje, kiedy zaczyna ona wchłaniać materię z sąsiadującego obiektu, przyspieszając jej strumień do niewiarygodnych prędkości. Taki strumień będzie wysyłać wysokoenergetyczne promieniowanie gamma, zanim zostanie całkowicie pochłonięty. Obserwacje podobnych efektów pozwalają na identyfikację wielkiej liczby czarnych dziur we wszechświecie, co stanowi mocne potwierdzenie ogólnej teorii względności.

Sukcesy empiryczne ogólnej teorii względności zostały przypieczętowane przypadkiem tzw. anomalnej precesji peryhelium Merkurego. Na ruch obiegowy po elipsie Merkurego – najbliższej Słońca planety Układu Słonecznego – nakłada się powolny obrót całej orbity, a zatem także najbliższego jej położenia względem Słońca, czyli peryhelium (rys. 6.10). Taki ruch może być wytłumaczony na gruncie mechaniki newtonowskiej zaburzającym wpływem innych planet, przede wszystkim Jowisza, największej planety naszego układu. Jednakże dokładne obliczenia wykonane na podstawie równań klasycznej teorii grawitacji odbiegały nieco od obserwowanej prędkości rotacji peryhelium. Różnica nie była wielka – 43 sekundy kątowe na stulecie (jedna sekunda kątowa jest równa 1/3600 stopnia). Dodajmy, że sama precesja odbywa się z prędkością 5600 sekund na sto lat, zatem rozbieżność teorii i eksperymentu wynosiła mniej niż 1%. Problem ten nie dawał jednak spokoju astronomom. Pojawiły się sugestie, że wykryta niezgodność jest rezultatem oddziaływania nowej nieznanej planety. Nie było to zupełnie rozwiązanie *ad hoc*. Parę dekad wcześniej podobny problem dotyczył niezgodności z teorią obserwowanej orbity planety Uran. Choć pojawiły się sugestie, że wspaniała i jak do tej pory, bezbłędna teoria Newtona może wymagać poprawki, astronomowie John Adams (Brytyjczyk) i Urbain Leverriere (Francuz) zasugerowali istnienie nowej planety, zaburzającej tor ruchu Urana, i obliczyli, gdzie się powinna znajdować. Astronom niemiecki Johann Galle wykorzystał ich obliczenia i odnalazł na nieboskłonie poszukiwaną planetę, nazwaną imieniem Neptuna, greckiego boga morza. Epizod ten został szybko uznany za największy, spektakularny sukces teorii Newtona. Mało kto przypuszczał, że teoria ta miała jeszcze tylko niewiele ponad pół wieku spokojnego istnienia, zanim została porzucona na korzyść obu teorii względności.



Rys. 6.10. Zjawisko precesji peryhelium w ruchu obiegowym Merkurego wokół Słońca

Skoro raz udał się podobny zabieg, było naturalne, aby powtórzyć go dla przypadku Merkurego. Zaczęto więc poszukiwać hipotetycznej planety Wulkan, która miała być tak blisko Słońca, że blask pochodzący od naszej gwiazdy uniemożliwiał jej obserwację gołym okiem. Jednakże mimo wykorzystania najnowocześniejszych instrumentów poszukiwania okazały się bezowocne. Dzisiaj jedynym wykorzystaniem „planety Wulkan" są rozważania filozofów na temat metafizyki nieistniejących obiektów. Problem jednakże pozostał do chwili, kiedy Einstein zabrał się za odpowiednie obliczenia, wykorzystujące równania OTW. Gdy w rezultacie pojawiła się liczba 43 sekund kątowych, stał się jasne, że odtąd Merkury będzie symbolem zarówno upadku starej dobrej teorii Newtona, jak i sukcesu młodej konkurentki. Dodajmy, że rozbieżności między teorią Einsteina a Newtona pojawiają się tylko dla silnych pól grawitacyjnych, jak w przypadku pola działającego na Merkurego. Pozostałe planety naszego układu znajdują się na tyle daleko od Słońca, że różnice w przewidywaniach obu teorii są zupełnie pomijalne i spokojnie można korzystać z równań mechaniki klasycznej np. do obliczania orbit sond kosmicznych wysyłanych z Ziemi.[6]

6.5. Ruch i czasoprzestrzeń

Przejdźmy teraz na grunt rozważań filozoficznych na temat natury czasu i przestrzeni oraz względności ruchu. Przypomnijmy, że idea relatywności ruchu (zgodnie z którą ruch przedmiotów ma sens jedynie w odniesieniu do innych ciał) pojawiła się po raz pierwszy u Galileusza w formie zasady względności. Jednakże zasada ta dotyczyła jedynie układów inercjalnych, tj. nieprzyspieszających. Przyspieszenie układu powoduje powstanie dodatkowych sił inercjalnych, które umożliwiają nam odróżnienie „prawdziwego" ruchu od ruchu względnego. Fakt istnienia efektów inercjalnych w układach przyspieszających został wykorzystany przez Newtona w jego słynnym argumencie za absolutnością przestrzeni. Newton twierdził, że jedynym wyjaśnieniem dla pojawienia się siły odśrodkowej w przypadku ruchu obrotowego ciała jest hipoteza, że ciało to obraca się naprawdę względem absolutnie nieru-

[6] Warto jednak pamiętać, że jeśli zależy nam na dużej dokładności, jak w wypadku dokładnego określania położenia obiektów na powierzchni Ziemi za pomocą satelitarnych systemów GPS, musimy wziąć pod uwagę poprawki wynikające z ogólnej teorii względności.

chomej przestrzeni. Argument Newtona został skontrowany przez Macha, który wytknął słynnemu uczonemu pominięcie roli pozostałej części wszechświata. W przypadku względnego obrotu wiadra w stosunku do wody, o tym, czy pojawi się siła odśrodkowa zmieniająca jej kształt decyduje to, czy woda obraca się w odniesieniu do reszty wszechświata, czy nie. Mach uważał, że każdy ruch jest względny, niezależnie, czy jest on przyspieszony, czy nie. Pojedyncze ciało znajdujące się w pustym wszechświecie nie będzie doznawało działania żadnych sił inercjalnych, gdyż nie ma w nim innych ciał, względem których mogłoby przyspieszać.

Jak wygląda spór między Newtonem a Machem na gruncie ogólnej teorii względności? Na pierwszy rzut oka wydawać się może, że OTW jest bliższa stanowisku Macha. Wynika to z przyjętej przez Einsteina zasady równoważności, która zrównuje lokalnie układy przyspieszające z układami spoczywającymi w polu grawitacyjnym. Można więc sugerować, że uczestniczenie w ruchu przyspieszonym nie jest absolutną własnością danego ciała, gdyż równie dobrze można powiedzieć, że to grawitacja wywołuje obserwowany efekt w postaci sił inercjalnych (pozornych). Jednakże sprawa nie jest prosta. Równoważność Einsteinowska dotyczy dwóch różnych fizycznie sytuacji, a nie dwóch opisów tej samej sytuacji. W jednej sytuacji mamy przyspieszające ciało bez pola grawitacyjnego, a w drugim to samo ciało w polu grawitacyjnym. Tylko że pole grawitacyjne musi mieć jakieś źródło. Zatem w drugim przypadku muszą istnieć jakieś dodatkowe ciała wytwarzające odpowiednie pole grawitacyjne. Teoretycznie możemy więc odróżnić obie sytuacje przez dokładne badanie otoczenia testowego ciała. Jeśli wykluczymy istnienie zewnętrznych ciał wytwarzających pole grawitacyjne, będziemy zmuszeni przyznać, że nasze badane ciało naprawdę przyspiesza.

Argument ten można przedstawić bardziej wyraziście, rozważając hipotetyczny scenariusz samotnego ciała próbnego we wszechświecie. W sytuacji braku źródeł pola grawitacyjnego najbardziej naturalnym (choć nie jedynym) rozwiązaniem równań Einsteina jest płaska czasoprzestrzeń Minkowskiego szczególnej teorii względności. Jednakże, jak wiemy, w STW istnieje pojęcie absolutnego przyspieszenia oraz związanych z nim sił inercjalnych. Zatem ogólna teoria względności, której szczególnym przypadkiem jest STW, dopuszcza występowanie sił pozornych umożliwiających nam określenie, czy układ naprawdę przyspiesza (np. obraca się), nawet jeśli nie jesteśmy w stanie takiego przyspieszonego ruchu odnieść do zewnętrznych obiektów.

Zagadnienie samotnych rotujących obiektów w OTW rozważył ogólnie Kurt Gödel, słynny logik i matematyk, którego długie dyskusje z Einsteinem w Institute for Advanced Studies w Princeton przeszły do legendy. Gödel udowodnił, że istnieje rozwiązanie równań Einsteina opisujące pojedynczy dysk, w którym linie światła wysyłanego pionowo w stosunku do dysku będą zakreślały spiralny kształt. Najbardziej naturalną interpretacją tego rezultatu jest, że przedstawia on obracający się dysk, z którego perspektywy linie światła, naprawdę będące liniami prostymi, wyglądają spiralnie. Zatem OTW dopuszcza istnienie absolutnie obracających się obiektów w pustej przestrzeni! Mach byłby bardzo niezadowolony z takiego (*nomen omen*) obrotu sprawy. Oczywiście pozostaje otwartą kwestią, czy rozwiązanie Gödla ma sens fizyczny. Jednakże sam fakt dopuszczenia takiego rozwiązania jest znaczący. Wielu fizyków usiłowało poprawić ogólną teorię względności w duchu Macha i relatywizmu, tak aby wykluczyć z niej wszelkie elementy ruchów absolutnych. Jedną z takich ostatnich prób jest tzw. dynamika kształtu, zaproponowana przez brytyjskiego fizyka Juliana Barboura. Za wcześnie jednak ocenić, czy podejście to spełni pokładane w nim nadzieje.

Przywołajmy teraz inny słynny filozoficzny spór między dwoma wybitnymi umysłami epoki: Newtonem i Leibnizem. Jak pamiętamy z rozdziału drugiego, przedmiotem tego sporu był ontologiczny status czasu i przestrzeni. Leibniz bronił poglądu o ontologicznej zależności czasu i przestrzeni od znajdujących się w nich obiektów fizycznych, podczas gdy Newton stanowczo twierdził, że czas i przestrzeń są substancjami zdolnymi do samodzielnego istnienia. W pewnym sensie Leibniz może być uznanym za zwycięzcę tego sporu, jako że na gruncie szczególnej teorii względności ani czas, ani przestrzeń nie mają charakteru absolutnego – są zależne od układów odniesienia, a zatem także od wyznaczających te układy obiektów fizycznych. Jednak spór między Newtonem a Leibnizem może być przedstawiony w nowej formie jako dotyczący statusu całej czasoprzestrzeni, a nie jej podziału na czas i przestrzeń. Ponieważ oczywiście w STW czasoprzestrzeń wraz z jej geometrią Minkowskiego jest niezmiennicza względem transformacji układowych, pozostaje otwarte pytanie, czy czasoprzestrzeń jest niezależną substancją, czy też jej istnienie zależy od przedmiotów i łączących je relacji, jak chciał tego Leibniz.

Na gruncie szczególnej teorii względności wydaje się, że Leibniz ma duże szanse na odniesienie sukcesu. Geometria Minkowskiego jest w pewnym sensie nieciekawym „tłem" – wygląda tak samo w każdej części wszechświata, nie podlega zmianom ani oddziaływaniom, to po prostu wygodny sposób na opisanie tego, co dzieje się z przedmiotami fizycznymi, takimi jak pręty miernicze czy zegary. Pamiętamy z poprzedniego rozdziału, jak podstawowe efekty relatywistyczne – skrócenie długości czy dylatacja czasu – zostały wyjaśnione za pomocą odpowiednich procedur pomiarowych. Wielu filozofów przyjmuje wobec teorii czasoprzestrzeni obecnej w STW postawę skrajnie empirystyczną czy też wręcz instrumentalistyczną. Traktują tę teorię jako syntetyczne ujęcie procedur pomiarowych, wykorzystujących promienie świetlne do określenia m.in. równoczesności zdarzeń. W takim ujęciu czasoprzestrzeń nie wydaje się głównym przedmiotem badań, a jedynie odzwierciedleniem pomiarów na obiektach fizycznych.

Ogólną teorię względności określa się często, choć nieco myląco, mianem „teorii niezależnej od tła" (*background independent*) – bardziej adekwatne byłoby nazwać ją „pozbawioną tła" (*background free*). Chodzi o to, że we wszystkich poprzednich teoriach, takich jak mechanika klasyczna czy STW, czasoprzestrzeń jest biernym tłem, umożliwiającym lokalizację obiektów i zdarzeń, lecz nieuczestniczącym w fizycznych oddziaływaniach. Natomiast na gruncie OTW czasoprzestrzeń przejmuje aktywną rolę, gdyż oddziałuje na ciała fizyczne, wyznaczając ich ruch po zakrzywionych geodezyjnych, i na nią samą oddziałują obiekty materialne obdarzone masą. Jednakże niektórzy filozofowie mają wątpliwość, czy nie jest to nadużycie pojęcia oddziaływania przyczynowego. Ciała obdarzone masą nie wpływają dosłownie na czasoprzestrzeń w sensie oddziaływania fizycznego; po prostu istnieje zależność między masą, energią i pędem a krzywizną. Podobnie poruszanie się po geodezyjnych nie jest formą „popychania", lecz syntetycznym ujęciem pewnej prawidłowości. Nie powiemy przecież, że pusta przestrzeń newtonowska „popycha" ciała swobodne, aby kontynuowały one jednostajny i prostoliniowy ruch.

Sytuacja ta ulega radykalnej zmianie na gruncie ogólnej teorii względności. Po pierwsze, geometria czasoprzestrzeni staje się tutaj istotnym aktorem rozważanych procesów. Jej własności mogą ulegać modyfikacjom. Czasoprzestrzeń podlega przeróżnym deformacjom, wyrażalnym w postaci krzywizny, która może przybrać ekstremalną postać w pobliżu osobliwości takich jak czarne dziury. Należy dodać, że deformacje takie mają istotny fizyczny wpływ

na zachowanie ciał. Podstawowa zasada dynamiki OTW, zgodnie z którą swobodne ciała fizyczne podążają wzdłuż linii geodezyjnych, w naturalny sposób daje się zinterpretować w kategoriach niemal fizycznego oddziaływania czy wpływu czasoprzestrzeni na znajdujące się w niej obiekty. Spektakularnym przykładem mogą być tutaj fale grawitacyjne, czyli rozchodzące się zaburzenia czasoprzestrzeni, niosące ze sobą energię. Odpowiednio silne fale grawitacyjne mogłyby np. strącić słynną skałę na Półwyspie Gibraltarskim. W świetle takich faktów trudno odmówić czasoprzestrzeni pełnej realności fizycznej. Syntetyczne ujęcie relacji między przedmiotami, czym w istocie była przestrzeń dla Leibniza, nie byłoby raczej w stanie zniszczyć fizycznego obiektu w rodzaju Skały Gibraltarskiej.

Jednak zwolennicy relacjonizmu nie znajdują się na straconej pozycji. Mają asa w rękawie, jakim jest argument Leibniza z przesunięcia. Argument ten musi być oczywiście dostosowany do matematyki i fizyki OTW. W oryginalnej postaci rozumowanie Leibniza przeprowadzone było przy założeniu Galileuszowskiej płaskiej czasoprzestrzeni, rozbitej na momentalną przestrzeń i absolutny układowo czas. Obecnie przedstawimy jego zmodyfikowaną wersję, znaną jako argument dziury (*hole argument*). Zacznijmy od przypomnienia podstawowej własności OTW, jaką jest jej tzw. ogólna kowariantność. Za tą trudną nazwą kryje się dość jasna idea, jawnie sformułowana przez Einsteina: równania teorii powinny obowiązywać jednakowo w każdym układzie współrzędnych, niezależnie od tego, czy jest on prostoliniowy, krzywoliniowy, czy porusza się względem innych układów itd. Formalnie taka niezależność wyraża się w uogólnionym sposobie przechodzenia z jednego układu do drugiego, zawartym w tzw. dyfeomorfizmach. Dyfeomorfizm to każde przekształcenie pewnej przestrzeni, które nie zmienia relację bliskości między punktami (jak wspomnieliśmy, takie relacje określa się mianem topologii). Możemy więc np. rozważyć przekształcenie, które rozciąga daną przestrzeń, deformując jej kształt, ale nie takie, które ją rozrywa.

Formalnie dyfeomorfizmy są gładkimi funkcjami z N-wymiarowych przestrzeni liczb rzeczywistych w takie same przestrzenie. Przestrzenie te reprezentują współrzędne punktów naszej wyjściowej przestrzeni. (W wypadku czasoprzestrzeni wymiar N będzie oczywiście równy 4.) Zatem dany dyfeomorfizm może być po prostu interpretowany jako zamiana danych współrzędnych punktów czasoprzestrzeni na inne współrzędne. Jest to tzw. interpretacja pasywna transformacji. Nie zmienia ona niczego w fizycznej sytuacji, a jedynie zastępuje jeden opis innym, równie dobrym.

Okazuje się jednak, że każdej transformacji pasywnej odpowiada pewna transformacja aktywna, dokonująca modyfikacji w samej czasoprzestrzeni. Rozważmy prosty przykład ilustrujący taką zmianę. Niech będą dane dwa punkty czasoprzestrzeni A i B, których współrzędne w pewnym układzie wynoszą (1, 0, 0, 0) i (2, 1, 0, 0). Rozważmy dyfeomorfizm, który przekształca współrzędne punktu A na współrzędne (2, 1, 0, 0), zmieniając jednocześnie współrzędne B na jakieś inne (nie jest istotne, na jakie).[7] Jednakże takiej pasywnej transformacji odpowiada transformacja aktywna, która nie zmienia współrzędnych, tylko same punkty, tzn. punkt A przesuwamy w miejsce punktu B. Oczywiście na „miejsce" punktu A (w sensie odpowiednich współrzędnych) wejdzie jakiś inny punkt, a punkt B przesunie się gdzie indziej.[8] Rezultatem będzie czasoprzestrzeń, która wygląda tak samo jak oryginalna ze zmienionymi współrzędnym, ale która ontologicznie jest już inna.

[7] Taki dyfeomorfizm może polegać np. na dodaniu 1 do współrzędnej czasowej t i do współrzędnej przestrzennej x: $d(t, x, y, z) = (t + 1, x + 1, y, z)$.

[8] Można wysunąć zarzut – całkiem zresztą słuszny – że aktywna interpretacja transformacji podana wyżej jest oparta na błędnym przekonaniu o istnieniu „miejsc", w których znajdują się punkty

Z punktu widzenia Leibniza taka transformacja aktywna nie ma większego sensu, skoro rezultat jest nieodróżnialny od stanu wyjściowego. W zasadzie odtworzyliśmy więc stary dobry argument z przesunięcia, w którym założenie substancjalizmu prowadziło do przyjęcia istnienia dwóch (w istocie nieskończenie wielu) ontologicznie różnych, lecz empirycznie nieodróżnialnych sytuacji. Jednak w wypadku OTW argument z przesunięcia posiada dodatkowy element. Przesunięcie leibnizjańskie musiało być jednorodne – we wszystkich punktach wszechświata przesuwamy materię o tę samą odległość. Natomiast przekształcenia dyfeomorficzne w OTW są bardziej elastyczne – jedne obszary mogą być „rozciągane" bardziej, inne mniej. W szczególności istnieje transformacja dyfeomorficzna czasoprzestrzeni, która przesuwa punkty tylko w pewnym wyróżnionym obszarze (w „dziurze"), pozostawiając resztę czasoprzestrzeni niezmienioną. Mamy zatem do czynienia z pewną szczególną formą indeterminizmu. Istnieją dwie możliwe ewolucje wszechświata, które wyglądają tak samo w pewnym okresie czasu, ale które różnią się w obszarze dziury. Różnica polega na tym, że te same punkty otrzymują inne własności fizyczne i geometryczne. Na przykład punkt, który w jednym „świecie" znajdował się w wnętrzu gwiazdy neutronowej, w innym możliwym świecie będzie pozbawiony wszelkiej materii. Zatem jego własności metryczne będą inne – w pierwszym wypadku będzie on charakteryzował się tensorem metrycznym (i tensorem krzywizny) znacznie odbiegającym od tensora płaskiej czasoprzestrzeni Minkowskiego, a w drugim będzie miał tensor metryczny praktycznie równy tensorowi Minkowskiego.

Oczywiście opisane wyżej różnice są niemożliwe do wykrycia, gdyż nie możemy „zaznaczyć" poszczególnych punktów markerem, aby następnie sprawdzić, jakie mają one własności w odpowiednich światach. Indeterminizm dziury nie prowadzi do żadnych konsekwencji empirycznych. Mimo to dla substancjalisty zmiana wewnątrz dziury jest realna, choć nieobserwowalna. Z kolei leibnizjański relacjonista powie, że punkty czasoprzestrzenne są wtórne względem koincydencji własności fizycznych. Nie ma sensu mówić, że punkt we wnętrzu gwiazdy neutronowej mógłby być tym samym co punkt w pustej przestrzeni. W każdym możliwym świecie z takim samym rozkładem materii jest jeden i ten sam punkt: ten, który jest wyznaczony określonymi warunkami fizycznymi panującymi we wnętrzu gwiazdy. Zatem wszystkie możliwości odpowiadające różnym transformacjom dyfeomorficznym sprowadzają się do jednej i tej samej możliwości.

Czy jednak substancjalista jest skazany na niedookreślenie sytuacji ontologicznej wewnątrz dziury? To zależy, jaki wariant substancjalizmu przyjmiemy. Argument dziury poddaje krytyce najbardziej radykalną wersję substancjalizmu, zgodnie z którą każdy punkt ma pewną wewnętrzną „tożsamość" czy też indywidualność, niezależną od wszelkich posiadanych własności. W języku metafizyki taką bezjakościową indywidualność określa się mianem *haecceitas*, od łacińskiego *haecce* czyli „ta oto" (wspomnieliśmy o tym pojęciu w ramce na s. 59). Posiadanie *haecceitas* pozwala na mówienie o różnych alternatywnych własnościach tego samego obiektu, np. o różnych możliwych lokalizacjach tego samego punktu względem obiektów materialnych, takich jak nasza przykładowa gwiazda neutronowa. Jednakże są filozofowie, którzy kwestionują zasadność wprowadzania tak rozumianej tożsamo-

---

i względem których te punkty mogą być przesuwane. Jednakże to same punkty są miejscami – pytanie o umiejscowienie miejsc jest pozbawione sensu. Ściśle rzecz biorąc, aktywna wersja transformacji dokonuje się nie względem miejsc, a względem pól fizycznych czy rozkładu materii. Jeśli np. punkt A w oryginalnej sytuacji był zajęty przez jakiś obiekt materialny – np. gwiazdę – a punkt B był pusty, to po transformacji punkt A będzie pusty, a B trafi gdzie indziej, być może do jakiegoś innego obiektu czy pola fizycznego.

ści. W alternatywnym ujęciu przedmioty charakteryzują się tzw. własnościami istotnościowymi, które determinują tożsamość każdego obiektu. Własność istotnościowa (zwana również esencjalną) to taka, bez której przedmiot nie może istnieć. Filozof fizyki Tim Maudlin zasugerował, aby za własności istotnościowe punktów czasoprzestrzennych uznać ich własności metryczne wyrażane tensorem metrycznym. Przy takim założeniu sytuacja, w której punkt znajdujący się wewnątrz gwiazdy neutronowej znalazłby się w pustym obszarze, jest metafizycznie niemożliwa, gdyż straciłby on wtedy swoją własność esencjalną w postaci silnie „wykrzywionego" tensora metrycznego. Zatem aktywna transformacja dyfeomorficzna, choć matematycznie dopuszczalna, jest wykluczona na mocy założenia metafizycznego.

Jak się zatem wydaje, substancjalizm esencjalistyczny (zwany również „wyrafinowanym" – *sophisticated*) jest w stanie odeprzeć argument dziury, w przeciwieństwie do substancjalizmu haecceistycznego. Oczywiście jak to zwykle bywa w filozofii, pojawiają się dodatkowe trudności, kontrargumenty i próby obrony. Jednym z problemów koncepcji Maudlina jest to, że dopuszcza ona inną formę indeterminizmu. Chociaż nasz wybrany punkt we wnętrzu gwiazdy neutronowej nie może przyjąć innych własności metrycznych, to jednak nie jest wykluczone istnienie alternatywnego świata możliwego, w którym we wnętrzu gwiazdy znalazłby się zupełnie nowy, nieistniejący realnie punkt o tych samych własnościach (w literaturze takie obiekty nazywa się „kretami", które podszywają się pod prawowite przedmioty). Rozwiązaniem tego problemu może być np. przyjęcie, że własności istotnościowe są zarówno konieczne, jak i wystarczające do bycia danym przedmiotem. Analiza konsekwencji takiego stanowiska wykracza jednak poza ramy niniejszego opracowania. Czytelników zainteresowanych głębszą metafizyką argumentu dziury odsyłam do literatury zamieszczonej na końcu rozdziału, my natomiast przejdziemy z abstrakcyjnych rozważań filozoficznych do nie mniej abstrakcyjnej matematyki.

6.6.* Podstawy geometrii różniczkowej

6.6.1. Wektory i tensory

Ogólna teoria względności w swojej warstwie formalnej opiera się na matematycznej teorii zwanej geometrią różniczkową. Z pewnymi pojęciami tej teorii zetknęliśmy się już w rozdziałach poświęconych teorii elektromagnetyzmu i szczególnej teorii względności, natomiast teraz przyjrzymy się nieco dokładniej i w bardziej uporządkowany sposób jej podstawowym pojęciom wykorzystywanym w fizyce. Geometria różniczkowa stosuje metody analizy matematycznej do opisu różnych własności geometrycznych przestrzeni takich jak np. krzywizna. Aby było możliwe zastosowanie analizy matematycznej, musimy umieć wyrazić własności geometryczne w języku funkcji na liczbach rzeczywistych. Zapewne większość czytelników jest zaznajomiona z geometrią analityczną, w której za pomocą kartezjańskiego układu współrzędnych możemy przypisać punktom ich liczbowe współrzędne, a następnie przy ich pomocy algebraicznie przedstawić różne geometryczne relacje, jak np. w równaniu prostej czy okręgu. Jednakże geometria analityczna jest geometrią Euklidesową, natomiast my potrzebujemy bardziej elastycznej (*omen nomen*) geometrii. Punktem wyjścia takiej geometrii jest pojęcie rozmaitości różniczkowalnej, w skrócie – rozmaitości.

Zacznijmy od wprowadzenia pojęcia mapy na pewnym zbiorze punktów M. Ogólnie mapą nazwiemy każde jedno-jednoznaczne odwzorowanie φ, które przekształca pewien pod-

zbiór $A \subseteq M$ na podzbiór (otwarty[9]) N-wymiarowej przestrzeni R^N (iloczyn kartezjański N zbiorów liczb rzeczywistych). Idea jest bardzo prosta – chodzi o to, aby mapa przypisywała każdemu punktowi w A jego unikalne współrzędne w postaci N-tki liczb rzeczywistych. Oczywiście w przypadku teorii czasoprzestrzeni wymiar N będzie równy 4, ale geometria różniczkowa operuje przestrzeniami arbitralnie wielo-wymiarowymi. Zbiór map C, które są ze sobą w pewnym sensie zgodne i które pokrywają łącznie całą przestrzeń M, nazywamy atlasem. Rozmaitość różniczkowalna będzie to więc zbiór M razem z jego atlasem: (M, C).[10] Wraz z wprowadzeniem atlasu w postaci szeregu map możemy przedstawiać liczbowe funkcje zdefiniowane na zbiorze punktów M jako funkcje z R^N w R, do których stosują się standardowe pojęcia z analizy matematycznej: ciągłość, różniczkowalność w stopniu d (istnienie pochodnych do stopnia d). Na przykład dowolna funkcja α z M w R jest „gładka", gdy dla każdej mapy φ z atlasu C, funkcja $\alpha \circ \varphi^{-1}$, której dziedziną jest podzbiór R^N, a przeciwdziedziną R, jest nieskończenie wiele razy różniczkowalna.

Niezmiernie ważnym pojęciem, umożliwiającym wprowadzenie całego szeregu matematycznych charakterystyk w odniesieniu do danej rozmaitości, jest pojęcie przestrzeni stycznej do M w danym punkcie p. Poglądowo możemy wyobrazić sobie przestrzeń styczną jako kartkę papieru przyklejoną np. do sfery w pewnym punkcie. Przestrzeń styczna jest „areną", na której działają takie obiekty jak wektory i tensory. Istnieje wiele równoważnych sposobów zdefiniowania przestrzeni stycznej. Najbardziej elegancką, choć nieco abstrakcyjną, jest następująca definicja. Niech $S(p)$ będzie zbiorem wszystkich gładkich funkcji zdefiniowanych w otoczeniu p. Na tym zbiorze rozważymy wszystkie funkcje (w istocie funkcjonały – por. paragraf 2.7) ξ w liczby rzeczywiste, które spełniają formalne warunki pochodnych (f_1 i f_2 są dowolnymi funkcjami z $S(p)$):

$$\xi(f_1 + f_2) = \xi(f_1) + \xi(f_2),$$

$$\xi(f_1 f_2) = f_1 \xi(f_2) + f_2 \xi(f_1),$$

Jeśli f = const, to $\xi(f) = 0$.

Być może pamiętacie, że pochodna z funkcji w punkcie spełnia wszystkie trzy powyższe warunki: pochodna z sumy funkcji jest sumą pochodnych, pochodna iloczynu dana jest powyższym wzorem, a pochodna z funkcji stałej jest oczywiście równa zeru. Każdy funkcjonał ξ spełniający dane warunki nazwiemy wektorem stycznym w punkcie p. Związek pochodnych ze stycznymi jest znany już ze szkolnego kursu matematyki (por. wpis w ramce na

[9] Zbiór otwarty to zbiór pozbawiony brzegu – np. wnętrze koła o promieniu r bez punktów na okręgu.

[10] Można zadać pytanie, dlaczego musimy rozważać więcej niż jedną mapę – czy nie wystarczy po prostu jedno odwzorowanie φ przypisujące każdemu punktowi w M jego współrzędne? Okazuje się, że w pewnych wypadkach jest to niemożliwe. Na przykład z geografii znany jest fakt, że nie istnieje jedna płaska mapa, która przedstawiałaby cały glob ziemski. Potrzebujemy co najmniej dwóch takich map – jedną np. zawierającą biegun północny, bez bieguna południowego, a drugą na odwrót. Warunek zgodności dwóch map dotyczy ich zachowania w obszarach „przecięcia" ich dziedzin (gdzie na siebie nachodzą). Najprościej byłoby przyjąć, że dwie mapy φ_1: $A \to \mathrm{R}^n$ i φ_2: $B \to \mathrm{R}^n$ są zgodne, gdy w obszarze $A \cap B$ są identyczne. To jednak za silny warunek, bo konkretne wartości współrzędnych mogą ulec zmianie przy zmianie mapy. Przyjmuje się, że dwie mapy są zgodne, gdy punkty uznane za „bliskie" przez jedną mapę będą bliskie w drugiej mapie. Dokładniej warunek ten stwierdza, że złożenia odwzorowań $\varphi_1 \circ \varphi_2^{-1}$ oraz $\varphi_2 \circ \varphi_1^{-1}$, które są funkcjami z podzbiorów R^n w R^n, są wystarczająco „gładkie" w danych dziedzinach (są różniczkowalne odpowiednią liczbę razy).

s. 43). W obecnie rozważanym abstrakcyjnym podejściu istnieje nieskończenie wiele operacji „pochodnych" dla danego punktu *p*, z których każda taka operacja wyznacza pewien abstrakcyjny kierunek styczny do danej rozmaitości. Ponadto zbiór wszystkich takich operacji (wektorów stycznych) tworzy przestrzeń wektorową. Oznacza to, że suma dwóch wektorów stycznych jest również wektorem stycznym (spełnia powyższe warunki) oraz iloczyn wektora stycznego przez liczbę jest wektorem stycznym. Przestrzeń wszystkich wektorów stycznych w punkcie *p* nazywamy właśnie przestrzenią styczną i oznaczamy jako T_p.[11]

Dodajmy, że wektory styczne należą do kategorii obiektów zwanych kontrawariantnymi. Istnieją również inne typy wektorów, określanych mianem kowariantnych (nazewnictwo to nie ma jakiegoś szczególnego uzasadnienia i jest pewną historyczną zaszłością). Aby je wprowadzić, rozważmy nową przestrzeń T_p^*, zwaną przestrzenią dualną do przestrzeni stycznej T_p. Przestrzeń dualna do danej przestrzeni wektorowej jest ogólnie definiowana jako przestrzeń złożona z wszystkich funkcjonałów liniowych operujących na danych wektorach. Funkcjonały liniowe tworzą przestrzeń wektorową (suma funkcjonałów jest funkcjonałem, podobnie iloczyn funkcjonału przez liczbę). Znowu zatem mamy do czynienia z sytuacją, gdzie identyfikujemy jako wektory obiekty matematyczne, które pozornie mają niewiele wspólnego z geometrią (są to funkcje przypisujące liczby rzeczywiste odpowiednim wektorom z T_p).

Przestrzenie T_p i T_p^* umożliwiają formalne wprowadzenie kolejnej niezmiernie ważnej kategorii obiektów, czyli tensorów. Tensorem kowariantnym nazwiemy każdą liniową funkcję, która *n* wektorom w przestrzeni stycznej T_p przypisuje liczbę. Innymi słowy, tensor kowariantny jest liniową funkcją (funkcjonałem) z iloczynu kartezjańskiego *n* przestrzeni $T_p \times T_p \times \ldots \times T_p$ w liczby rzeczywiste R. Liczba *n* jest nazywana rzędem (*rank*) tensora. Warto zauważyć, że tensor kowariantny rzędu 1 to po prostu wektor kowariantny. Z kolei tensory kontrawariantne to funkcje działające na iloczynach kartezjańskich przestrzeni dualnych: $T_p^* \times T_p^* \times \ldots \times T_p^*$. Można zatem w skrócie powiedzieć, że tensory i wektory kontrawariantne „mieszkają" w przestrzeni stycznej i działają na elementach przestrzeni dualnej,[12] a tensory i wektory kowariantne zamieszkują przestrzeń dualną i działają na elementach przestrzeni stycznej.

[11] Mam nadzieję, że jako filozofowie docenicie subtelne piękno tej konstrukcji, która na pierwszy rzut oka może wydawać się aberracją. Definiujemy tutaj czysto geometryczne pojęcie wektora przy pomocy zaczerpniętego z analizy matematycznej (liczb rzeczywistych) pojęcia funkcjonału zachowującego się jak pochodna. Może to budzić sprzeciw: w szkole jesteśmy przyzwyczajeni patrzeć na obiekty matematyczne typu wektor czy funkcja jako na odrębne rodzaje bytów. Wektor nie jest przecież funkcją, tylko strzałką o określonym kierunku, zwrocie i długości (jeśli dobrze zapamiętałem szkolną definicję). Okazuje się jednak, że w wyższej matematyce pozornie odrębne klasy obiektów często się przecinają. Wektor nie jest po prostu strzałką, ale przedmiotem „zachowującym" się w określony sposób, a dokładniej wchodzącym w określone relacje z innym wektorami. Wektorami mogą być funkcje, funkcjonały czy jeszcze inne abstrakcyjne obiekty.

[12] Wytłumaczenie, dlaczego wektory kontrawariantne mogą być interpretowane jako funkcje (funkcjonały) na przestrzeni dualnej T_p^*, jest trochę skomplikowane. Bierze się to stąd, że przestrzeń dualna do przestrzeni dualnej, czyli T_p^{**}, jest w pewnym sensie „identyczna" z przestrzenią wyjściową T_p (dokładniej, w matematyce mówimy o izomorficzności między tymi przestrzeniami). Zatem każdy element przestrzeni T_p, który jak wiemy jest wektorem kontrawariantnym, może być interpretowany jako pewien element T_p^{**}, czyli funkcjonał linowy na T_p^*. Zostawiam najodważniejszym czytelnikom do rozważenia, jak może wyglądać relacja „tożsamości" między T_p a T_p^{**}.

Wprowadzone przez nas pojęcia wektorów i tensorów kowariantnych i kontrawariantnych są dość abstrakcyjne (jest to metoda preferowana przez matematyków). Fizycy jednak wolą myśleć o tych obiektach w nieco bardziej „przyziemny" sposób, przez odwołanie się do pewnego konkretnego układu współrzędnych. Niech $x^1(p), x^2(p), \ldots, x^N(p)$ będą współrzędnymi punktu p odpowiadającymi pewnej mapie. Rozważmy N odwzorowań z R^N w R, z których każde przypisuje każdemu ciągowi współrzędnych (x^1, x^2, …, x^N) jedną wybraną współrzędną x^i (będziemy oznaczać tę funkcję takim samym symbolem x^i). Weźmy dowolny wektor styczny ξ z przestrzeni T_p, zdefiniowany jak wyżej. Wektor ten, jako funkcjonał liniowy, przypisuje każdej funkcji x^i pewną liczbę ξ^i. Ciąg liczb (ξ^1, ξ^2, …, ξ^N) nazwiemy składowymi wektora ξ w układzie współrzędnych x^1, x^2, …, x^N. Jak widać, przechodzimy teraz z abstrakcyjnego ujęcia wektorów stycznych jako pewnych funkcjonałów do „konkretnej" i dobrze znanej charakterystyki w postaci N-tek liczb. Można udowodnić (nie będziemy tego robić), że działanie wektora-funkcjonału ξ na dowolnej funkcji f z $S(p)$ wyglądać będzie tak, jak poniżej. Zatem rzeczywiście wektor ξ działa jak pochodna (dodajmy, pochodna w kierunku wektora o składowych ξ^i).

$$\xi(f) = \sum_{i=1}^{N} \xi^i \frac{\partial f}{\partial x^i}. \tag{6.3}$$

Należy podkreślić, że chociaż w każdym układzie współrzędnych wektory styczne reprezentowane są przez ciągi liczb, to jednak przy przejściu z jednego układu do drugiego liczby te będą ulegać zmianie. Wektory styczne (kontrawariantne) transformują się według ściśle określonych reguł. Regułę transformacji można wyprowadzić z wzoru (6.3). Niech $y^1, y^2, \ldots y^N$ będą współrzędnymi w nowym układzie odniesienia. Współrzędne te będą oczywiście powiązane pewnymi zależnościami funkcyjnymi ze „starymi" współrzędnymi x^i. Zgodnie z regułą łańcuchową, pochodną funkcji f po współrzędnej x^i można wyrazić jako iloczyn pochodnej po nowych współrzędnych razy pochodna nowej współrzędnej po starej współrzędnej:

$$\frac{\partial f}{\partial x^i} = \frac{\partial f}{\partial y^j} \frac{\partial y^j}{\partial x^i}.$$

W powyższej formule zastosowaliśmy powszechnie wykorzystywaną konwencję (tzw. konwencję sumacyjną Einsteina), zgodnie z którą indeksy powtarzające się (w naszym wypadku j) należy posumować po wszystkich wartościach. Upraszcza to zapis odpowiednich formuł (bez tej konwencji musielibyśmy napisać prawą stronę równania jako $\sum_j \frac{\partial f}{\partial y^j} \frac{\partial y^j}{\partial x^i}$).

Stosując powyższą formułę, łatwo się przekonamy, że wzór na transformacje składowych wektora przy przejściu z jednego układu x do innego y ma postać:

$$\xi^i_{(y)} = \frac{\partial y^i}{\partial x^j} \xi^j_{(x)}, \tag{6.4}$$

gdzie indeksy dolne (x) i (y) oznaczają relatywizację do danego układu współrzędnych. W podręcznikach do ogólnej teorii względności można często spotkać sformułowanie, że wektor kontrawariantny definiujemy po prostu jako obiekt, który transformuje się w powyżej podany sposób. Z kolei wektory kowariantne, które jak pamiętamy, należą do przestrzeni dualnej T_p^*, mają inny sposób transformacji. Zacznijmy od pytania, jak wyglądają składowe

wektora kowariantnego (dualnego) w danym układzie współrzędnych. Aby je obliczyć, działamy danym wektorem (który jest funkcjonałem liniowym na zbiorze wektorów kontrawariantnych) na wektory o postaci (1, 0, …, 0), (0, 1, …, 0) itd. (Takie wektory nazywamy bazowymi, gdyż przy ich pomocy możemy wyrazić każdy inny wektor.) Liczby uzyskane w wyniku takich operacji będą właśnie składowymi danego wektora kowariantnego – oznaczamy je symbolami z indeksem dolnym: ξ_i. Okazuje się, że jeśli zastosujemy regułę transformacji (6.4) dla wektorów kontrawariantnych do definicji wektora kowariantnego, odpowiednia reguła transformacji przyjmie postać:

$$\xi_i^{(y)} = \frac{\partial x^j}{\partial y^i}\,\xi_j^{(x)}.$$

Jest to w pewnym sensie „odwrotność" reguły (6.4). Przechodząc z układu x do y, musimy obliczyć, jak szybko zmieniają się współrzędne „stare" względem nowych, a nie nowe względem starych, jak to miało miejsce w przypadku wektorów kontrawariantnych.

Podobne reguły transformacji stosują się do tensorów, zarówno kowariantnych, jak i kontrawariantnych. Do określenia tensora w danym układzie współrzędnych nie wystarcza N liczb – potrzebna jest do tego cała tablica (macierz). Na przykład tensor kowariantny drugiego rzędu jest, jak pamiętamy, funkcją liniową przypisującą parom wektorów stycznych liczby. Żeby zatem określić sposób działania tego tensora na wektorach bazowych (1, 0, …, 0), (0, 1, …, 0) itd.., potrzebujemy $N \times N$ liczb (każda liczba daje rezultat dla dwóch wektorów bazowych). Aby zaznaczyć, że mamy do czynienia z tensorem 2. rzędu, stosujemy dwa indeksy, np. T_{ij}. Ponadto przyjmujemy konwencję, że tensor z dolnymi indeksami oznacza tensor kowariantny („zamieszkujący" przestrzeń dualną i działający na elementach przestrzeni stycznej), a górne indeksy oznaczają tensor kontrawariantny (działający na elementach przestrzeni dualnej). Podobnie jak w wypadku wektorów, tensory kowariantne i kontrawariantne różnią się sposobem transformacji przy przejściu z jednego układu współrzędnych do drugiego. Oto odpowiednie reguły dla tensorów kowariantnych i kontrawariantnych drugiego rzędu:

$$T_{nm}^{(y)} = \frac{\partial x^r}{\partial y^n}\frac{\partial x^s}{\partial y^m}T_{rs}^{(x)}, \tag{6.6}$$

$$T_{(y)}^{nm} = \frac{\partial y^n}{\partial x^r}\frac{\partial y^m}{\partial x^s}T_{(x)}^{rs}. \tag{6.7}$$

Dodajmy, że możliwe jest (w istocie jest to niezbędne) rozważanie „mieszanych" tensorów, dla których część indeksów jest kowariantna, a część kontrawariantna. Oznacza to, że tensory takie działają na odpowiednich kombinacjach wektorów kowariantnych i kontrawariantnych. Przykładami takich tensorów mogą być: T_n^m (kowariantny w indeksie n i kontrawiariantny w indeksie m) lub T_a^{bc} (kontrawariantny w dwóch indeksach i kowariantny w jednym). Reguły transformacji dla takich tensorów będą zawierały odpowiednie kombinacje pochodnych cząstkowych jednych współrzędnych względem drugich.[13]

[13] Spróbujcie napisać te reguły dla tensorów T_n^m i T_a^{bc}.

6.6.2. Pochodna kowariantna, geodezyjne i krzywizna

Studiujący geometrię różniczkową uczą się metod operowania na tensorach – dodawania, mnożenia, a także bardziej „egzotycznych" operacji, jak zwężanie (zmniejszanie liczby niezależnych indeksów) czy opuszczanie bądź podnoszenie indeksów. Nie będziemy oczywiście w stanie omówić tych zagadnień, zainteresowanych odsyłając do literatury. Natomiast powinniśmy się przyjrzeć jednej ważnej operacji na tensorach czy wektorach – mianowicie operacji różniczkowania. Jest bardzo istotne, abyśmy umieli poprawnie zróżniczkować np. wektor wzdłuż danej krzywej. Jak wiadomo, pochodne reprezentują liczbowo to, jak zmienia się dana funkcja wraz ze zmianą parametru różniczkowania. W szczególności możemy przy pomocy pochodnej wyrazić fakt, że np. dany wektor nie zmienia się wzdłuż pewnej krzywej – jest tak wtedy, gdy jego pochodna wzdłuż tej krzywej jest równa zeru. Umożliwi to nam wprowadzenie pojęcia przeniesienia równoległego, które jak wiemy jest kluczowe w geometrii o niezerowej krzywiźnie. Sprawa może wydawać się prosta. Rozważmy dowolne pole wektorowe $V_m(p)$ na pewnym obszarze rozmaitości M, rozpisane na składowe w jakimś układzie współrzędnych (w pewnej mapie). Oznacza to, że $V_m(p)$ jest w istocie funkcją z R^N (współrzędne punktów) w R^N (składowe wektora), którą można w standardowy sposób różniczkować, a zatem możemy obliczyć wartości pochodnych cząstkowych $\frac{\partial V_m}{\partial x^n}$ w danym punkcie p. Utworzą nam one tablicę liczb numerowanych indeksami m, n. Wydaje się, że otrzymaliśmy dobrze znany obiekt – tensor drugiego rzędu.

Pojawia się jednak pewien problem. Zestaw liczb $\frac{\partial V_m}{\partial x^n}$ nie tworzy tensora, gdyż liczby te nie transformują się we właściwy sposób przy przejściu z jednego układu współrzędnych do drugiego. Można to sprawdzić bezpośrednim rachunkiem, stosując znane zasady różniczkowania, takie jak reguła różniczkowania iloczynu funkcji, i pamiętając, że przy zmianie układu z (x) do (y) musimy zmienić składowe wektora V_m, zgodnie z regułą (6.5). W rezultacie otrzymamy następującą „regułę transformacji":

$$\frac{\partial V_m^{(y)}}{\partial y^n} = \frac{\partial x^r}{\partial y^m}\frac{\partial V_r^{(x)}}{\partial y^n} + \frac{\partial^2 x^r}{\partial y^n \partial y^m} V_r^{(x)}.$$

Pierwszy człon sumy po prawej stronie powyższego równania reprezentuje transformację, jakiej należałoby się spodziewać, gdyby liczby $\frac{\partial V_m}{\partial x^n}$ były składowymi tensora. Widać to dobrze, jeśli przepiszemy ten człon przy pomocy reguły łańcuchowej jako $\frac{\partial x^r}{\partial y^m}\frac{\partial x^s}{\partial y^n}\frac{\partial V_r^{(x)}}{\partial x^s}$ – jest to dokładnie reguła transformacji tensora kowariantnego drugiego rzędu, zapisana w formule (6.6). Niestety, sprawę „psuje" obecność drugiego członu w sumie.

Rozwiązaniem jest przyjęcie nowej reguły różniczkowania, która będzie „kompensować" wystąpienie czynnika $\frac{\partial^2 x^r}{\partial y^n \partial y^m} V_r^{(x)}$. Zauważmy, że czynnik ten ma postać liczb o trzech indeksach n, m, i r, które „działają" na wektor $V_r^{(x)}$. Zatem nowa reguła różniczkowania, zwana pochodną kowariantną, powinna wyglądać następująco:

$$\nabla_n V_m = \frac{\partial V_m}{\partial y^n} + \Gamma^r_{mn} V_r. \tag{6.8}$$

Oczywiście należy znaleźć odpowiednią formę dla liczb Γ^r_{mn}, tak aby cała operacja tworzyła poprawnie transformujący się tensor. Liczby te noszą nazwę symboli Christoffela (nazywa

się je też często koneksją afiniczną). Należy podkreślić, że Γ^r_{mn} nie jest tensorem, co zrozumiałe, jako że liczby $\frac{\partial V_m}{\partial y^n}$ także nie tworzą współrzędnych tensora. Dopiero dodanie obu składników tworzy poprawny tensor.

Pochodna kowariantna ∇_n może być zastosowana do dowolnych tensorów, nie tylko wektorów. Jednym z najważniejszych zastosowań tej operacji jest definicja linii geodezyjnych. Idea jest wystarczająco przejrzysta: wektory styczne do linii geodezyjnej powinny być do siebie równoległe ze względu na pochodną kowariantną, czyli pochodna kowariantna z wektorów stycznych jest równa zeru. Musimy jeszcze tylko zdefiniować wektory styczne do danej linii. Nie jest to trudne zadanie. Przede wszystkim wyprecyzujmy, co należy rozumieć przez linię krzywą w rozmaitości *M*: jest to gładkie odwzorowanie z odcinka liczb rzeczywistych *I* w zbiór *M* (można sobie wyobrazić, że numerujemy punkty na danej linii liczbami rzeczywistymi). Niech $\gamma(s)$: $I \subseteq \mathrm{R} \rightarrow M$ będzie gładką funkcją definiującą pewną krzywą w *M*. Wektorem stycznym do krzywej γ w punkcie *p* (dla którego $\gamma(s_0) = p$) nazwiemy funkcjonał liniowy γ_p, taki że

$$\gamma_p(f) = \frac{d(f \circ \gamma)}{ds}(s_o),$$

dla dowolnej funkcji *f* ze zbioru $S(p)$ (zbioru gładkich funkcji z otoczenia punktu *p* w liczby rzeczywiste). Znowu mamy bliskie pokrewieństwo pojęcia styczności z pochodną – tym razem pochodną „wzdłuż" danej krzywej.[14]

Obliczmy składowe wektora stycznego do krzywej γ w danym punkcie *p*. Jak pamiętamy, składowe wektora w danym układzie współrzędnych (x) powstają w wyniku zastosowania tego wektora (który jest funkcjonałem) do pojedynczych współrzędnych (funkcji przypisujących całej *N*-tce współrzędnych jedną wybraną współrzędną x^i). Korzystając z powyższej definicji wektora-funkcjonału γ_p, zauważamy, że jego *i*-ta składowa będzie miała postać $\frac{d(x^i \circ \gamma)}{ds}$. Wyrażenie to często upraszcza się do postaci $\frac{dx^i}{ds}$, z zaznaczeniem, że różniczkujemy funkcję współrzędnej x^i „wzdłuż" danej krzywej γ parametryzowanej zmienną *s*.

Będziemy teraz chcieli obliczyć pochodną kowariantną dowolnego wektora V_m wzdłuż krzywej γ. Robi się to przez wzięcie iloczynu pochodnej kowariantnej po współrzędnych ze współrzędnymi wektora stycznego do krzywej. Rezultatem jest operacja pochodnej kowariantnej wzdłuż krzywej ∇_γ, która wygląda następująco:

$$\nabla_\gamma V_m = (\nabla_n V_m)\frac{dx^n}{ds} = \frac{\partial V_m}{\partial x^n}\frac{dx^n}{ds} + \Gamma^r_{mn} V_r \frac{dx^n}{ds}.$$

Zauważmy, że pierwszy składnik sumy można uprościć do postaci całkowitej pochodnej z wektora V_m po zmiennej *s* (znowu jest to reguła łańcuchowa w działaniu):

$$\nabla_\gamma V_m = \frac{dV_m}{ds} + \Gamma^r_{mn} V_r \frac{dx^n}{ds}.$$

[14] Okazuje się, że każdy wektor styczny do rozmaitości w punkcie (wektor w przestrzeni stycznej T_p) można przedstawić jako wektor styczny do pewnej krzywej γ. Mamy zatem alternatywną i równoważną definicję przestrzeni stycznej jako przestrzeni wszystkich wektorów stycznych do jakiejś krzywej przechodzącej przez punkt *p*.

Załóżmy teraz, że chcemy zbadać zmienność samego wektora stycznego wzdłuż krzywej γ. W tym celu musimy wstawić do powyższego równania $\frac{dx^m}{ds}$ w miejsce V_m:

$$\nabla_\gamma \frac{dx^m}{ds} = \frac{d^2x^m}{ds^2} + \Gamma^r_{mn}\frac{dx^r}{ds}\frac{dx^n}{ds}.$$

Jak już powiedzieliśmy, linia geodezyjna to taka, której wektor styczny nie zmienia się przy przesuwaniu się po tej linii. Oznacza to, że jego pochodna kowariantna wzdłuż geodezyjnej jest równa zeru. Mamy zatem tzw. równanie geodezyjnej:

$$\frac{d^2x^m}{ds^2} + \Gamma^r_{mn}\frac{dx^r}{ds}\frac{dx^n}{ds} = 0. \tag{6.9}$$

Ciekawe, że w równaniu tym występuje druga pochodna położenia po pewnym parametrze, który w wypadku linii świata może być traktowany jako czas (własny), co daje przyspieszenie. Nasuwa to skojarzenie z drugim prawem dynamiki Newtona, jeśli utożsamimy człon $-\Gamma^r_{mn}\frac{dx^m}{ds}\frac{dx^n}{ds}$ z „siłą". Okazuje się, że w geometrii o zerowej krzywiźnie (płaskiej), takiej jak geometria Minkowskiego, symbole Christoffela Γ^r_{mn} są równe zeru. Równanie geodezyjnej przechodzi więc w formułę $a = 0$, co istotnie definiuje linię prostą, zgodnie z pierwszą zasadą dynamiki. Natomiast w geometrii zakrzywionej przyspieszenie nie będzie zerowe – przedmioty poruszające się po geodezyjnych będą się zachowywały tak, jakby działała na nie siła.

Umiemy już w języku geometrii różniczkowej wyrazić pojęcie linii „prostej" (geodezyjnej). Nadal jednak brakuje nam bardzo ważnego pojęcia długości. W jaki sposób możemy matematycznie przedstawić długość krzywych? Potrzebna jest tutaj pewna miara zwana metryką. Najogólniejszy sposób wprowadzenia metryki opiera się na pojęciu tensora metrycznego. Formalnie rzecz ujmując, zaczynamy od wprowadzenia operacji iloczynu skalarnego na wektorach stycznych. Iloczyn skalarny, który omówiliśmy już wcześniej, znany jest ze szkolnego kursu geometrii: przedstawia się go zwykle jako iloczyn długości wektorów razy sinus kąta między nimi, lub też w rozpisaniu na składowe jako sumę iloczynów poszczególnych składowych. W ujęciu abstrakcyjnym iloczynem skalarnym wektorów A i B może być dowolna funkcja w liczby rzeczywiste $g(A, B)$, spełniająca pewne warunki – symetryczności, liniowości oraz (zazwyczaj) warunek dodatniej określoności. Ten ostatni warunek może być jednak pominięty – np. w wypadku geometrii Minkowskiego, jak pamiętamy z poprzedniego rozdziału iloczyn dwóch wektorów może być liczbą ujemną (metrykę opartą na takim pojęciu iloczynu nazywa się pseudo-Riemannowską).

Formalnie, tensor metryczny jest to tensor kowariantny drugiego rzędu, który każdej parze wektorów stycznych (kontrawariantnych) przypisuje ich iloczyn skalarny. Rozważmy wektor styczny do danej krzywej $\gamma(s)$, którego składowe w pewnym układzie wynoszą $\frac{dx^i}{ds}$. Iloczyn skalarny takiego wektora z samym sobą (czyli jego długość do kwadratu) można najogólniej przedstawić w postaci kombinacji liniowej:

$$g(\gamma(s), \gamma(s)) = g_{ij}\frac{dx^i}{ds}\frac{dx^j}{ds},$$

gdzie g_{ij} są pewnymi liczbami. Liczby te będziemy interpretować jako składowe tensora metrycznego w pewnym układzie współrzędnych. Zakładając konwencjonalnie, że wektory styczne mają długość jednostkową, możemy powyższe równanie przepisać jako:

$$1 = g_{ij}\frac{dx^i}{ds}\frac{dx^j}{ds},$$

a stąd stosując zwykłą algebrę do elementów infinitezymalnych *ds*, otrzymujemy najbardziej znaną formułę na tensor metryczny (tzw. element liniowy):[15]

$$ds^2 = g_{ij}dx^i dx^j.$$

Wyrażenie *ds* interpretuje się jako infinitezymalne (nieskończenie małe) przesunięcie wzdłuż pewnej krzywej. W takiej interpretacji tensor metryczny mówi nam, jak takie przesunięcie zależy od przesunięć wzdłuż poszczególnych współrzędnych. W geometrii Euklidesowej przedstawionej we współrzędnych kartezjańskich element liniowy jest oczywiście dany za pomocą twierdzenia Pitagorasa:

$$ds^2 = \sum_i \left(dx^i\right)^2, \tag{6.10}$$

skąd wynika, że składowe tensora metrycznego mają postać $g_{ij} = 1$ dla $i = j$ oraz zero dla pozostałych przypadków.

Przy pomocy tensora metrycznego możemy obliczyć długość danej krzywej. W każdym punkcie należy określić infinitezymalne przesunięcie za pomocą długości wektora stycznego, a następnie takie przesunięcia posumować. Oczywiście pamiętamy, że sumowanie nieskończenie małych odcinków odbywa się za pomocą całkowania. Zatem formuła wyrażająca długość krzywej między dwoma punktami *p* i *q* będzie miała następującą postać:

$$\int_a^b |\gamma(s)|ds,$$

gdzie granice całkowania *a* i *b* są dane przez $\gamma(a) = p$, $\gamma(b) = q$; $\gamma(s)$ oznacza wektor styczny do danej krzywej w punkcie parametryzowanym przez *s*, a pionowe linie – długość. Korzystając z faktu, że długość wektora dana jest przez pierwiastek z jego iloczynu skalarnego z samym sobą, mamy następujące określenie długości krzywej, zawierające jawnie tensor metryczny g_{ij}:

$$\int_a^b \sqrt{g_{ij}\frac{dx^i}{ds}\frac{dx^j}{ds}}\,ds.$$

Nawiasowo wspomnijmy, że geodezyjne można alternatywnie zdefiniować jako linie, dla których długość obliczona zgodnie z powyższym wzorem jest minimalna (przy założeniu, że metryka jest Riemannowska, tj. nieujemna). Pamiętamy z rozdziału poświęconego mechanice analitycznej, że warunkiem minimalizacji całki takiej jak powyższa jest spełnienie równania Eulera-Lagrange'a przez całkowaną funkcję. Okazuje się (nie jest to trudne do sprawdzenia), że równanie Eulera-Lagrange'a dla powyższej całki ma identyczną formę jak równanie geodezyjnych (6.9), co pokazuje równoważność dwóch pojęć linii geodezyjnych – jednego opartego na warunku minimalizacji długości, a drugiego na zerowaniu się pochodnej kowariantnej z wektora stycznego.

[15] Można sprawdzić, że liczby g_{ij} transformują się we właściwy sposób dla tensorów kowariantnych (6.6). Wskazówka: infinitezymalne przesunięcie dx^i można przedstawić w nowym układzie jako $dy^j \frac{\partial x^i}{\partial y^j}$.

Tensor metryczny pełni fundamentalną rolę przy opisie podstawowych własności geometrycznych danej przestrzeni. Przy jego pomocy można np. stwierdzić, czy dana geometria jest zakrzywiona, czy płaska. Tensor ten wyznacza też inne istotne parametry geometryczne przestrzeni. Na przykład istnieje zależność między tensorem metrycznym g_{ij} a symbolami Christoffela Γ^{i}_{jk}. Zależność ta wynika z założenia tzw. kompatybilności pochodnej kowariantnej (koneksji afinicznej) z tensorem metrycznym. Pochodna kowariantna jest kompatybilna z tensorem metrycznym, gdy jej zastosowanie do tensora metrycznego daje zero (tensor metryczny nie zmienia się od punktu do punktu). Okazuje się, że dla danego tensora metrycznego istnieje dokładnie jedna kompatybilna pochodna kowariantna, a zatem także jeden zestaw symboli Christoffela. Matematyczna zależność między tymi symbolami a tensorem metrycznym jest dość skomplikowana – można ją sprawdzić w każdym podręczniku do ogólnej teorii względności.

Przejdźmy teraz do kolejnego ważnego pojęcia, jakim jest krzywizna przestrzeni. Dokładniej chodzi nam o tzw. krzywiznę wewnętrzną, czyli taką, która nie zależy od tego, w jakiej przestrzeni jest „zanurzona" nasza przestrzeń. Na przykład powierzchnia walca jest zakrzywiona, jeśli patrzymy na nią z perspektywy trójwymiarowej przestrzeni, natomiast nie charakteryzuje się ona krzywizną wewnętrzną (obowiązują w niej wszystkie zasady geometrii Euklidesowej). Z kolei powierzchnia sfery ma niezerową krzywiznę wewnętrzną. Zakrzywienie wewnętrzne przestrzeni można zidentyfikować, badając postać tensora metrycznego. Okazuje się, że w przestrzeni płaskiej (niezakrzywionej) tensor metryczny da się przedstawić w postaci delty Kroneckera, gdzie elementy diagonalne tensora g_{ii} są wszystkie równe jedności (zakładamy na razie, że przestrzeń jest Riemannowska, tzn. iloczyn skalarny wektorów jest dodatnio określony). Innymi słowy, istnieje układ współrzędnych (kartezjański), w którym element liniowy ma postać wynikającą z twierdzenia Pitagorasa (6.10). Warto przy tym zauważyć, że zmieniając współrzędne z kartezjańskich na krzywe (np. biegunowe), możemy otrzymać postać tensora metrycznego niespełniającą tego warunku (w której diagonalne niezerowe składniki nie będą równe 1). Jednakże istotne jest tylko to, że w pewnym układzie tensor metryczny może mieć postać macierzy jednostkowej. Natomiast przestrzenie z wewnętrzną krzywizną (jak np. powierzchnia sfery) nie mają tej własności. Tensor metryczny w *każdym* układzie współrzędnych będzie różny od macierzy jednostkowej.

Najbardziej „syntetyczna" matematyczna reprezentacja krzywizny wewnętrznej w danym punkcie przestrzeni dana jest w postaci tensora zwanego tensorem krzywizny Riemanna. Nie będziemy szczegółowo wyprowadzać postaci tego tensora; poprzestaniemy na poglądowym przedstawieniu odpowiedniej procedury. Zasadnicza idea jest dość prosta; przedstawiliśmy ją zresztą już wcześniej w ogólnych zarysach w paragrafie (6.2). Krzywizna w danym obszarze jest niezerowa, gdy dokonując operacji przeniesienia równoległego danego wektora wzdłuż zamkniętej krzywej, stwierdzimy różnicę między wektorem początkowym a końcowym. Nieco ściślej, wybieramy dwie współrzędne x^{μ} i x^{ν} i rozważamy bardzo mały czworokąt, którego boki dane są odpowiednio przez infinitezymalne przesunięcie δx^{μ} wzdłuż współrzędnej x^{μ}, infinitezymalne przesunięcie δx^{ν} wzdłuż współrzędnej x^{ν}, a następnie przesunięcia w odwrotnym kierunku, tak aby powrócić do punktu wyjścia. Wybieramy następnie dowolny wektor V^{β} (ściśle rzecz biorąc, pole wektorowe) w punkcie wyjścia i staramy się obliczyć, jak będzie wyglądała zmiana tego wektora wzdłuż czterech boków wybranego czworokąta. Ogólna formuła reprezentująca taką infinitezymalną zmianę będzie miała następującą postać:

$$\delta V^{\alpha} = \delta x^{\mu} \delta x^{\nu} (\ldots) V^{\beta}, \tag{6.11}$$

gdzie nawias (…) symbolizuje jakiś matematyczny obiekt, którego dokładną postać należy dopiero obliczyć. Zauważmy, że aby zachować spójność całej formuły, obiekt ten musi charakteryzować się czterema wskaźnikami α, β, μ, ν (trzy zostaną „wchłonięte" za pomocą sumowania przez indeksy obiektów po prawej stronie równania, a czwarty będzie indeksem tożsamym z indeksem obiektu po lewej stronie). Zatem można przypuszczać, że poszukiwanym obiektem jest tensor czwartego rzędu. Jest to tzw. tensor krzywizny Riemanna. Należy się również spodziewać, że tensor ten będzie wyrażalny przy pomocy pochodnych kowariantnych, gdyż infinitezymalne zmiany obiektów wzdłuż danych współrzędnych zwykle przedstawia się w postaci pochodnych razy zmiana tej współrzędnej. Istotnie, okazuje się, że wyrażenie w nawiasie (…) da się przedstawić syntetycznie w postaci tzw. komutatora operacji pochodnych kowariantnych wzdłuż współrzędnych:

$$[\nabla_\mu, \nabla_\nu],$$

którego definicja jest następująca: $\nabla_\mu \nabla_\nu - \nabla_\nu \nabla_\mu$. Zauważmy, że wyrażenie to daje zerową wartość, jeśli operacje brania pochodnych w różnych kierunkach są przemienne, czyli „komutują". Okazuje się, że w przestrzeni płaskiej (bez zakrzywienia) operacje pochodnych kowariantnych istotnie nie zależą od kolejności ich wykonania. To znaczy, rezultat operacji zróżniczkowania danego wektora wzdłuż współrzędnej x^μ, a następnie wzdłuż x^ν, jest taki sam, jak wykonanie tych operacji w odwrotnej kolejności. Łatwo się domyślić, że fakt komutowania tych operacji powoduje, że zmiana wektora wzdłuż krzywej zamkniętej będzie zerowa – wrócimy do punktu wyjścia.

Jednakże w przestrzeni zakrzywionej pochodne kowariantne w różnych kierunkach nie komutują, czyli ich komutator jest niezerowy. Jest to właśnie tensor krzywizny Riemanna. Chociaż w powyższej formule występują jawnie tylko dwa indeksy μ i ν, to jednak „ukryte" są w nim dodatkowe dwa indeksy wynikające z definicji pochodnej kowariantnej (6.8). Zachowując spójność z formułą (6.11), zapiszemy tensor Riemanna jako $R_{\mu\nu}{}^{\alpha}{}_{\beta}$. Można również spotkać wariant tego tensora z samymi indeksami dolnymi $R_{\mu\nu\alpha\beta}$, powstały w wyniku operacji opuszczenia indeksu. Tensor krzywizny Riemanna da się przedstawić w postaci odpowiedniej kombinacji symboli Christoffela $\Gamma^{\alpha}_{\beta\gamma}$, przez podstawienie formuły (6.8) do wyrażenia (6.11) i rozwinięcie całego komutatora. W rezultacie otrzymujemy dość skomplikowaną kombinację, w której występują zarówno pochodne cząstkowe symboli Christoffela, jak i ich iloczyny.

Wspomnijmy jeszcze na koniec o innym ważnym tensorze – tzw. tensorze Ricciego. Powstaje on przez „zwężenie" dwóch indeksów tensora Riemanna. Zwężenie polega na zastąpieniu dwóch indeksów jednym, przy zastosowaniu konwencji Einsteina nakazującej zsumowanie wszystkich składowych tensora po wartościach powtarzających się indeksów. W ten sposób z tensora o czterech indeksach otrzymujemy tensor drugiego rzędu:

$$R_{\mu\nu} = R_{\mu\alpha}{}^{\alpha}{}_{\nu}.$$

Tensor Ricciego odgrywa kluczową rolę w sformułowaniu podstawowego prawa ogólnej teorii względności, łączącego własności geometryczne czasoprzestrzeni z rozkładem mas i energii. Dodatkowo w prawie tym będzie występować jeszcze jedna wielkość, powstająca w wyniku kolejnego zwężenia dwóch pozostałych indeksów tensora Ricciego. Oczywiście w wyniku tej operacji powstanie obiekt z zerową liczbą „wolnych" indeksów, czyli skalar.

Zwany jest on skalarem krzywizny i oznaczany literą R.[16] Przejdziemy teraz do omówienia matematycznych i fizycznych szczegółów tego prawa, wyrażonego w tzw. równaniach Einsteina.

6.7.* Fizyka w zakrzywionej czasoprzestrzeni

Cały poprzedni paragraf poświęcony był wprowadzeniu pojęć matematycznych niezbędnych do opisania geometrii zakrzywionych przestrzeni. Do pojęć tych zaliczają się przede wszystkim: pojęcie tensora metrycznego i oparte na nim obiekty matematyczne, takie jak pochodna kowariantna (wraz z symbolami Christoffela), linie geodezyjne (o najmniejszej możliwej długości), pojęcie przeniesienia równoległego oraz tensory krzywizny (Riemanna i Ricciego). W niniejszym paragrafie wprowadzimy obiekty opisujące nie geometrię, lecz własności fizyczne obiektów znajdujących się w czasoprzestrzeni. Dokładniej, zaczniemy od wytłumaczenia pojęcia tensora energii-pędu, aby następnie podać heurystyczne wytłumaczenie i sens fizyczny podstawowego równania Einsteina, łączącego fizykę z geometrią. Zanim to jednak zrobimy, musimy dokonać niezbędnej modyfikacji części geometrycznej naszej teorii. Do tej pory zakładaliśmy milcząco, że nasza przestrzeń jest zasadniczo zbliżona do przestrzeni Euklidesowej, o ile jej krzywizna jest pomijalna. W szczególności, kiedy krzywizna przestrzeni jest dokładnie równa zeru, zakładaliśmy, że tensor metryczny przyjmie postać delty Kroneckera δ_{nm}, czyli macierzy zawierającej same jedynki na przekątnej oraz zera wszędzie indziej. Innymi słowy, w płaskiej przestrzeni infinitezymalna odległość ds dana jest w postaci twierdzenia Pitagorasa, charakterystycznego dla geometrii Euklidesowej.

Jednakże, jak pamiętamy z rozdziału poświęconego szczególnej teorii względności, czasoprzestrzeń posiada inną strukturę metryczną. Jej inwariantem nie jest odległość Euklidesowa, a interwał czasoprzestrzenny, który uwzględnia współrzędną czasową, ale w sposób odmienny od współrzędnych przestrzennych (wchodzą one z przeciwnym znakiem do znaku współrzędnej czasowej):

$$ds^2 = (dt)^2 - (dx)^2 - (dy)^2 - (dz)^2.$$

Zatem tensor metryczny w tym wypadku będzie miał postać macierzy z liczbami 1 oraz –1 na przekątnej (pamiętajmy ponadto o konwencji $c = 1$). Przyjmując notację wprowadzoną w poprzednim rozdziale, zgodnie z którą współrzędną czasową t oznaczamy zerowym indeksem x^0, a współrzędne przestrzenne zapisujemy jako x^1, x^2, x^3, możemy podać składowe tensora metrycznego w płaskiej czasoprzestrzeni Minkowskiego w następujący sposób:

$$g_{00} = 1;\ g_{11} = -1;\ g_{22} = -1;\ g_{33} = -1$$

oraz zero we wszystkich pozostałych przypadkach. Taki tensor oznacza się często symbolem $\eta_{\mu\nu}$.

Pozostałe definicje podane w poprzednim paragrafie nie ulegają zmianie. Przejdźmy teraz do problemu matematycznej reprezentacji fizycznych źródeł pola grawitacyjnego, czyli w obecnym ujęciu krzywizny czasoprzestrzeni. Zgodnie z teorią grawitacji Newtona, źródłem grawitacji jest masa obiektów, czyli pewien skalar. To podejście musi ulec podwójnej

[16] Formalnie skalar krzywizny R definiowany jest jako R^{μ}_{μ}. Powstaje on w wyniku zastosowania dwóch operacji do tensora Ricciego $R_{\mu\nu}$: najpierw podnosimy jeden z dwóch indeksów, a następnie zwężamy go z dolnym indeksem.

modyfikacji. Po pierwsze, jak wiemy ze szczególnej teorii względności, masa jest równoważna energii, zgodnie ze słynnym wzorem $E = mc^2$. Zatem energia powinna również wnosić wkład do powstania krzywizny czasoprzestrzeni, którą obserwujemy jako pole grawitacyjne. Pojawia się jednak tutaj pewien problem. Energia nie jest inwariantem transformacji międzyukładowych. Energia przedmiotu znajdującego się w spoczynku jest równa mc^2, natomiast z punktu widzenia układu poruszającego się z prędkością v jego energia wzrasta o czynnik $\frac{1}{\sqrt{1-\frac{v^2}{c^2}}}$. Aby temu zaradzić, wprowadza się jeszcze dodatkowo pęd. Jak pamiętamy, w szczególnej teorii względności energia i pęd występują w tzw. czterowektorze energii-pędu P^μ (energia jest jego pierwszą składową, a trzy pozostałe składowe reprezentują trzy składowe pędu). Długość czterowektora energii-pędu (oczywiście liczona w zgodzie z metryką Minkowskiego) jest niezmiennikiem transformacji.

Jednak to nie długość czterowektora P^μ jest wielkością, która charakteryzuje fizyczne źródło pola grawitacyjnego. Drugą modyfikacją teorii grawitacji Newtona będzie przyjęcie, że wielkość fizyczna „odpowiedzialna" za powstanie krzywizny czasoprzestrzeni jest tensorem drugiego rzędu, oznaczanym jako $T^{\mu\nu}$. Dlaczego? Odpowiedź znajdziemy w szczegółowej postaci składowych tego tensora. Zacznijmy jednakże od przypomnienia innego ważnego pojęcia, jakie pojawiło się przy okazji analizy zjawisk elektromagnetycznych w języku szczególnej teorii względności (par. 5.9). Chodzi tutaj o czterowektor gęstości prądu J^μ, który syntetycznie ujmuje dwa źródła pola elektromagnetycznego: ładunek elektryczny i prąd. Czterowektor ten można zapisać w postaci (ρ, j_x, j_y, j_z), gdzie ρ jest gęstością ładunku w danym punkcie czasoprzestrzeni, a j_x, j_y, j_z są odpowiednimi składowymi wektora prądu wzdłuż kierunków x, y i z. Dokładniej, składową j_x przedstawiamy jako strumień ładunku przepływający przed pewną infinitezymalną powierzchnię prostopadłą do kierunku x w jednostce czasu. Ponieważ pole takiej powierzchni to iloczyn $dydz$, ogólnie składową prądu j_x da się zapisać jako (Q jest ładunkiem):

$$j_x = \frac{dQ}{dydzdt}$$

i podobnie dla pozostałych składowych:

$$j_y = \frac{dQ}{dxdzdt};$$

$$j_z = \frac{dQ}{dxdydt}.$$

Zauważmy, że gęstość ładunku można przedstawić w analogiczny sposób, gdyż będzie to ładunek podzielony przez jednostkę objętości:

$$\rho = \frac{dQ}{dxdydz}.$$

Obserwujemy tu pewną prawidłowość. Każdy ze składników czterowektora J^μ jest wyrażony jako iloraz infinitezymalnej zmiany ładunku oraz trójwymiarowej infinitezymalnej „powierzchni" w czterowymiarowej czasoprzestrzeni. Składniki reprezentujące przepływ prądu zawierają „powierzchnie" zbudowane z dwóch współrzędnych przestrzennych i jednej czasowej, podczas gdy składnik gęstości ładunku ma w mianowniku powierzchnię czysto prze-

strzenną, bez elementu czasowego. Można zatem interpretować gęstość ładunku ρ jako „strumień" ładunku płynący wzdłuż współrzędnej czasowej t, tak samo jak j_x oznaczało strumień ładunku wzdłuż współrzędnej x.

Dodatkowym ważnym elementem analizy czterowektora prądu jest równanie ciągłości, które wyraża uogólnioną („lokalną") zasadę zachowania ładunku. Równanie to mówi, że zmiana ilości ładunku w danej objętości przestrzennej musi być „skompensowana" przepływem prądu przez ścianki tej objętości. Skoro gęstość ładunku można zinterpretować jako strumień wzdłuż osi czasu, oznacza to po prostu, że całkowita ilość ładunku w czterowymiarowej objętości nie ulega zmianie, albo też, że całkowity „wypływ" ładunku uwzględniający cztery wymiary jest zerowy. Matematyczna postać równania ciągłości dla ładunku, jak już wcześniej pisaliśmy w paragrafie 5.9, jest następująca:

$$\frac{\partial \rho}{\partial t} = -\nabla \cdot \boldsymbol{j}.$$

W jednorodnej formie czterowymiarowej równanie to przyjmuje postać:

$$\frac{\partial J^{\mu}}{\partial x^{\mu}} = 0.$$

Wyrażenie po lewej stronie to czterowymiarowa dywergencja, oznaczająca wypływ czterowektora J^{μ}.

Analogiczną procedurę można zastosować do energii, a następnie do każdej składowej pędu. W przypadku energii możemy mówić zarówno o jej gęstości, jak i o przepływie energii przez odpowiednie powierzchnie, prostopadłe do osi x, y i z. Zatem energia będzie wymagała czterech składowych. Ponadto każda z trzech składowych pędu p_x, p_y i p_z również wymaga czterech parametrów. Są to: gęstość danej składowej oraz jej strumień we wszystkich trzech kierunkach przestrzennych. Zatem łącznie mamy $16 = 4 \times 4$ parametrów, co daje nam tensor drugiego rzędu $T^{\mu\nu}$.[17]

	Oś t	Oś x	Oś y	Oś z
Energia	T^{00}	T^{01}	T^{02}	T^{03}
Pęd p_x	T^{10}	T^{11}	T^{12}	T^{13}
Pęd p_y	T^{20}	T^{21}	T^{22}	T^{23}
Pęd p_z	T^{30}	T^{31}	T^{32}	T^{33}

Przedstawmy dokładniej poszczególne składowe tego tensora, zwanego tensorem energii-pędu. Składowa T^{00} oznacza gęstość energii (energię na jednostkę objętości), czyli strumień energii wzdłuż osi czasu t. Składowa T^{01} to z kolei strumień energii wzdłuż osi x. Podobnie T^{02} oraz T^{03} będą reprezentować strumienie energii wzdłuż kolejnych osi y i z; składowa tensora T^{10} to gęstość składowej p_x pędu, a składowe T^{11}, T^{12} i T^{13} to strumienie tej składowej wzdłuż kierunków x, y i z. To samo będzie dotyczyć kolejnych składowych pędu p_y i p_z. Zestawmy składowe tensora energii-pędu w postaci macierzy (por. wyżej), w której wiersze odpowiadają energii i składowym pędu, a kolumny reprezentują poszczególne osie

[17] Ściśle rzecz biorąc, aby $T^{\mu\nu}$ był tensorem, jego składowe muszą się transformować w odpowiedni sposób. Tak jest w istocie, ale nie będziemy tego udowadniać.

(czasową i trzy przestrzenne). Na przecięciu odpowiedniego wiersza i kolumny mamy strumień danej wielkości w wybranym kierunku.

W analogii do równania ciągłości dla ładunku możemy zapisać równanie ciągłości dla każdej z czterech wielkości uwzględnionych w tensorze energii-pędu. Na przykład równanie ciągłości energii (zasada zachowania energii) będzie miało postać:

$$\frac{\partial T^{00}}{\partial t}+\frac{\partial T^{01}}{\partial x}+\frac{\partial T^{02}}{\partial y}+\frac{\partial T^{03}}{\partial z}=0.$$

Ujmując sprawę syntetycznie, równanie ciągłości dla tensora energii-pędu przyjmuje formę (pamiętajmy o konwencji sumacyjnej):

$$\frac{\partial T^{\mu\nu}}{\partial x^{\nu}}=0.$$

Jednakże powyższe równanie obowiązuje tylko w czasoprzestrzeni pozbawionej krzywizny. Jak pamiętamy, w ogólnym przypadku pochodne cząstkowe muszą zostać zastąpione pochodnymi kowariantnymi. Zatem ogólne równanie ciągłości dla tensora energii-pędu wygląda następująco:

$$\nabla_{\nu}T^{\mu\nu}=0.$$

Mamy już w zasadzie gotowe wszystkie składniki równania Einsteina – podstawowego równania ogólnej teorii względności. Spróbujmy teraz prześledzić w ogólnych zarysach drogę do tego równania, zaczynając od przypomnienia formalnego opisu grawitacji w teorii Newtona. W rozdziale drugim wprowadziliśmy powszechnie znane prawo grawitacji Newtona, opisujące oddziaływanie między dwoma ciałami punktowymi. W ogólnym wypadku rozkład mas może być jednak dowolny, nie tylko punktowy, a zatem prawo grawitacji musi uwzględniać także taką możliwość. Rozwiązaniem problemu jest podanie równania analogicznego do pierwszego równania Maxwella w wersji różniczkowej (4.7), omówionego w rozdziale czwartym. Równanie to łączy dywergencję pola elektrycznego $\boldsymbol{E}$ z gęstością ładunku w danym punkcie. Z kolei pole elektryczne może być przedstawione jako gradient potencjału elektrycznego (ze znakiem minus). Dywergencja z gradientu jest operacją różniczkową polegającą na dwukrotnym zróżniczkowaniu danego skalara wzdłuż każdej z trzech współrzędnych przestrzennych i dodaniu rezultatów. W wyniku tej operacji zastosowanej do pola grawitacyjnego otrzymujemy równanie, zwane równaniem Poissona:

$$\nabla^2\varphi=-4\pi G\rho,$$

gdzie φ jest potencjałem pola grawitacyjnego, ρ – gęstością masy, a operator ∇^2, jak już wcześniej definiowaliśmy, jest operatorem Laplace'a, $\nabla^2=\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}$.

Idea jest taka, aby nowe równanie opisujące grawitację przechodziło w przybliżeniu w równanie Poissona dla niewielkich prędkości i niewielkich mas. Jednakże w nowym ujęciu nie posługujemy się już pojęciem pola grawitacyjnego i jego potencjału, a tylko geometrycznym ujęciem kinematycznego zachowania ciał w postaci zasady, iż wszystkie ciała poruszają się po liniach geodezyjnych. Będziemy więc musieli znaleźć sposób na „przetłumaczenie" języka pola i potencjału grawitacyjnego na język geometrii. W poprzednim paragrafie wprowadziliśmy opis linii geodezyjnych za pomocą założenia, że pochodna kowariantna wektora stycznego do linii geodezyjnej powinna być tożsamościowo równa zeru (tj. wektor styczny do geodezyjnej nie zmienia swojego kierunku przy przeniesieniu równoległym). Prowadzi to

do równania geodezyjnych, zawierającego symbole Christoffela i odpowiednie pochodne współrzędnych po parametrze charakteryzującym daną linię. Obecnie rozważamy nie dowolne linie, ale tzw. linie świata dla ciał, czyli linie, których wektor styczny jest zawsze wektorem czasopodobnym (skierowanym „ku górze" w stosunku do osi czasu). Linię świata danego obiektu parametryzujemy jego czasem własnym τ (czasem liczonym w układzie, w którym ten obiekt spoczywa). Zatem wektor styczny do danej linii świata ma postać $\frac{dx^\mu}{d\tau}$, a równanie geodezyjnej przyjmie następującą formę:

$$\frac{d^2x^\mu}{d\tau^2} + \Gamma^\mu_{\alpha\beta}\frac{dx^\alpha}{d\tau}\frac{dx^\beta}{d\tau} = 0.$$

Rozważmy powyższe równanie tylko w odniesieniu do współrzędnych przestrzennych, pomijając współrzędną czasową $x^\mu = x^0$ (w przybliżeniu małych prędkości czas własny τ jest praktycznie równy x^0, a zatem druga pochodna z x^0 po τ wynosi 0). Pochodna $\frac{dx^0}{d\tau}$ jest bliska 1, podczas gdy prędkości $\frac{dx^i}{d\tau}$ (gdzie zgodnie z przyjętą konwencją wskaźnik łaciński i przebiega wartości 1, 2, 3) są pomijalne w stosunku do $\frac{dx^0}{d\tau}$ – pamiętajmy o przyjętym założeniu, że $c = 1$. Zatem równanie geodezyjnej przechodzi w następującą zależność:

$$\frac{d^2x^i}{dt^2} + \Gamma^i_{00} = 0.$$

Jest to, jak się można domyślić, wariant klasycznej zasady dynamiki Newtona, gdzie $\frac{d^2x^i}{dt^2}$ jest przyspieszeniem, a $-\Gamma^i_{00}$ reprezentuje siłę (podzieloną przez masę). Symbole Christoffela w ogólności są, jak już pisaliśmy, skomplikowanymi funkcjami tensora metrycznego (zawierają pochodne składników tensora metrycznego po różnych współrzędnych). Jednakże w przybliżeniu, gdy tensor metryczny $g_{\mu\nu}$ odbiega w niewielkim stopniu od tensora płaskiej czasoprzestrzeni Minkowskiego $\eta_{\mu\nu}$, symbol Christoffela Γ^i_{00} można przedstawić jako:

$$\Gamma^i_{00} = -\frac{1}{2}\frac{\partial g_{00}}{\partial x^i}.$$

Ponieważ w mechanice newtonowskiej przyspieszenie ciała w polu grawitacyjnym (tj. siła) jest dane w postaci gradientu potencjału:

$$\frac{d^2x^i}{dt^2} = -\frac{\partial\varphi}{\partial x^i},$$

ostatecznie mamy związek potencjału grawitacyjnego z geometrią, czyli tensorem metrycznym, obowiązujący dla słabych pól grawitacyjnych i małych prędkości:[18]

$$\varphi = \frac{1}{2}g_{00}.$$

[18] Oczywiście do potencjału możemy zawsze dodać dowolną stałą (zgodnie z transformacją cechowania). Nie zmienia to jednak naszych rozważań, gdyż pochodna stałej jest zerowa.

Wstawiając powyższe do równania Poissona oraz pamiętając, że składowa T^{00} tensora energii-pędu reprezentuje gęstość energii – czyli masy – otrzymujemy pierwsze przybliżenie równania Einsteina, łączącego fizykę z geometrią:

$$\frac{1}{2}\nabla^2 g_{00} = 4\pi G T^{00}.$$

Powyższe równanie wskazuje na kluczowe dla ogólnej teorii względności powiązanie własności fizycznej materii (reprezentowanej przez składową tensora energii-pędu) i własności czasoprzestrzeni (składowa tensora metrycznego). Nie jest to jednak zadowalające rozwiązanie dla ogólnego przypadku. Nasze równanie zawiera bowiem wybraną składową, a nie całość tensora energii-pędu. Jak pamiętamy, składowe tensorów zmieniają się przy zmianie układu współrzędnych. Fundamentalne równanie teorii względności powinno być niezmiennicze względem dowolnych transformacji współrzędnych, a więc powinno zawierać pełne tensory, a nie ich wyróżnione składowe. Do tego oczywiście dochodzi fakt, że zaproponowane równanie obowiązuje tylko w szczególnych układach odniesienia (poruszających się z niewielką prędkością) i dla niewielkich mas-energii. Zatem należy poszukiwać ogólnego równania o następującej postaci:

$$G^{\mu\nu} = kT^{\mu\nu}, \tag{6.12}$$

którego szczególnym przypadkiem jest powyższa równość.

W powyższym zapisie $G^{\mu\nu}$ jest pewnym tensorem reprezentującym własności geometryczne czasoprzestrzeni, a k odpowiednią stałą. Z wcześniejszych rozważań wynika, że tensor $G^{\mu\nu}$ powinien zawierać drugie pochodne tensora metrycznego. Takim tensorem mógłby być tensor krzywizny Riemanna, który istotnie spełnia taki warunek (zawiera on bowiem pochodne symboli Christoffela, które z kolei definiowane są przy pomocy kombinacji pochodnych tensora metrycznego). Jednakże tensor ten jest czwartego rzędu, więc matematycznie nie może być zrównany z tensorem energii-pędu, bo ten zawiera dwa indeksy. Natomiast tensor Ricciego $R^{\mu\nu}$, który jest zwężeniem tensora Riemanna, mógłby być tutaj zastosowany. Niestety na przeszkodzie stoi fakt, że tensor energii-pędu musi spełniać równania ciągłości, wyrażające zasadę zachowania pędu i energii. Skoro pochodna kowariantna prawej strony równania jest równa zeru, to samo powinno dotyczyć lewej strony równania. Można jednak obliczyć, że pochodna kowariantna tensora Ricciego nie jest w ogólności zerowa. Wygląda ona następująco:

$$\nabla_\mu R^{\mu\nu} = \frac{1}{2} g^{\mu\nu} \frac{\partial R}{\partial x^\mu},$$

gdzie R to skalar krzywizny. Uwzględniając wspomniany wcześniej fakt, że pochodna kowariantna tensora metrycznego jest zawsze równa zeru[19], oraz że pochodna kowariantna każdego skalara jest tożsama ze zwykłą pochodną cząstkową, możemy przedstawić prawą stronę powyższego równania jako $\nabla_\mu \frac{1}{2} g^{\mu\nu} R$. Zatem znaleźliśmy obiekt, którego pochodna kowariantna jest tożsamościowo równa zeru. Jest to tzw. tensor Einsteina:

$$G^{\mu\nu} = R^{\mu\nu} - \frac{1}{2} g^{\mu\nu} R.$$

[19] Zerowanie pochodnej kowariantnej z tensora metrycznego jest warunkiem kompatybilności koneksji afinicznej z tensorem metrycznym.

Wstawiając tę formułę do wzoru (6.12), otrzymujemy wreszcie słynne równanie Einsteina, będące kwintesencją ogólnej teorii względności:

$$R^{\mu\nu} - \frac{1}{2} g^{\mu\nu} R = kT^{\mu\nu}.$$

Teoretycznie równanie powyższe zawiera wszystko, co jest nam potrzebne do obliczenia własności metrycznych czasoprzestrzeni przy założeniu danego rozkładu mas (energii) oraz pędu. W praktyce jednak rozwiązanie równania (a raczej równań, ponieważ rozpada się ono na wiele niezależnych równań dla poszczególnych składowych) Einsteina jest niewykonalne ze względu na wysoki stopień matematycznej złożoności. Stosuje się więc różnego rodzaju założenia upraszczające, np. w postaci nałożonych warunków symetrii.

W przypadku niewielkich mas i prędkości, dominującą składową tensora energii-pędu jest składowa T^{00}, reprezentująca gęstość energii (występuje w niej czynnik c^2, który jest oczywiście ogromną wielkością). Można również pokazać, że lewa strona równania, czyli tensor Ricciego minus tensor metryczny razy R, zbiega w tym wypadku do laplasjanu z zerowej składowej tensora metrycznego: $\nabla^2 g_{00}$. Zatem aby równanie Einsteina przeszło w odpowiednik klasycznego równania Poissona, musimy przyjąć następującą wartość stałej k: $k = 8\pi G$.

Rozważmy jeszcze pytanie, co się stanie, jeśli tensor energii-pędu będzie równy zeru w całym obszarze czasoprzestrzeni. Odpowiada to oczywiście sytuacji całkowicie pustego wszechświata, pozbawionego wszelkiej materii. Zgodnie z równaniem Einsteina, lewa strona musi też się zerować, co daje nam następującą zależność:

$$R^{\mu\nu} = \frac{1}{2} g^{\mu\nu} R.$$

Na pozór wydaje się, że równanie to może mieć rozwiązania w postaci niezerowego tensora Ricciego. Tak jednak nie jest. „Mnożąc" obie strony przez tensor metryczny $g_{\mu\nu}$, a ściślej rzecz ujmując zwężając obie strony z tym tensorem, otrzymujemy równanie zawierające sam skalar krzywizny:

$$R = 2R.$$

Oczywiście wynika z tego, że $R = 0$, a zatem także tensor Ricciego wynosi zero. Czy oznacza to, że czasoprzestrzeń będzie pozbawiona krzywizny? Może nawet się wydawać, zgodnie z sugestią Leibniza, że w takiej sytuacji czasoprzestrzeń w ogóle przestanie istnieć, jako że jest ona jedynie epifenomenem obiektów materialnych. Jednakże matematyka mówi nam co innego. Zerowanie tensora Ricciego, który jest zwężeniem Riemannowskiego tensora krzywizny, nie implikuje znikania tego ostatniego. Istnieją niezerowe formy tensora Riemanna, dla których tensor Ricciego jest równy zeru. Jedna z takich form reprezentuje „bezźródłowe" fale grawitacyjne. Zatem matematyka ogólnej teorii względności dopuszcza istnienie zakrzywionej czasoprzestrzeni nawet w wypadku całkowitego braku materii. Oczywiście pytanie, czy taka możliwość jest fizycznie sensowna, pozostaje otwarte. Większość fizyków zdaje się uważać, że tak jest, ale z oczywistych powodów nie będziemy mogli poddać tej hipotezy bezpośredniej weryfikacji empirycznej.

Pytania i problemy

1. Omów zasadę równoważności Einsteina, posługując się przykładami spadku swobodnego oraz rakiety kosmicznej poruszającej się ze stałym przyspieszeniem. Porównaj zasadę Einsteina z zasadą względności Galileusza.

2. Co to są linie geodezyjne? Podaj nieformalną interpretację dwóch matematycznych definicji pojęcia geodezyjnych. Jakie linie na sferze są liniami geodezyjnymi?

3. Omów pojęcie krzywizny wewnętrznej, operując pojęciem przeniesienia równoległego. Podaj przykład dwuwymiarowej powierzchni zakrzywionej w trzech wymiarach, której krzywizna wewnętrzna jest zerowa, oraz takiej, dla której krzywizna wewnętrzna jest większa lub mniejsza od zera.

4. Jaka zasada w ogólnej teorii względności opisuje kinematyczne zachowanie w polu grawitacyjnym ciał niepoddanych działaniu żadnych innych sił? Jaki jest związek tej zasady z pierwszą zasadą dynamiki Newtona? Jak będzie wyglądała trajektoria ciała swobodnego w polu grawitacyjnym?

5. Omów zasadniczą ideę fizyczną zawartą w równaniu pola Einsteina. Jaka własność czasoprzestrzeni jest uzależniona od rozkładu materii?

6. Jakie dwa postulaty doprowadziły Einsteina do sformułowania konkretnej postaci tzw. tensora Einsteina i całego równania łączącego rozkład materii z krzywizną czasoprzestrzeni?

7. Czy geometryzacja oddziaływania grawitacyjnego jest unikalną cechą ogólnej teorii względności, nieobecną w fizyce klasycznej? Jaka jest zasadnicza różnica między OTW a geometryczną wersją mechaniki newtonowskiej?

8. Omów najważniejsze testy empiryczne potwierdzające słuszność ogólnej teorii względności.

9. Porównaj historyczny problem obserwowanych anomalii w ruchu planety Uran i jego rozwiązanie z analogicznym problemem anomalii w ruchu Merkurego.

10. Czy ogólna teoria względności realizuje ideę Macha całkowitej relatywizacji wszelkich ruchów, w tym przyspieszonych?

11. Jakie argumenty zaczerpnięte z ogólnej teorii względności przemawiają na korzyść tezy o substancjalnym charakterze czasoprzestrzeni?

12. Przedstaw w ogólnych zarysach ideę argumentu dziury za relacjonizmem w kwestii natury czasoprzestrzeni. Porównaj ten argument z argumentem Leibniza z przesunięcia. Wykorzystaj pojęcia aktywnej i pasywnej interpretacji transformacji układu odniesienia.

Literatura uzupełniająca

Klasycznym podręcznikiem do ogólnej teorii względności jest: B.F. Schutz, *Wstęp do ogólnej teorii względności*, PWN, Warszawa 1995.

Godna polecenia jest niewielka książka: W. Kopczyński, P. Trautman, *Czasoprzestrzeń i grawitacja*, PWN Warszawa 1984.

Przystępne, lecz pogłębione omówienie czasoprzestrzeni w szczególnej i ogólnej teorii względności zawiera książka znanego fizyka amerykańskiego: R. Geroch, *General Relativity from A to B,* The University of Chicago Press, Chicago 1978.

Bardzo eleganckie przedstawienie matematycznych i filozoficznych aspektów teorii względności znajdziemy w pracy: M. Friedman, *Foundations of Space-Time Theories*, Princeton University Press, Princeton 1983.

Gruntowny przegląd problematyki czasu i przestrzeni, od Newtona do obu teorii względności, zawiera monumentalna praca: L. Sklar, *Space, Time and Spacetime*, University of California Press, Berkeley 1974.

Nowsze dyskusje na temat statusu czasoprzestrzeni, uwzględniające argument dziury i wersje substancjalizmu, znajdziemy w artykule: O. Pooley, „Relationist and substantivalist approaches to spacetime", w: R. Batterman (red.) *The Oxford Handbook of Philosophy of Physics*, Oxford University Press, Oxford, s. 522–586.

Kolejna książka z cyklu „Co musisz wiedzieć, żeby zacząć zajmować się fizyką", która właśnie została przetłumaczona na język polski, to: L. Susskind, A. Cabannes, *Ogólna teoria względności. Teoretyczne minimum*, Prószyński i S-ka, Warszawa 2024.

ROZDZIAŁ 7. MECHANIKA KWANTOWA

Kwantowa rewolucja w fizyce, która miała miejsce na początku dwudziestego wieku, odbiła się w kręgach filozoficznych jeszcze szerszym echem niż powstanie obu teorii względności. Powodów zainteresowania filozofów mechaniką kwantową jest wiele. Po pierwsze, teoria ta sięga swoimi konsekwencjami głęboko do zagadnień *stricte* ontologicznych i epistemologicznych, takich jak problem determinizmu i przyczynowości, realizm czy zagadnienie identyczności i trwania w czasie. Co więcej, mechanika kwantowa jako bodaj jedyna z obecnie uznawanych teorii w fizyce zawiera fundamentalną lukę w swoich teoretycznych podstawach, znaną jako problem pomiaru. Choć wielu fizyków kwestionuje wagę czy wręcz istnienie tego problemu, to jednak znaczna część teoretyków dostrzega konieczność odniesienia się do tego zagadnienia. W rezultacie mechanika kwantowa doczekała się znacznej liczby tzw. interpretacji, których głównym celem jest uzupełnienie owej zasadniczej luki, a przy okazji rozwiązanie pomniejszych problemów, jak np. domniemane istnienie nielokalnych oddziaływań czy radykalny indeterminizm. Interpretacje te zwykle uzupełniają podstawowe założenia teorii o pewne dodatkowe hipotezy czy też reguły interpretacyjne. Do najbardziej znanych interpretacji mechaniki kwantowej należą: teoria wielu światów, mechanika Bohmowska oraz teoria spontanicznej lokalizacji, choć istnieje wiele innych ciekawych podejść.

W niniejszym rozdziale przedstawimy przede wszystkim drogę dojścia do nowej teorii kwantowej, która radykalnie zerwała z obrazem świata obecnym w mechanice klasycznej, a także w obu teoriach względności. Wskażemy na eksperymentalne źródła teorii kwantów i jej nieklasycznego charakteru. Następnie omówimy najważniejsze pojęcia nowej teorii, takie jak pojęcie stanu, superpozycji stanów, stanów splątanych oraz prawdopodobieństwa pomiarowego. Wspomnimy także o podstawowym prawie ewolucji stanów, wyrażonym w równaniu Schrödingera. Analizę filozoficznych problemów mechaniki kwantowej rozpoczniemy od przestawienia słynnego argumentu EPR, sformułowanego przez Einsteina i jego współpracowników. Celem argumentu było wykazanie, że mechanika kwantowa wraz ze swoim nieusuwalnym użyciem prawdopodobieństw nie może być „ostatnim słowem" nauki. Innymi słowy, jest niekompletna i musi być zastąpiona doskonalszą teorią, opartą na zasadzie determinizmu i przewidywalności. Kluczowym założeniem argumentu EPR jest teza lokalności; wspominaliśmy już o niej w kontekście teorii elektromagnetyzmu i szczególnej teorii względności. Zasada lokalności, według Einsteina, wymusza istnienie głębszej realności,

której obecna mechanika kwantowa nie jest w stanie opisać. Jednakże zaskakujące twierdzenie Bella pokazuje, że istnienie takiej głębszej realności (wyrażonej w tzw. parametrach ukrytych) wraz założeniem lokalności prowadzi do niezgodności z dobrze potwierdzonymi doświadczalnie przewidywaniami mechaniki kwantowej.

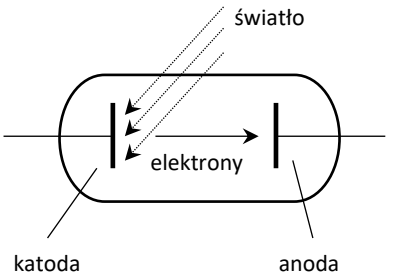
Po omówieniu twierdzenia Bella i jego konsekwencji filozoficznych przejdziemy do centralnego zagadnienia, jakim jest problem pomiaru w mechanice kwantowej. Popularne jego ujęcie przyjmuje formę słynnego paradoksu kota Schrödingera. Alternatywne interpretacje mechaniki kwantowej, o których wspomnimy w niniejszym rozdziale, można przedstawić w formie odmiennych sposobów rozwiązania tego paradoksu. Pokażemy także, jak mechanika kwantowa nakazuje opisywać układy złożone z wielu cząstek tego samego rodzaju, jak np. elektrony czy fotony. Okazuje się, że opis ten ma bardzo interesujące konsekwencje dotyczące kwestii tożsamości i odróżnialności obiektów kwantowych. Zatem znów spotkamy się z filozoficzną zasadą tożsamości przedmiotów nieodróżnialnych, a także z problemem identyfikacji przedmiotów w czasie. Tak jak poprzednio, w paragrafie z gwiazdką bardziej dociekliwi czytelnicy znajdą więcej szczegółów na temat pięknej matematyki wykorzystywanej w teorii kwantowej, opartej na przestrzeniach Hilberta, wektorach i operatorach.

7.1. Od eksperymentów do kwantów

Powstanie mechaniki kwantowej wiąże się nierozerwalnie z rozwojem wiedzy na temat struktury materii. W dziewiętnastym wieku dominującym podejściem w nauce (chemii i fizyce) stał się atomizm, opierający się na poszukiwaniu najprostszych składników materii. Ich fundamentalne własności i wzajemne oddziaływania powinny wyjaśnić złożoność obserwowanego świata materialnego. Podejście to spotkaliśmy przy okazji omawiania redukcji zjawisk cieplnych do procesów molekularnych. Fundamentalnym założeniem tej redukcji była teza, że elementarne składniki materii (molekuły, atomy) podlegają prawom mechaniki newtonowskiej. Jak się za chwilę przekonamy, wraz z coraz dokładniejszą analizą podstawowych składników materii, przekonanie to musiało ulec radykalnej modyfikacji. Przełomem stało się tutaj odkrycie wewnętrznej struktury atomów. Koncepcja atomów jako nieprzenikliwych „grudek" materii została odrzucona na korzyść modelu niewielkiego jądra atomowego i otaczających go elektronów. Odkrycie elektronu i wyznaczenie jego podstawowych parametrów (masy, ładunku) to kolejny krok milowy w kierunku nowej koncepcji materii.

Jednakże pierwszym wskaźnikiem „nieklasyczności" mikroświata była analiza promieniowania elektromagnetycznego, a konkretnie jego oddziaływania z materią. Niemal każdy wstęp do mechaniki kwantowej rozpoczyna się od przedstawienia epizodu z historii fizyki dotyczącego promieniowania tzw. ciała doskonale czarnego. W kontekście tego dość szczególnego zagadnienia po raz pierwszy zostało wprowadzone pojęcie „kwantu", które stało się osnową nowej teorii. Na przełomie dziewiętnastego i dwudziestego wieku fizycy zetknęli się z problemem, jak teoretycznie opisać termiczne promieniowanie wysyłane przez przedmiot o określonej temperaturze. Z doświadczenia wiemy, że promieniowanie takie ma pewien rozkład częstotliwości – najwięcej energii wypromieniowane jest dla pewnej częstotliwości pośredniej, a dla wyższych i niższych częstotliwości energia ta spada do zera. Próbując teoretycznie wyprowadzić wzór na rozkład takiego promieniowania, naukowcy założyli, w zgodzie z klasyczną teorią elektromagnetyzmu Maxwella, że fala stojąca o danej częstotliwości (tzw. mod drgania) może przenosić dowolną wartość energii pomiędzy zerem a nieskończonością, zależną tylko od jej amplitudy.

Niestety, wyprowadzony na podstawie tego założenia wzór miał absurdalną konsekwencję: energia wypromieniowywana przez fale o coraz wyższych częstotliwościach rosła do nieskończoności (tzw. katastrofa w nadfiolecie), co jest oczywiście fizycznie niemożliwe. Niemiecki fizyk Max Planck zauważył, stosując metodę prób i błędów, że jeśli przyjmie się założenie, że energia fali o określonej częstotliwości ν może przyjmować tylko wyróżnione wartości, będące wielokrotnością pewnej stałej razy częstotliwość: $nh\nu$, to wyprowadzony na tej podstawie wzór będzie doskonale zgadzał się z obserwacjami empirycznymi. Porcje energii dostępne dla danego promieniowania nazwał Planck „kwantami", a stała h jest do dzisiaj określana jego imieniem: stała Plancka.

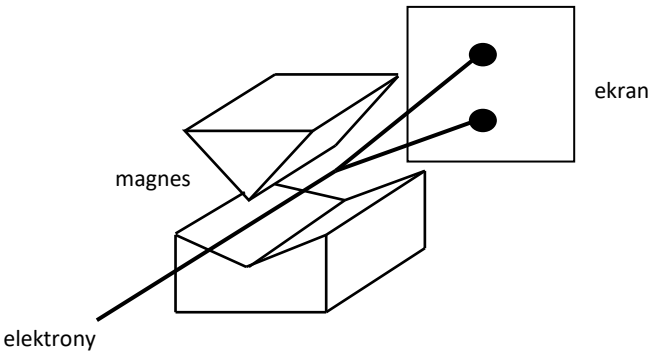


Rys. 7.1. Efekt fotoelektryczny

Hipoteza Plancka dotycząca kwantów promieniowania może być uznana za *ad hoc* i sam jej twórca pierwotnie traktował ten pomysł jedynie jako matematyczny trick, umożliwiający rozwiązanie palącego problemu. Wkrótce jednak pojawiły się dodatkowe doświadczalne fakty wskazujące na istnienie fotonów jako indywidualnych nośników energii promieniowania elektromagnetycznego. Jednym z nich był tzw. efekt fotoelektryczny. Opiera się on na stwierdzonej w eksperymencie możliwości „wybicia" wolnych elektronów z metalu przez promieniowanie elektromagnetyczne (światło). W opróżnionej z powietrza rurze (rys. 7.1) umieszcza się dwie elektrody pod napięciem: dodatnią (anodę) i ujemną (katodę). W normalnych warunkach przez rurę nie przepływa prąd, gdyż w próżni nie ma nośników elektryczności. Jednakże kiedy oświetlimy katodę, pojawia się przepływ prądu, co tłumaczy się tym, że energia dostarczona przez promieniowanie świetlne umożliwia elektronom wydostanie się z katody (pokonanie bariery energetycznej) i przepływ w kierunku dodatnio naładowanej anody. Jak pokazano doświadczalnie, prąd rośnie zarówno ze wzrostem natężenia promieniowania, jak i jego częstotliwości. Jest to wytłumaczalne na gruncie klasycznej teorii elektromagnetyzmu, bo zarówno natężenie (proporcjonalne do amplitudy fali), jak i częstość fali (liczba drgań na sekundę) określają dostarczaną przez falę energię. Im większa energia, tym więcej wybitych elektronów i tym większy prąd. Natomiast trudny do klasycznego wytłumaczenia był inny fakt – zaobserwowano mianowicie istnienie częstotliwości progowej, poniżej której nie pojawiał się żaden prąd, mimo zwiększania natężenia światła. Rozwiązaniem, zaproponowanym przez Einsteina, było przyjęcie za Planckiem, że energia promieniowania przekazywana jest w porcjach równych $h\nu$. Każdy pojedynczy elektron oddziałuje z jednym kwantem (fotonem), zatem aby elektron został uwolniony, porcja energii $h\nu$ musi być większa od pewnej wartości progowej, niezależnie od tego, jak wiele takich porcji znajduje się w wiązce.

Doświadczalne fakty dotyczące zachowania światła, czy też szerzej – fal elektromagnetycznych, ujawniają zasadniczy problem interpretacyjny. Z jednej strony światło wykazuje bezsporne cechy propagacji falowej – podlega zjawiskom interferencji i dyfrakcji, które są jednoznacznie związane z falowym charakterem. Z drugiej strony jednak przywołane wyżej odkrycia empiryczne pokazują, że w pewnych okolicznościach światło zachowuje się analogicznie do roju cząsteczek. Jak można pogodzić te wykluczające się wzajemnie modele? Jest to jedno z wielu wyzwań, przed którymi stanęła nowa teoria mikroświata.

Inne kontrowersyjne fakty zostały ujawnione podczas badania własności elementarnych składników materii, a konkretnie elektronów. Jednym z bardziej spektakularnych doświadczeń wykonanych we wczesnych latach dwudziestego wieku było doświadczenie Sterna-Gerlacha, badające oddziaływanie momentu magnetycznego elektronu z polem magnetycznym. Zgodnie z klasyczną teorią elektromagnetyzmu, obracające się ciała obdarzone ładunkiem elektrycznym wytwarzają pole magnetyczne – w skrócie zachowują się jak małe magnesy o określonym ustawieniu w przestrzeni, równoległym do osi obrotu. Ten efekt opisany jest przy pomocy wektora momentu magnetycznego, który jest funkcją zarówno prędkości rotacji (ściślej momentu pędu), jak i ładunku. Ciało obdarzone momentem magnetycznym, poruszające się w niejednorodnym polu magnetycznym, dozna działającej siły równoległej do kierunku linii sił zewnętrznego pola. Wartość tej siły zależy od ustawienia wektora momentu magnetycznego w stosunku do linii sił: jest ona największa, kiedy moment magnetyczny jest ustawiony równolegle, a spada do zera dla kierunku prostopadłego. Zatem przepuszczając strumień cząstek o losowo ustawionych momentach magnetycznych przez odpowiednio ukształtowane pole magnetyczne, powinniśmy uzyskać całe spektrum odchyleń: od wartości maksymalnej „w górę", przez wartość zerową, do maksymalnej „w dół".



Rys. 7.2. Doświadczenie Sterna-Gerlacha

Celem doświadczenia Sterna-Gerlacha było zbadanie interakcji momentu magnetycznego elektronu z polem magnetycznym. W tym celu zastosowano nie wolne elektrony, a atomy srebra, które posiadają jeden „niesparowany" elektron na najwyższej orbicie. Ze względu na „kasowanie się" momentów pędu pozostałych elektronów całkowity wewnętrzny moment pędu atomu srebra jest równy momentowi pojedynczego elektronu. W dalszym opisie będziemy jednak upraszczająco mówić o pojedynczych elektronach. Strumień elektronów przepuszczono zatem pomiędzy biegunami odpowiednio ukształtowanego magnesu, rejestrując ich odchylenie na ekranie (kliszy fotograficznej). Uzyskany rezultat różnił się znacząco od przewidywań klasycznych – zamiast ciągłego rozkładu zaobserwowano jedynie dwa mak-

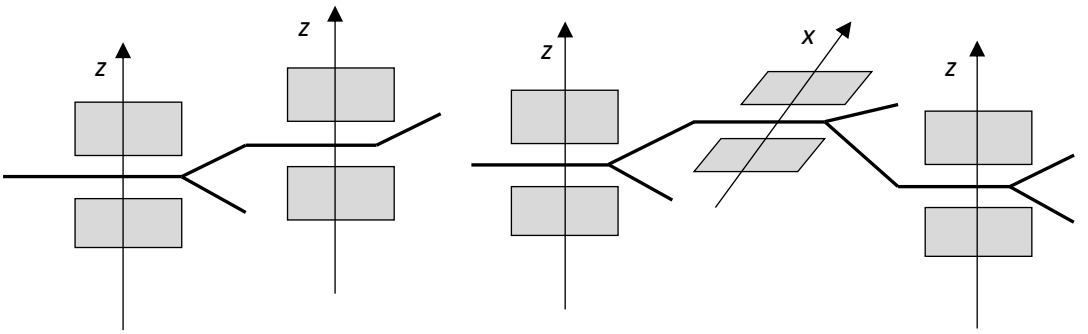
symalne odchylenia „do góry" i „w dół" (rys. 7.2). Wygląda to tak, jakby wszystkie elektrony ustawiły swoje momenty magnetyczne równolegle do linii sił pola, w górę lub w dół. Wydaje się to jednak absurdalne. Dodatkowo, powtarzanie eksperymentów dla różnych orientacji magnesu prowadzi do dokładnie takiego samego rezultatu. Jednakże jest matematyczną niemożliwością, aby składowe danego wektora we wszystkich kierunkach przyjmowały taką samą, maksymalną wartość.

Pierwszy wniosek wyciągnięty z doświadczenia Sterna-Gerlacha był taki, że wewnętrzny moment pędu elektronu jest wielkością nieklasyczną, której nie można opisać w języku wektorów w trójwymiarowej przestrzeni fizycznej. Dla podkreślenia nowego, nieklasycznego charakteru, moment pędu cząstek określa się mianem spinu (*to spin* – obracać się). Dokładną matematyczną analizę spinu podał Wolfgang Pauli (przy pomocy macierzy – szczegóły w paragrafie z gwiazdką), a później Paul Dirac (za pomocą pojęcia spinorów). Co więcej, spin wykazuje cechę zwaną *skwantowaniem*. Oznacza to, że dopuszczalne są tylko szczególne, dyskretne wartości dla danej wielkości, a nie ich ciągłe spektrum, jak przy wielkościach klasycznych. W wypadku spinu elektronów dla danego kierunku istnieją dokładnie dwie dopuszczalne wartości, równe $\frac{1}{2}\hbar$ oraz $-\frac{1}{2}\hbar$ (taki spin nazywa się często połówkowym). Inne cząstki posiadają więcej możliwych wartości składowych spinu.[1] Skwantowanie, czy też dyskretyzacja wielkości fizycznych jest unikatową cechą mechaniki kwantowej. Charakteryzuje ona wiele innych wielkości, takich jak orbitalny moment pędu czy też energia w układach związanych. Skwantowanie poziomów energetycznych w atomie (w szczególności w atomie wodoru) dostarcza wyjaśnienia innego ważnego faktu obserwacyjnego, mianowicie istnienia dyskretnych linii widmowych w promieniowaniu charakterystycznym dla danego rodzaju atomu.

Jednakże to wszystko nie wyczerpuje przełomowego charakteru doświadczenia Sterna-Gerlacha. Dodatkowy fascynujący aspekt świata kwantowego został ujawniony przy wykonaniu sekwencyjnych pomiarów spinu wzdłuż różnych kierunków. Zacznijmy może od prostego przypadku, w którym ustawione jeden po drugim zostały dwa magnesy o tej samej orientacji (rys. 7.3 po lewej stronie). Po przejściu elektronów przez pierwszy magnes jedna z wiązek wyjściowych (np. dolna) została zablokowana, a górna skierowana do drugiego magnesu. Wynik takiego eksperymentu jest zgodny z oczekiwaniami – na wyjściu z drugiego magnesu pojawia się jedynie górna wiązka. Interpretacja tego rezultatu wydaje się oczywista. Górna wiązka opuszczająca pierwszy magnes zawiera tylko elektrony o spinie skierowanym do góry, a zatem drugi pomiar nie może ujawnić elektronów o przeciwnie skierowanym spinie. Jak na razie wszystko przebiega zgodnie z oczekiwaniami. Zobaczmy jednak, co się stanie, jeśli ustawimy trzy magnesy jeden za drugim w taki sposób, że pierwszy i trzeci będą miały taką samą orientację, a drugi będzie zorientowany do nich prostopadle (rys. 7.3 po prawej). Dla ustalenia uwagi przyjmijmy, że osie pierwszego i trzeciego magnesu są równoległe do osi z, a drugiego do osi x. Zatem mamy do czynienia z sekwencyjnym pomiarem spinów s_z, s_x, a następnie znowu s_z. Wiązka górna wychodząca z pierwszego magnesu zostaje ponownie rozszczepiona na dwie wiązki („lewą" i „prawą") po przejściu magnesu środko-

[1] Zwróćmy uwagę, że wprowadza się tutaj nie jeden, a w istocie nieskończenie wiele spinów, każdy zdefiniowany względem jednego kierunku w przestrzeni. Co prawda określa się te spiny mianem „składowych", ale nie są to składowe normalnego wektora. Natomiast można dodatkowo zdefiniować tzw. spin całkowity, którego kwadrat jest sumą kwadratów trzech składowych w trzech prostopadłych kierunkach: $s_x^2 + s_y^2 + s_z^2$. Jak się okazuje, dostajemy wtedy wielkość, która przyjmuje tylko jedną wartość dla wszystkich elektronów, równą $\frac{3}{4}\hbar^2$.

wego. Jest to intuicyjne – w obrębie cząstek o spinie s_z zwróconym do góry powinno się znaleźć mniej więcej tyle samo cząstek o spinie s_x skierowanym w jedną stronę, co i w drugą. Jeśli natomiast jedna z tych wiązek przejdzie ponownie przez magnes ustawiony wzdłuż osi z, otrzymamy zaskakujący rezultat: pojawią się znowu dwie wiązki, górna i dolna, o porównywalnej intensywności.



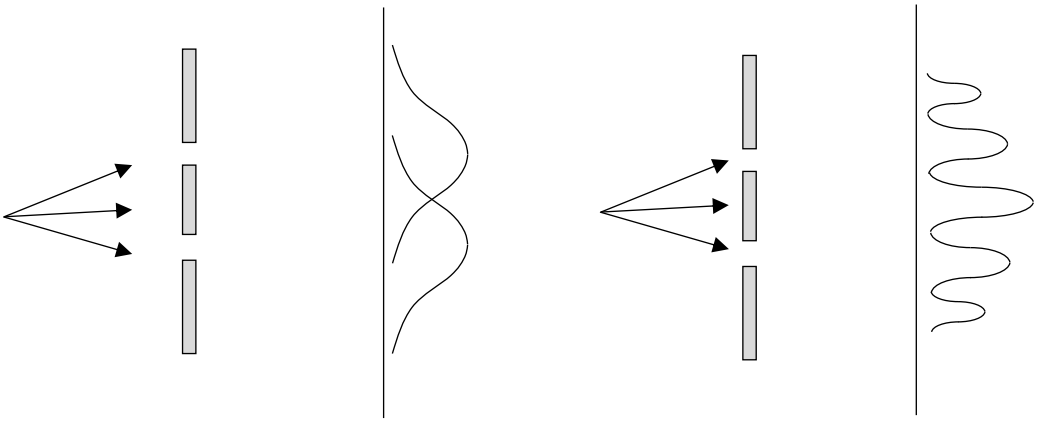
Rys. 7.3. Podwójne i potrójne doświadczenie Sterna-Gerlacha

Jest to zastanawiające. Cząstki wlatujące do drugiego magnesu mają spin s_z równy $\frac{1}{2}\hbar$. Po wykonaniu pomiaru spinu s_x i wyselekcjonowaniu jednej z dwóch opcji otrzymujemy cząstki o spinie $-\frac{1}{2}\hbar$ wzdłuż osi x. Jednakże powtórne sprawdzenie spinu s_z daje znów dwie możliwe wartości. Wygląda to tak, jakby w wyniku drugiego pomiaru połowa elektronów „odwróciła" swoje uprzednio zmierzone spiny. Takie zachowanie trudno wytłumaczyć z klasycznego punktu widzenia. Nie znamy żadnego mechanizmu odpowiedzialnego za odwrócenie momentów magnetycznych. Wyjaśnienie zaproponowane na gruncie nowej teorii jest takie, że pewne wielkości fizyczne wzajemnie się „kasują" w takim sensie, że dokładne wyznaczenie wartości jednej z nich zaburza wartość drugiej. Wielkości takie nazywa się „niekompatybilnymi" – zaliczają się do nich składowe spinu w różnych kierunkach, a także pęd i położenie. Dodatkowo zwróćmy uwagę, że proponowana analiza sytuacji *implicite* zakłada, że dana wielkość fizyczna (np. spin) może w danym momencie nie przyjmować żadnej dobrze określonej wartości. Cząstki opuszczające drugi magnes nie są w żaden sposób określone pod względem spinu wzdłuż osi z. Dopiero kolejny pomiar „tworzy" odpowiednie wartości dla poszczególnych elektronów (przy okazji jednak „niszcząc" uzyskaną wcześniej wartość spinu s_x).

Przegląd eksperymentalnych faktów ujawniających nieklasyczne zachowanie składników mikroświata zakończymy doświadczeniem z dwiema szczelinami. To w pewnym sensie eksperyment myślowy, którego praktyczna realizacja jest dużo bardziej złożona,[2] ale dla celów ilustracyjnych wykorzystuje się go właśnie w tej mocno wyidealizowanej formie. Rozważmy zatem przesłonę z dwiema szczelinami oraz umieszczony za nią ekran (rys. 7.4). Jest to typowe ustawienie dla eksperymentu interferencyjnego z falami – np. możemy taki eksperyment łatwo przeprowadzić na wodzie, ustawiając odpowiednio przygotowaną barierę

[2] Najbliższym tej idealizacyjnej wersji jest tzw. doświadczenie Davissona-Germera z rozpraszaniem elektronów na siatkach krystalicznych.

z dwoma otworami i tworząc falę na powierzchni wody. Obecnie doświadczenie będzie jednak dotyczyło nie fal, a cząstek, np. elektronów. W pierwszym etapie zasłaniamy jedną z dwóch szczelin i wysyłamy wiązkę elektronów w kierunku bariery z jedną otwartą szczeliną. Uzyskamy w ten sposób na ekranie gładki rozkład padających elektronów z maksimum wypadającym naprzeciwko szczeliny (rys. 7.4). Jest to normalne zachowanie cząsteczek, które podążają po liniach prostych od szczeliny w pewnym zakresie kierunków, ze względu na możliwe niewielkie odbicia od krawędzi szczeliny. Podobny rozkład uzyskamy przy otwartej drugiej szczelinie i zamkniętej pierwszej. Co się jednak stanie, kiedy otworzymy obie szczeliny?



Rys. 7.4. Doświadczenie z dwiema szczelinami. Rozkład elektronów przechodzących osobno przez pojedyncze szczeliny (po lewej) i przez obie otwarte szczeliny (po prawej)

Okazuje się, że w takiej sytuacji dystrybucja elektronów na ekranie będzie zasadniczo odmienna od poprzednich rozkładów (czy też ich sumy). Pojawią się mianowicie dobrze znane z efektów falowych maksima i minima natężeń (tzw. prążki interferencyjne). Oznacza to, że elektrony „preferują" pewne kierunki, a w innych rozpraszają się mniej chętnie. Efekt ten pokazuje, że materia na elementarnym poziomie wykazuje własności falowe, analogiczne do własności światła. Podobnie bowiem jak światło, które w niektórych wypadkach przejawia cechy falowe, a w innych cząsteczkowe (np. w efekcie fotoelektrycznym), tak i elektrony ujawniają swoiście „dualną" naturę. Jednakże kwestia tzw. dualizmu korpuskularno-falowego jest ściśle związana z interpretacjami mechaniki kwantowej – w różnych interpretacjach przyjmuje ona różną formę. Na przykład w mechanice Bohmowskiej cząstki są całkowicie klasycznymi obiektami, mającymi dobrze określone trajektorie, natomiast efekty falowe są tłumaczone istnieniem dodatkowego „pola" reprezentowanego matematycznie przez tzw. funkcję falową, które interferuje ze sobą i jednocześnie wpływa na zachowanie pojedynczych cząstek. Z kolei zgodnie z hipotezą fal materii francuskiego fizyka Ludwiga de Broglie'a, elektrony „naprawdę" są paczkami fal, gdzie rozkład danej fali jest rozkładem gęstości materii. Zostawmy jednak na razie kwestie interpretacyjne i spróbujmy przeanalizować uzyskane rezultaty pomiarowe w języku cząstek.

Pierwsze możliwe wyjaśnienie zachowania elektronów sugeruje, że być może efekty „interferencyjne" pojawiają się wskutek wzajemnego oddziaływania cząstek w strumieniu przechodzącym przez szczeliny. Niewykluczone, że elektrony przechodzące przez jedną szcze-

linę zderzają się z elektronami z drugiej szczeliny, co prowadzi do powstania rozkładu odmiennego od sumy dwóch rozkładów dla każdej szczeliny z osobna. Aby zweryfikować tę hipotezę, wykonano podobne doświadczenie z bardzo słabymi wiązkami, gdzie w praktyce każdy elektron przechodził indywidualnie przez barierę. Efekt pozostał niezmieniony – nadal po zebraniu odpowiedniej liczby zarejestrowanych elektronów pojawiał się znajomy obraz prążków interferencyjnych. Zatem to nie wzajemne oddziaływanie jest odpowiedzialne za efekty falowe. Możemy więc rozważyć zachowanie pojedynczego elektronu. Fizyka klasyczna mówi nam, że elektron musi przejść przez jedną z dwóch szczelin – albo pierwszą, albo drugą. Jednakże każda z dwóch dróg prowadzi do gładkiego rozkładu, bez maksimów i minimów. Prowadzi to do zaskakującego, ale trudnego do uniknięcia wniosku, że elektron nie przechodzi przez żadną ze szczelin osobno, ale w pewnym sensie przez dwie szczeliny na raz.

Czy jednak nie da się poddać tej hipotezy testowi empirycznemu? Możemy np. oświetlić jedną ze szczelin strumieniem skoncentrowanego światła tak, aby zlokalizować przechodzący przez nią elektron. W takiej sytuacji będziemy mogli każdorazowo się przekonać, czy dany elektron zarejestrowany na ekranie przeszedł przez pierwszą czy drugą szczelinę. Zarówno zebrany bogaty materiał empiryczny, jak i teoretyczne rozważania jednoznacznie wskazują, że w takiej sytuacji rozkład elektronów będzie zupełnie klasyczny, tj. nie pojawią się żadne efekty interferencyjne. Wystąpienie tych efektów jest ściśle skorelowane z naszą zasadniczą niewiedzą co do tego, którą drogę wybrał elektron. Standardowo fakt ten przedstawia się za pomocą pojęcia superpozycji. W momencie przejścia przez barierę położenie elektronu nie jest dobrze określone – nie znajduje się on ani w pobliżu jednej szczeliny, ani drugiej. Możemy mówić jedynie o pewnym prawdopodobieństwie znalezienia elektronu w danym obszarze. Taki stan określa się mianem superpozycji dwóch stanów, z których każdy ma dobrze określoną lokalizację. Superpozycja stanów „redukuje się" do jednego stanu podczas pomiaru, np. przy oświetleniu jednej z dwóch szczelin promieniowaniem. Zatem pomiar nieuchronnie prowadzi do zmiany stanu obserwowanego obiektu.

7.2. Stany kwantowe, prawdopodobieństwa, superpozycje

Droga od danych empirycznych do pełnej teorii kwantowej nie była prosta. W celu wyjaśnienia obserwowanych zjawisk proponowano różnego rodzaju „półklasyczne" modele, jak np. znany model atomu Bohra-Sommerfelda. Rozwiązania te były jednak połowiczne. Do pracy nad podstawami zupełnie nowej teorii włączyli się zarówno fizycy (Werner Heisenberg, Wolfgang Pauli, Niels Bohr, Erwin Schrödinger), jak i matematycy (Johann von Neumann). W niniejszym paragrafie przedstawimy w najogólniejszych zarysach podstawowe pojęcia tej teorii, podkreślając różnice w stosunku do fizyki klasycznej. W paragrafie oznaczonym gwiazdką przyjrzymy się bliżej matematycznej stronie nowego formalizmu. Zacznijmy od przypomnienia pojęcia stanu układu fizycznego z mechaniki klasycznej. Jak pamiętamy, szczegółowy opis układu N klasycznych cząstek wymaga podania ich momentalnych położeń oraz prędkości (pędów). Mając te dane, możemy obliczyć wszystkie inne wielkości charakteryzujące układ w danej chwili, np. całkowitą energię kinetyczną. Zatem ogólnie możemy przyjąć, że kompletny stan układu jest dany w postaci funkcji, która każdemu fizycznemu parametrowi przypisuje dokładną wartość. W przeciwieństwie do mechaniki klasycznej, stan układu w teorii kwantowej operuje pojęciem prawdopodobieństwa. W danej chwili określony

jest rozkład prawdopodobieństwa dla wszystkich możliwych wartości każdej wielkości fizycznej.

Stany kwantowe reprezentujemy matematycznie przy pomocy wektorów w pewnej abstrakcyjnej przestrzeni, zwanej przestrzenią Hilberta. Szczegóły nie są w tej chwili istotne – ważne tylko, aby pamiętać, że na wektorach określone są pewne podstawowe operacje matematyczne, takie jak dodawanie wektorowe, mnożenie przez liczbę oraz iloczyn skalarny wektorów, który daje pewną liczbę. Dzięki wprowadzeniu iloczynu możemy mówić o bardzo ważnym pojęciu prostopadłości wektorów – w języku matematyki nazywa się to ortogonalnością (dwa wektory są ortogonalne, gdy ich iloczyn jest równy zeru). Pojedynczy wektor zawiera w sobie „ukrytą" informację na temat wszystkich wchodzących w grę prawdopodobieństw. Można je odtworzyć korzystając z pewnej zasady zwanej regułą Borna[3]. Nie będziemy tej reguły na razie formułować ogólnie, a jedynie dla szczególnego przypadku. Zacznijmy od zaznaczenia, że chociaż w ogólnym wypadku stan kwantowy przypisuje wartościom danej wielkości „nietrywialne" prawdopodobieństwa (tj. prawdopodobieństwa różne od 0 i 1), to jednak w pewnych okolicznościach możliwe jest, aby jedna wartość uzyskała prawdopodobieństwo równe 1, podczas gdy pozostałe 0. Oznacza to, że dana wielkość (energia, pęd czy spin) ma w tym stanie dobrze określoną wartość. Taki stan nazywa się stanem własnym tej wielkości. Oczywiście dla każdej możliwej wartości tej wielkości stan własny będzie inny (wektory reprezentujące takie stany są do siebie ortogonalne).

Rozważmy teraz dwa stany własne dla pewnej wielkości, odpowiadające dwóm różnym wartościom pomiarowym. Jako przykład zastosujmy omówiony w poprzednim paragrafie spin w pewnym kierunku. Niech z_+ oznacza stan, w którym spin cząstki wzdłuż pewnej osi z jest skierowany do góry, a z_- stan ze spinem w dół. Podstawowa zasada formalizmu kwantowego, zawarta w założeniu o wektorowym charakterze stanów, stwierdza, że dowolne dwa stany można ze sobą „łączyć" (reprezentujemy tę operację dodawaniem). Otrzymamy w ten sposób nowy stan, którego fizyczną interpretacją obecnie się zajmiemy. Ogólna formuła na liniową kombinację dwóch wektorów z_+ i z_- jest następująca:

$$az_+ + bz_-,$$

gdzie a i b to pewne liczby.[4] Formuła ta opisuje stan, w którym prawdopodobieństwo uzyskania rezultatu „spin do góry" wynosi $|a|^2$, a prawdopodobieństwo rezultatu „spin do dołu" jest równe $|b|^2$. Oczywiście, aby zachować zasadę, że prawdopodobieństwo zdarzenia pewnego jest równe jedności, musimy założyć, że $|a|^2 + |b|^2 = 1$, co określa się mianem „normalizacji" wektora (jest to równoważne ze stwierdzeniem, że długość wektora liczona iloczynem skalarnym z samym sobą wynosi jeden). Zatem powyżej zapisany wektor określa nowy, nieznany w fizyce klasycznej stan, w którym wartości spinu charakteryzują się jedynie pewnymi prawdopodobieństwami. Taki stan określa się często mianem superpozycji stanów z dobrze określonym spinami.

[3] Max Born – niemiecko-brytyjski fizyk i matematyk pierwszej połowy dwudziestego wieku. Jako ciekawostkę podam, że Born był dziadkiem ze strony matki niedawno zmarłej piosenkarki i aktorki, Olivii Newton-John.

[4] W ogólności będą to liczby zespolone, dlatego też podnosząc je do kwadratu, musimy wziąć tzw. moduł.

Szczególnym przypadkiem superpozycji jest sytuacja, kiedy oba superponowane stany wchodzą do niej z równymi prawdopodobieństwami. W takiej sytuacji powyższy wektor przyjmuje postać „symetrycznej" kombinacji:

$$\frac{1}{\sqrt{2}}(z_+ + z_-).$$

Rozważane w poprzednim paragrafie „potrójne" doświadczenie Sterna-Gerlacha może być opisane właśnie przy pomocy analogicznej superpozycji. Elektrony w górnej wiązce wychodzącej z pierwszego magnesu znajdują się w stanie symetrycznej superpozycji stanów „spin w kierunku *x* do góry" i „spin w kierunku *x* do dołu", co pokazuje rezultat drugiego pomiaru, gdzie dokładnie połowa elektronów odchyliła się w górę, a połowa w dół.[5] *Nota bene* dokładnie ten sam stan, będący superpozycją spinów w kierunku *x*, można opisać jako stan, w którym spin w kierunku *z* jest dobrze określony. Pokazuje to, że istnieją mocne związki między spinami w różnych kierunkach. Określony stan spinu w dowolnym kierunku jednoznacznie określa rozkłady prawdopodobieństw dla spinów w każdym innym kierunku.

Musimy odróżnić stan superpozycji od statystycznej mieszaniny złożonej z takiej samej liczby cząstek o spinie do góry i spinie w dół. Choć w takiej mieszaninie prawdopodobieństwo tego, że losowo wybrana cząstka będzie miała spin wzdłuż osi *x* skierowany do góry wynosi ½, to jednak prawdopodobieństwa wartości spinów wzdłuż innych osi będą inne niż w stanie superpozycji: będą one mianowicie równe ½ dla każdej osi. Natomiast w stanie superpozycji spinów, jak już zauważyliśmy, zawsze istnieje oś, wzdłuż której prawdopodobieństwo uzyskania danego rezultatu jest równe jeden. Stany mieszane reprezentowane są w formalizmie mechaniki kwantowej nie za pomocą wektorów, a za pomocą pewnego rodzaju operatorów zwanych operatorami gęstości (lub operatorami statystycznymi).

Innego przykładu superpozycji dostarcza poprzednio opisany przypadek doświadczenia z dwiema szczelinami. W trakcie przechodzenia przez barierę stan elektronu opisany jest superpozycją dwóch funkcji falowych – jednej opisującej przejście przez pierwszą szczelinę, a drugiej przez drugą szczelinę. Ten matematyczny fakt wyjaśnia także, dlaczego na ekranie pojawiają się efekty interferencyjne. Ponieważ funkcja falowa jest „wektorem" opisującym położenie danej cząstki (dokładniej tzw. gęstość prawdopodobieństwa znalezienia cząstki), aby obliczyć prawdopodobieństwo znalezienia cząstki w danym obszarze należy wziąć kwadrat tej funkcji i scałkować (zsumować) go po tym obszarze. Skoro kwadrat sumy liczb nie jest równy sumie kwadratów, prawdopodobieństwo znalezienia elektronu w danym miejscu na ekranie nie będzie prostą sumą prawdopodobieństwa znalezienia go przy założeniu przejścia przez jedną szczelinę i prawdopodobieństwa przy założeniu przejścia przez drugą.

Ontologicznymi aspektami superpozycji zajmiemy się bliżej przy okazji analizy paradoksu kota Schrödingera. Na razie zwróćmy uwagę na istotny fakt dotyczący pomiarów na układach przygotowanych w stanach superpozycji. Skoro przed pomiarem np. spinu układ

[5] Można przypuszczać, że elektrony wchodzące do pierwszego magnesu Sterna-Gerlacha znajdują się również w stanie superpozycji spinu wzdłuż osi *z* „do góry" i „do dołu", skoro dokładnie połowa z nich odchyliła się w górę, a połowa w dół. Niestety, jest to przypuszczenie błędne. Wyjściowy stan elektronów należy do innej kategorii tzw. stanów mieszanych (por. wpis w ramce). Aby przejść z takiego stanu do „zwykłego" (zwanego stanem czystym), należy wykonać pomiar spinu w dowolnym kierunku. Takie pomiary nazywa się często „przygotowującymi".

znajdował się w superpozycji stanów z wartościami „do góry" i „do dołu", a po pomiarze każdy elektron znajduje się w jednej z dwóch wiązek z dobrze określonym spinem, to naturalne jest przyjąć, że w rezultacie pomiaru elektrony zmieniły swój wyjściowy stan $az_+ + bz_-$ na jeden z dwóch stanów z_+ lub z_-. Zmiana taka jest procesem probabilistycznym – prawdopodobieństwo, że elektron „przeskoczy" do stanu z_+ wynosi $|a|^2$, a tego, że przeskoczy do z_- – $|b|^2$. W języku mechaniki kwantowej taki proces nazywa się „kolapsem" lub „redukcją" wyjściowego stanu. Jak się okaże, interpretacja kolapsu pomiarowego w mechanice kwantowej jest jednym z najbardziej kontrowersyjnych i najzagorzalej dyskutowanych zagadnień.

7.3. Wielkości pomiarowe i zasada nieoznaczoności

Nawet przy pobieżnym kontakcie z fizyką kwantową musimy w pewnym momencie natknąć się na słynną zasadę nieoznaczoności Heisenberga. W pierwotnej wersji zasada ta była wyprowadzona z falowego ujęcia procesów kwantowych. Z klasycznej analizy zjawisk falowych wiemy, że każdą tzw. paczkę falową (czyli falę, która zajmuje pewne dobrze określone położenie w przestrzeni) można przedstawić w postaci sumy („superpozycji") fal sinusoidalnych o różnych częstotliwościach oraz liczbach falowych (odwrotność długości fali). Można udowodnić, że szerokość danej paczki jest odwrotnie proporcjonalna do „rozmycia" liczb falowych w rozkładzie tej paczki na sinusoidalne składowe (tzw. harmoniczne) – czyli im lepiej zlokalizowana w przestrzeni paczka, tym szerszy zakres różnych liczb falowych dla harmonicznych składników. Ponieważ jednak szerokość paczki falowej powiązanej z daną cząstką (fotonem czy elektronem) jest miarą nieokreśloności jej położenia w przestrzeni, a liczba falowa wyznacza pęd cząstki-fali, mamy stąd zależność, że rozmycie, czyli nieokreśloność położenia, jest odwrotnie proporcjonalne do nieokreśloności pędu. Innymi słowy, im lepiej określony jest pęd cząstki, tym bardziej nieokreślone jest jej położenie (i *vice versa*). Zasadę tę ujmuje się w znanej nierówności:

$$\Delta x \Delta p \geq \hbar,$$

gdzie Δx i Δp oznaczają niepewność pomiarową związaną z określeniem odpowiednio położenia i pędu, a $\hbar$ jest stałą Plancka podzieloną przez 2π.

Okazuje się, że powyższa zależność między pędem i położeniem jest tylko jednym z wielu przykładów powszechnego w mechanice kwantowej zjawiska *niekompatybilności* między pewnymi mierzalnymi wielkościami. Jak już wspominaliśmy we wcześniejszym paragrafie, dotyka ono także m.in. składowych spinu w różnych kierunkach. Formalizm mechaniki kwantowej dostarcza eleganckiego sposobu na wyrażenie owej zależności. Aby to jednak omówić, musimy sprecyzować, jak w aparacie pojęciowym teorii kwantowej należy matematycznie reprezentować wielkości mierzalne (które w tej teorii zwane są również obserwablami). Wiemy już, że dla danej obserwabli można wyróżnić pewne szczególne wektory zwane wektorami własnymi. Określają one stany, w których ta obserwabla ma pewną dobrze określoną wartość. Wspomnieliśmy też, że wektory własne odpowiadające różnym wartościom tej samej wielkości są wzajemnie ortogonalne. Wynika to z oczywistego faktu, że jeśli układ charakteryzuje się daną wartością określonej obserwabli, to prawdopodobień-

stwo uzyskania w pomiarze innej wartości jest oczywiście zerowe.[6] Przyjmijmy teraz dla uproszczenia, że wybrana obserwabla posiada skończoną liczbę różnych wartości (jak np. spin połówkowy z dwiema możliwymi wartościami). W takim wypadku obserwabla ta pozostaje jednoznacznie skorelowana z układem wzajemnie prostopadłych wektorów własnych w przestrzeni Hilberta (taki układ wektorów jest też zwany „spektrum" danej obserwabli). Zwróćmy uwagę, że każdy wektor nienależący do spektrum może być przedstawiony jako kombinacja (superpozycja) wektorów własnych. W stanie reprezentowanym przez ten wektor dana obserwabla nie posiada więc określonej wartości; istnieją jedynie prawdopodobieństwa, że wartości będą odpowiednie.

Jeśli teraz wybierzemy jakąś inną obserwablę, ona również będzie reprezentowana przez zestaw prostopadłych wektorów, ale w ogólności innych niż poprzednie. Co oznacza fakt, że dwie obserwable A i B mają różne wektory własne? Niech v_a będzie wektorem własnym dla obserwabli A, odpowiadającym pewnej wartości a. Załóżmy, że v_a nie jest wektorem własnym dla B. W takim razie układ znajdujący się w stanie v_a ma dobrze zdefiniowaną wartość dla A, ale nie dla B. Czyli wielkość B jest w tym stanie „rozmyta". Jest to dokładnie sytuacja opisana w zasadzie nieoznaczoności Heisenberga – dobrze określone położenie, źle określony pęd. Obserwable, dla których nie wszystkie wektory własne się pokrywają, nazywamy niekompatybilnymi. Na odwrót, jeśli dwie obserwable mają dokładnie takie same wektory własne, określamy je jako kompatybilne.

Istnieje jeszcze bardziej elegancki i „syntetyczny" sposób na przedstawienie obserwabli i ich relacji kompatybilności. Dokładniej opowiemy o nim w paragrafie z gwiazdką. Obecnie wspomnimy jedynie, że standardową reprezentacją wielkości mierzalnych w mechanice kwantowej są obiekty zwane operatorami liniowymi na wektorach (są to funkcje transformujące jedne wektory w inne i spełniające odpowiednie warunki liniowości). Okazuje się, że z każdym zestawem ortogonalnych wektorów własnych i odpowiadających im wartości (są one zwane wartościami własnymi) skorelowany jest pewien operator liniowy. Jego działanie na dowolnym wektorze jest następujące: najpierw rozkładamy ten wektor na składowe równoległe do wektorów własnych, następnie każdą taką składową mnożymy przez odpowiednią wartość własną i sumujemy z powrotem tak otrzymane wektory. Wydaje się to dość skomplikowane i mało czytelne, ale taka procedura prowadzi do prostych i eleganckich reprezentacji wielu fizycznych faktów i prawidłowości. W szczególności istnieje bardzo zgrabny sposób na wyrażenie relacji kompatybilności. Dwie obserwable są kompatybilne, jeśli odpowiadające im operatory komutują, czyli ich działanie jest takie samo, niezależnie od kolejności wykonania: $AB = BA$. Jak się okazuje, operatory reprezentujące różne składowe spinu nie komutują ze sobą, podobnie jak operatory pędu i położenia.

[6] Zgodnie z regułą Borna, omówioną powyżej, prawdopodobieństwo rezultatów pomiaru wielkości A w określonym stanie obliczamy, rozkładając wektor tego stanu na składowe odpowiadające poszczególnym wartościom A, a następnie biorąc kwadraty współczynników tego rozkładu. Ponieważ wektor prostopadły do danego nie wchodzi do rozkładu tego wektora, prawdopodobieństwo uzyskania rezultatu z nim związanego musi być zerowe.

7.4. Argument EPR i niekompletność mechaniki kwantowej

Z powyższego szkicu wyłania się obraz nowej teorii, która opiera się na nieusuwalnym użyciu prawdopodobieństwa, ograniczającym możliwość dokładnego i kompletnego poznania mikroświata. Taka postać teorii naukowej była nie do zaakceptowania dla Einsteina. Był on w pełni świadomy ważnej roli, jaką prawdopodobieństwo i rozważania statystyczne odgrywają w nauce. Metody statystyczne stosujemy, gdy mamy do czynienia z wielką liczbą obiektów, których nie jesteśmy w stanie śledzić indywidualnie. Z podobną sytuacją zetknęliśmy się przy okazji statystycznej analizy zjawisk termodynamicznych, gdzie musieliśmy posługiwać się np. uśrednionymi wielkościami dla ogromnego zbioru cząstek. Jednakże zakładaliśmy, że każda poszczególna cząstka charakteryzuje się jednoznacznymi parametrami. Natomiast w wypadku teorii kwantowej prawdopodobieństwa pomiarowe stosujemy bezpośrednio do indywidualnych obiektów, przyjmując, że przed pomiarem obiekty te nie są w pełni scharakteryzowane fizycznymi wielkościami. Takie podejście budziło głęboki sprzeciw Einsteina.

W swoich polemikach ze zwolennikami nowej teorii (głównie z Bohrem), próbował pokazać, że teoretycznie możliwe jest pokonanie ograniczeń nakładanych na możliwość poznania dokładnych własności układów kwantowych (w tym tych wynikających z zasady nieoznaczoności). Próby te były zwykle kontrowane niezwykle subtelnymi replikami Bohra. Najsłynniejszy argument w tej polemice został opublikowany w kluczowej pracy z 1935 r. autorstwa Einsteina i jego dwóch współpracowników: Borisa Podolsky'ego i Nathana Rosena. Argument ten znany jest do dzisiaj pod nazwą EPR od nazwisk autorów. Jego główną ideę można przedstawić w postaci następującej historyjki. Wyobraźmy sobie, że pewien sklep zaczyna sprzedawać zabawki reklamowane jako „magiczne pudełka". Są to zwykłe niewielkie pudełka, które po otwarciu ujawniają zawartość – czarną lub białą kulkę. Nikt ze sprzedawców ani producentów nie wie, jakiego koloru kulkę znajdziemy w pudełku. Porównując ze sobą rezultaty dużej liczby otwartych pudełek dochodzimy do wniosku, że statystycznie szanse ujawnienia czarnej lub białej kulki są jednakowe. Co więcej, sprzedawcy zapewniają nas, że nie ma żadnej możliwości dowiedzieć się przed otwarciem, jaki jest kolor kulki. Najlepiej wyobrażać sobie, że kulka „decyduje", jaki kolor przybrać dopiero w momencie otwarcia pudełka.

Pewnego dnia sklep wprowadza nowość: zamiast jednego sprzedaje dwa pudełka. Choć nadal nie wiadomo, jakiego koloru kulki ujrzymy po otwarciu, to jednak doświadczenie poucza nas, że w obu pudełkach kulki ujawniają przeciwne kolory. Nigdy się nie zdarzyło, aby w danej parze pojawiły się dwie czarne lub dwie białe kulki. Wyobraźmy sobie teraz, że pudełka z danej pary zostały ofiarowane dwóm osobom: Alicji i Robertowi. Alicja zabiera swoje pudełko w daleką podróż, po czym wiedziona ciekawością otwiera je i widzi białą kulkę. Natomiast Robert się powstrzymuje i nie dotyka w ogóle swojego pudełka. Oczywiście Alicja w momencie zobaczenia białej kulki wie już, że pudełko Roberta musi zawierać czarną kulkę. Jednakże dowiedziała się o tym bez potrzeby otwarcia drugiego pudełka! Co więcej, ze względu na dzielącą ją odległość od Roberta możemy przyjąć, że otwarcie jej pudełka nie mogło w żaden sposób wpłynąć na stan drugiego pudełka. Zatem wyciągamy wniosek, że kulka Roberta nawet przed otwarciem pudełka musiała być koloru czarnego. Cała historia z magicznymi pudełkami okazała się oszustwem (co zresztą podejrzewaliśmy od samego początku). Każde pudełko po prostu zawiera czarną lub białą kulkę, a pudełka „czarne"

i „białe” są losowo mieszane w jednakowej proporcji, aby sprawić wrażenie tajemniczego i przypadkowego wybierania koloru w momencie otwarcia.

Możemy się domyślić, że kulki z naszej historyjki to obiekty kwantowe (np. elektrony), a pudełka i ich otwieranie reprezentują pomiary. Zgodnie z założeniem promowanym przez zwolenników nowej teorii kwantowej, przed pomiarem cząstka kwantowa nie posiada cechy, która zostaje w nim ujawniona (np. elektrony przed wejściem do urządzenia Sterna-Gerlacha nie mają określonego spinu). Pozostaje jednak pytanie, co w teorii kwantowej odpowiada parom pudełek zawierającym kulki o przeciwnym kolorze. Czy możliwe jest, żeby wybrać parę cząstek o idealnie skorelowanych własnościach bez jednoczesnego dowiedzenia się, jakie są te własności? Okazuje się, że mechanika kwantowa dopuszcza istnienie takich układów. Nazywamy je układami splątanymi.

Splątanie kwantowe to jedno z najbardziej fascynujących zjawisk, z jakimi zetknęła się fizyka mikroświata. Opiszmy je najpierw z teoretycznego punktu widzenia, a potem przedstawmy jego konkretne realizacje. Zacznijmy od pytania, jak wyrazić w języku mechaniki kwantowej stan układu składającego się z dwóch (lub więcej) obiektów, np. dwóch elektronów. W najprostszym wypadku, kiedy każda z cząstek z osobna posiada swój własny stan, stan całości jest prostą „kombinacją” tych dwóch stanów. W zasadzie to nic innego jak uporządkowana para: jeśli stan pierwszej cząstki reprezentowany jest przez wektor a, a stan drugiej przez wektor b, to łączny ich stan może być na przykład zapisany jako (a, b). W formalizmie kwantowo-mechanicznym taką parę przedstawia się w postaci pewnego iloczynu – nie zwykłego iloczynu skalarnego, a tzw. iloczynu tensorowego (ma to coś wspólnego z tensorami, o których mówiliśmy w poprzednich rozdziałach). Oznacza się go symbolem $\otimes$, choć niekiedy symbol ten się opuszcza, jak w wypadku zwykłego iloczynu dwóch liczb, który można, ale nie trzeba przedstawiać jako kropkę. Stan naszych dwóch cząstek zapiszemy zatem jako $a \otimes b$ lub w skrócie ab.

Należy pamiętać, że iloczyn tensorowy nie jest „odwracalny” (przemienny). Stan cząstek przedstawiony jako $b \otimes a$ jest różny od $a \otimes b$ – oznacza on, że pierwsza z cząstek ma własność b, a druga a.[7] Na razie wszystko wygląda zupełnie klasycznie. Pamiętajmy jednak, że w mechanice kwantowej obowiązuje zasada superpozycji – jeśli dwa wektory reprezentują pewne stany układu fizycznego, to dodając je, otrzymamy nowy stan, różniący się od każdego ze stanów z osobna. Możemy zatem rozważyć następujący wektor-stan:

$$\frac{1}{\sqrt{2}}(a \otimes b + b \otimes a).$$

Co reprezentuje taki wektor? Jest to stan, w którym z prawdopodobieństwem 50% pierwsza cząstka jest w stanie a, a druga w b, jak również z prawdopodobieństwem 50% pierwsza jest w stanie b, a druga w a. Stan żadnej z cząstek z osobna nie jest więc dokładnie określony, a jedynie dany z pewnym prawdopodobieństwem. Natomiast istnieje jednoznaczna korelacja między stanami cząstek – jeśli ujawnimy, że pierwsza cząstka znajduje się w stanie a, to druga musi być w stanie b. I na odwrót – jeśli pierwsza będzie w stanie b, druga nie ma

[7] W przypadku cząstek tego samego rodzaju, np. dwóch elektronów, które są w pewnym sensie nieodróżnialne, to założenie może być kwestionowane. Wrócimy do tego problemu w dalszych paragrafach. Na razie zakładamy milcząco, że każda z dwóch cząstek może być w jakiś sposób „zindywidualizowana” przez umieszczenie na nich etykietek „numer 1” i „numer 2”.

wyjścia tylko zająć stan *a*. Odnaleźliśmy więc w teorii kwantowej magiczne pudełka z opowiastki!

Fizyczna realizacja tej teoretycznej możliwości, którą wykorzystali Einstein, Podolsky i Rosen w swoim argumencie, obejmowała dwie niekompatybilne wielkości: pęd i położenie. Wykorzystali oni splątany stan dwóch cząstek, w którym ani pęd, ani położenie nie były dobrze określone, ale zarówno różnica ich położeń, jak i suma pędów miały ustalone wartości. Zatem mierząc pęd jednej z cząstek, można się dowiedzieć, jaki pęd posiada druga cząstka. Podobnie sytuacja wygląda w wypadku położenia. Niels Bohr zaproponował nawet bardzo pomysłową quasi-techniczną realizację stanu EPR za pomocą przesłony z dwiema szczelinami, umieszczonej na sprężynie, dzięki której przelatujące cząstki charakteryzowałyby się zarówno stałą różnicą położeń (równą odległości między szczelinami, jak i całkowitym pędem wzdłuż osi przesłony (równym co do wartości pędowi przekazanemu całej przesłonie). Jednakże praktyczna realizacja takiego układu jest zasadniczo niewykonalna.

Obecnie realizację splątanego stanu EPR ilustruje się przykładem elektronów ze spinami lub też fotonów i ich tzw. polaryzacji (ten drugi sposób jest najłatwiejszy do implementacji w laboratorium). Posłużmy się zatem raz jeszcze naszymi znajomymi stanami spinowymi z_+ (spin *z* do góry) i z_- (spin *z* w dół) w wybranym kierunku osi *z*. Rozważmy następujący stan dwóch elektronów, zwany singletowym:

$$\frac{1}{\sqrt{2}}(z_+ \otimes z_- - z_- \otimes z_+). \tag{7.1}$$

Znak minus w powyższym zapisie nie wpływa na prawdopodobieństwa znalezienia cząstek w odpowiednich stanach spinu w kierunku *z*, ponieważ i tak podnosimy współczynniki do kwadratu. Ma on natomiast fundamentalne znaczenie, jeśli chodzi o obliczanie prawdopodobieństw spinów wzdłuż innych osi (np. *x* czy *y*), ale dla naszych obecnych celów jest to nieistotne. Podobnie jak w poprzednio opisanej ogólnej sytuacji stanów *a* i *b*, mamy tutaj do czynienia ze ścisłą korelacją. Spiny obu cząstek nie są określone, ale jest pewne, że będą miały przeciwne wartości. Jakąkolwiek wartość ujawnimy na jednym z elektronów, drugi będzie posiadał wartość przeciwną.

Możemy teraz przedstawić rozumowanie Einsteina i jego współpracowników. Załóżmy, że elektrony przygotowane w powyższym stanie zostały rozdzielone na taką odległość, że żadne oddziaływanie fizyczne rozchodzące się z prędkością nie większą od prędkości światła nie może dotrzeć z jednego elektronu do drugiego (pamiętamy, że taka separacja nazywa się przestrzenno-podobna w szczególnej teorii względności). Następnie dokonujemy pomiaru spinu na jednej cząstce. Uzyskany wynik pozwala nam na precyzyjne przewidzenie wyniku pomiaru dla drugiego elektronu. Jednocześnie stan drugiego elektronu nie został w żaden fizyczny sposób zaburzony przez wykonany pomiar. Zatem jego stan byłby taki sam, nawet gdybyśmy nie poddali pierwszego elektronu pomiarowi. Wynika z tego jednoznaczny wniosek: odległy elektron ma ustalony spin w danym kierunku, nawet jeśli teoria kwantowa nie jest w stanie tego spinu przewidzieć. Mechanika kwantowa dostarcza zatem niekompletnego opisu rzeczywistości.[8]

[8] Jak się okazuje, powyższy stan elektronów (7.1) implikuje istnienie identycznych korelacji dla dowolnych kierunków spinu. Zatem analogiczne rozumowanie można powtórzyć dla każdego kierunku spinu, co pokazuje, że wszystkie spiny drugiej cząstki są jednoznacznie ustalone. Łamie to oczywiście zasadę nieoznaczoności (fakt niekompatybilności spinów).

Logiczna struktura argumentu EPR jest niezmiernie prosta. Korzystając z ogólnego założenia lokalności (istnieje górna granica prędkości oddziaływań fizycznych) oraz przewidywań mechaniki kwantowej (istnienie stanów splątanych), wyprowadzamy wniosek, że pewien obiekt fizyczny posiada własność, której nasza teoria nie jest w stanie przewidzieć. Jedyny logiczny sposób na uniknięcie wniosku o niekompletności mechaniki kwantowej to odrzucenie założenia lokalności. Jak się wydaje, taką drogę obrał Niels Bohr w swojej odpowiedzi na argument EPR. Jednakże uznał on, że nielokalność powiązania między odległymi cząstkami ma inny charakter niż zwykłe oddziaływanie fizyczne, jak np. oddziaływanie grawitacyjne czy elektromagnetyczne. Niestety Bohr był bardzo enigmatyczny, jeśli chodzi o szczegóły tego nielokalnego kwantowego powiązania (mówiąc o wpływie lokalnego pomiaru jedynie na *warunki sensowności* przypisania pewnych wartości drugiej cząstce). Do dzisiaj kwestia pozostaje mocno dyskusyjna. Natomiast w latach sześćdziesiątych ubiegłego stulecia sprawy przybrały zupełnie zaskakujący obrót za sprawą młodego i nikomu wcześniej nieznanego fizyka Johna Bella.

7.5. Twierdzenie Bella i nielokalność

Jak widzieliśmy w poprzednim paragrafie, idealna teoria fizyczna powinna według Einsteina spełniać dwa warunki. Po pierwsze, musi zasadniczo przewidywać wszystkie możliwe do uzyskania rezultaty pomiarowe. Po drugie, nie powinna ona implikować istnienia nielokalnych oddziaływań, czyli takich, które rozchodzą się z dowolnie dużą prędkością lub nawet są natychmiastowe. Jak wiemy, mechanika kwantowa nie spełnia pierwszego z tych warunków (o drugim powiemy później). Jednakże Einstein był przekonany, że w toku rozwoju nauki może ona być zastąpiona przez nową teorię, która dostosuje się do obu wymogów. Nie doczekał powstania takiej teorii, ale nadzieja na jej sformułowanie pozostała nawet po jego śmierci. Nadzieja ta została jednak zdruzgotana w wyniku przełomowego odkrycia, sformułowanego w postaci tzw. twierdzenia Bella. W niniejszym paragrafie przedstawimy pewną wersję tego twierdzenia znaną pod nazwą twierdzenia Clausera-Horne'a-Shimony'ego-Holta (w skrócie CHSH).

Rozumowanie Bella było w istocie bardzo proste. Przyjął on hipotetyczne założenie, że istnieje nowa teoria, rozszerzająca obecną mechanikę kwantową, która operuje pewnym nieznanym nam parametrem λ (nazywa się ten parametr „ukrytym"). Zastosowanie ukrytego parametru (czy też parametrów) λ pozwala na przewidzenie rezultatów pomiarowych dowolnych wielkości. Bell rozważył następnie splątany układ złożony z dwóch cząstek oraz serię pomiarów wykonanych na tych cząstkach. Zgodnie z założeniem lokalności przyjął, że wykonanie pomiaru na jednej z dwóch cząstek nie może zmienić żadnych własności drugiej z nich. Na podstawie tych założeń wyprowadził pewną nierówność (nazwaną jego imieniem), która łączy statystyczne rozkłady rezultatów uzyskanych w seriach pomiarowych. Istotne jest, że chociaż nowa hipotetyczna teoria operuje nieznanymi nam danymi, to jednak obliczone na jej podstawie rozkłady statystyczne są jak najbardziej weryfikowalne eksperymentalnie. Bell następnie pokazał, że standardowa mechanika kwantowa łamie wyprowadzoną przez niego nierówność dla pewnych wybranych mierzalnych własności. Zatem program Einsteina okazał się niewykonalny. Mechanika kwantowa nie może być niesprzecznie rozszerzona o pewne parametry ukryte z zachowaniem zasady lokalności.

Oczywiście nadal pozostaje teoretyczna możliwość, że mechanika kwantowa nie jest do końca poprawną teorią. Gdyby tak było, doświadczenie powinno pokazać, że nierówność Bella obowiązuje w rzeczywistości, wbrew przewidywaniom mechaniki kwantowej. Eksperymentalne badania tego problemu, za które w 2022 r. przyznano Nagrodę Nobla, pokazały jednoznacznie, że nierówność Bella jest łamana przez zjawiska kwantowe. Zatem idealna teoria Einsteina jest nie tylko niezgodna z obecnie przyjmowaną teorią kwantową – jest ona także niezgodna z doświadczeniem. Świat kwantowy nie może być zarazem w pełni określony i lokalny.

Przedstawmy teraz samo rozumowanie prowadzące do twierdzenia CHSH. Załóżmy, że mamy dwie cząstki A i B, przygotowane w pewnym stanie kwantowym i odseparowane od siebie przestrzennie. Na każdej z nich możemy dokonać jednego z dwóch pomiarów pewnych wielkości: A_1 i A_2 dla pierwszej z nich oraz B_1 i B_2 dla drugiej (zakładamy, że wielkości te są parami niekompatybilne, choć ściśle rzecz biorąc, nie jest to potrzebne do wyprowadzenia samej nierówności). Każdy z pomiarów może ujawnić jedną z dwóch możliwych wartości: 1 lub –1 (możemy myśleć o mierzonych wielkościach jako o spinach w różnych kierunkach). Istnieją cztery „globalne" ustawienia eksperymentalne: A_1B_1, A_1B_2, A_2B_1 i A_2B_2. Niech a_{11} i b_{11} oznaczają wyniki pomiarów uzyskane w pierwszym ustawieniu, a_{12}, b_{12} w drugim i tak dalej. Możemy teraz zdefiniować następujący parametr

$$\gamma = a_{11}b_{11} + a_{12}b_{12} + a_{21}b_{21} - a_{22}b_{22}\,.$$

Oczywiście obecna teoria kwantowa nie jest w stanie obliczyć wartości tego parametru, gdyż operuje on alternatywnymi rezultatami pomiarowymi, które nie mogą być łącznie ujawnione w doświadczeniu. Na przykład nie da się zarazem ustalić rezultatu a_{11} i a_{22}, bo wymagałoby to dokonania jednoczesnego pomiaru dla dwóch niekompatybilnych wielkości A_1 i A_2.[9] Natomiast możemy operować wielkościami uśrednionymi. Obliczmy średnią wartość parametru γ, oznaczoną jako $\langle\gamma\rangle$:

$$\langle\gamma\rangle = \langle a_{11}b_{11}\rangle + \langle a_{12}b_{12}\rangle + \langle a_{21}b_{21}\rangle - \langle a_{22}b_{22}\rangle.$$

Średnia z sumy danych liczb jest równa sumie średniej tych liczb. Zauważmy, że każdą z czterech wartości po prawej stronie równania można wyznaczyć eksperymentalnie. Na przykład wartość średniej $\langle a_{11}b_{11}\rangle$ otrzymujemy, powtarzając odpowiednią liczbą pomiarów w ustawieniu A_1B_1. Zatem także liczba $\langle\gamma\rangle$ jest wyznaczalna doświadczalnie. Może także być ona obliczona teoretycznie na podstawie reguł standardowej mechaniki kwantowej, jeśli tylko znany jest stan, w jakim zostały przygotowane cząstki.

Wprowadźmy teraz zasadę lokalności. W obecnym wypadku mówi nam ona, że wartość uzyskana w pomiarze wykonanym na jednej z cząstek nie powinna zależeć od tego, jakiego pomiaru dokonano na odległej, przestrzennie odseparowanej cząstce. Oznacza to, że np. liczby a_{11} i a_{12} powinny być sobie równe, gdyż pierwsza z nich określa rezultat pomiaru wielkości A_1 w sytuacji, kiedy na drugiej cząstce zmierzono B_1, a druga – rezultat tego samego pomiaru przy alternatywnej wielkości B_2 zmierzonej dla drugiej cząstki. Jednakże drugi pomiar nie powinien mieć żadnego wpływu, więc obie wartości należy zrównać. To samo do-

[9] Jak wiemy, seryjny pomiar najpierw wielkości A_1, a potem A_2 nie rozwiązuje problemu, bo wykonanie drugiego pomiaru skasuje wartość uzyskaną w pierwszym (pamiętamy sekwencyjne doświadczenia Sterna-Gerlacha).

tyczy pozostałych parametrów – np. b_{12} i b_{22} powinny być równe. W efekcie otrzymamy uproszczoną wersję wzoru na parametr γ:

$$\gamma = a_1 b_1 + a_1 b_2 + a_2 b_1 - a_2 b_2.$$

Możemy teraz dokonać prostego przekształcenia algebraicznego:

$$\gamma = a_1(b_1 + b_2) + a_2(b_1 - b_2).$$

Maksymalna możliwa wartość liczby $b_1 + b_2$ to 2, przy czym wtedy $b_1 - b_2 = 0$. Z kolei wartość minimalna to –2. Łatwo się przekonać, że wynika z tego, iż wartość parametru γ musi mieścić się w przedziale od –2 do 2. To samo oczywiście dotyczy średniej, a zatem otrzymaliśmy nierówność:

$$-2 \leq \langle\gamma\rangle \leq 2.$$

Jest to właśnie nierówność Bella w wersji CHSH. Okazuje się, że jeśli za A_1 i A_2 oraz B_1 i B_2 podstawimy spiny w odpowiednio dobranych kierunkach, a stan cząstek będzie stanem singletowym (7.1), to obliczona na podstawie reguł mechaniki kwantowej wartość średniej $\langle\gamma\rangle$ wynosi –2,5, co jest liczbą spoza dopuszczonego przedziału. Zatem mechanika kwantowa jest niezgodna z co najmniej jednym z założeń prowadzących do powyższej nierówności.

Zwykle winą za powyższą niezgodność obarcza się założenie o całkowitym zdeterminowaniu mierzalnych wielkości (zwane także realizmem posiadanych własności lub hipotezą parametrów ukrytych). Jeśli odrzucimy przypuszczenie, że obiekty kwantowe posiadają dobrze określone wartości dla wszystkich fizycznych wielkości, to jak się wydaje, możemy przynajmniej uniknąć drugiej z niepokojących Einsteina cech teorii, mianowicie nielokalności. Jednak sprawa nie jest taka prosta. Nielokalność rozumiana jako uzależnienie rezultatu pomiarowego na jednej cząstce od tego, jaką wielkość zmierzono na drugiej, jest istotnie zbędna. Natomiast mechanika kwantowa charakteryzuje się inną nielokalnością, która chociaż w pewnym sensie słabsza od poprzedniej, nadal stanowi odejście od Einsteinowskiego ideału. Jak widzieliśmy na przykładzie stanu singletowego, ujawniony rezultat pomiarowy na jednej z cząstek zależy silnie od rezultatu otrzymanego na drugiej cząstce. Jeśli jeden z rezultatów jest „do góry", drugi musi być „w dół". Gdyby natomiast pierwszy był „w dół", drugi byłby „w górę". Zatem sytuacja w odległym rejonie w jakiś sposób „wpływa" na to, co się dzieje z drugą cząstką.

Mamy więc dwa rodzaje nielokalności, jakie mogą się pojawić w teorii kwantowej. Nielokalność mocniejsza, zwana czasem „zależnością od parametru", przejawia się w tym, że fizyczny stan danego układu zależy od wyboru wielkości pomiarowej przez odległego eksperymentatora. Na przykład wybór orientacji aparatu Sterna-Gerlacha, czyli wybór spinu wzdłuż pewnej osi, zmieniłby jakąś własność drugiego obiektu. Natomiast nielokalność słabsza (zwana zależnością od rezultatu) to efekt uzależnienia stanu danej cząstki od rezultatu pomiarowego uzyskanego w odległym obszarze. Zasadniczą różnicą między nimi jest to, że w pierwszym wypadku mamy kontrolę nad przyczyną odległej zmiany, gdyż ustawienia aparatury zależą od decyzji badacza. Zatem teoretycznie moglibyśmy wykorzystać taki nielokalny związek do przesłania informacji na odległość, łamiąc zasadę nieistnienia sygnałów szybszych od światła. Natomiast drugi rodzaj nielokalności nie daje nam możliwości przesyłania sygnałów. Eksperymentator nie ma żadnej kontroli nad tym, jaki rezultat ujawni jego pomiar. Można domniemywać, że informacja na temat wyniku pomiarowego w jednym układzie musi zostać przesłana do drugiego układu, ale obserwator nie jest w stanie z niej sko-

rzystać. Mówiąc żartobliwie, jest to informacja wyłącznie do wewnętrznego użytku obiektów kwantowych.

W teorii informacji kwantowej dowodzi się prostego twierdzenia znanego pod nazwą „twierdzenia o niemożliwości przesyłania sygnału" (*no-signaling theorem*). Mówi ono, że nie jest możliwe przesłanie wiadomości z jednej części splątanego układu do drugiej. Wiadomość taka mogłaby być dana np. w postaci rozkładu prawdopodobieństwa mierzalnego na jednej z dwóch cząstek. Gdyby można było wykonać dany pomiar na jednej cząstce w taki sposób, że rozkład prawdopodobieństwa rezultatów dla cząstki drugiej uległby zmianie, mielibyśmy możliwość natychmiastowego porozumiewania się na dowolną odległość. Jednakże prawdopodobieństwa wyników pomiarowych są niezależne od tego, jakiego pomiaru dokonano na drugiej z cząstek. W wypadku prostego stanu singletowego (EPR) prawdopodobieństwo każdego rezultatu dla dowolnie wybranego spinu pojedynczej cząstki jest zawsze równe ½, niezależnie od tego, czy druga cząstka została poddana pomiarowi, czy nie. Jeśli nawet na odległej cząstce mierzymy spin w tym samym kierunku, to ponieważ tam również prawdopodobieństwa dwóch różnych rezultatów są równe, cząstka lokalna będzie miała równe prawdopodobieństwa, jakby nic się nie stało. Każdy pojedynczy rezultat będzie oczywiście jednoznacznie determinował drugi, zgodnie ze ścisłą korelacją, ale determinacja przez zdarzenie, które samo jest indeterministyczne, daje w rezultacie również zdarzenie indeterministyczne. To samo dotyczy pomiarów spinu wzdłuż różnych osi na obu cząstkach.

7.6. Problem pomiaru

Przystąpimy teraz do omówienia bodaj najbardziej kontrowersyjnego problemu leżącego u podstaw mechaniki kwantowej. Żadne popularne ujęcie kwantów nie może się obyć bez słynnego stworzenia, jakim niewątpliwie stał się kot Schrödingera. Popularność tego zwierzaka jest tak ogromna, że zdesperowany Roger Penrose miał kiedyś powiedzieć „Kiedy słyszę o kocie Schrödingera, odbezpieczam rewolwer". Spróbujemy może wyjść nieco poza omówienie smutnego losu ulubieńca Schrödingera, przypominając jedynie pokrótce jego historię, zapewne znaną czytelnikom. Kota tego uwięziono (na szczęście tylko fikcyjnie) w diabelskiej skrzynce, która zawiera urządzenie rozbijające flaszeczkę z trucizną. Urządzenie to podłączone jest do jakiegoś kwantowego układu – na przykład może to być nietrwały atom, ulegający rozpadowi promieniotwórczemu. Podłączenie wykonane zostało w taki sposób, że stan atomu przed rozpadem jest skorelowany z nienaruszoną buteleczką, natomiast stan po rozpadzie – z buteleczką rozbitą. Stan rozpadającego się atomu w danej chwili jest opisany jako superpozycja dwóch stanów, przy czym współczynniki tej superpozycji zmieniają się w czasie – współczynnik stanu przed rozpadem maleje, a stanu po rozpadzie rośnie. Oznacza to, że wraz z upływem czasu prawdopodobieństwo znalezienia atomu w stanie po rozpadzie rośnie. Dopóki jednak nie otworzymy pudełka z urządzeniem i kotem, stan atomu jest nieokreślony. Powstaje zatem natrętne pytanie: w jakim stanie znajduje się kot przed otwarciem pudełka? Superpozycja kota żywego i martwego jest raczej trudna do wyobrażenia, pozostaje więc szerokie pole dla różnych dywagacji.

Przykład z kotem ilustruje zasadniczą trudność z pogodzeniem zachowania mikroobiektów, które mogą zajmować stany superponowane, z zachowaniem dobrze nam znanych obiektów makroskopowych, takich jak kot, które raczej nie znajdują się w niedookreślonych

stanach. Problem ten staje się szczególnie dotkliwy w kontekście pomiarów, opierających się na korelacji między pewnymi stanami kwantowymi a stanami obiektów makroskopowych, takich jak położenia wskaźnika odpowiedniego urządzenia pomiarowego. Zanim jednak przejdziemy do szczegółowej analizy zagadnienia pomiaru, musimy uzupełnić zasadniczą lukę w naszej prezentacji mechaniki kwantowej. Jak do tej pory pomijaliśmy milczeniem tę część teorii, która jest najważniejsza dla praktykującego fizyka. Podstawowym celem każdej teorii fizycznej jest przewidywanie przyszłej ewolucji układów fizycznych. Jakie prawo rządzi ewolucją obiektów kwantowych? Na pewno nie będzie to Newtonowska druga zasada dynamiki, bo ta przecież stosuje się do przedmiotów o dobrze określonych własnościach, jak położenie i pęd. Musimy znaleźć nową regułę mówiąca, jak kwantowy stan układu, opisany w postaci pewnego abstrakcyjnego wektora, będzie zmieniał się w czasie. Reguła taka nosi nazwę równania Schrödingera.

Równanie Schrödingera powstało w ramach tzw. mechaniki falowej, w której zjawiska kwantowe opisywane miały być w analogii do zjawisk falowych z mechaniki klasycznej. Jak pamiętamy, wiele efektów kwantowych, takich jak np. interferencja elektronów w doświadczeniu z dwiema szczelinami, może być wyjaśnionych za pomocą założenia istnienia pewnej fali, reprezentującej prawdopodobieństwo znalezienia cząstki w danym obszarze. Fala ta zwana jest funkcją falową dla podkreślenia jej abstrakcyjnego charakteru (nie jest to żadna „fizyczna" fala rozchodząca się w pewnym ośrodku, takim jak powietrze, a nawet – tak jak fale elektromagnetyczne – w samej próżni). W abstrakcyjnym ujęciu funkcje falowe są po prostu pewnymi wektorami w nieskończenie wielowymiarowej przestrzeni Hilberta. W każdym razie, opierając się na pewnych analogiach z sytuacją klasyczną, a także postulując pewne matematyczne własności, Schrödinger sformułował poniższe równanie, które powinny spełniać funkcje falowe, a szerzej wszelkie reprezentacje stanów kwantowych. Oznaczając przez ψ funkcję falową danego układu, mamy:

$$i\hbar \frac{\partial \psi}{\partial t} = H\psi.$$

Symbol H oznacza pewien operator różniczkowy. Jego fizyczna interpretacja jest taka, że reprezentuje on całkowitą energię układu. Nie będzie zatem zaskoczeniem, że operator ten nosi nazwę hamiltonianu, podobnie jak w mechanice klasycznej. Zatem jak widać, ewolucja układu zależy od jego całkowitej energii, tak samo jak w mechanice Hamiltonowskiej. Matematyczne szczegóły równania Schrödingera nie są dla nas w tej chwili istotne. Natomiast niezmiernie ważne jest, że równanie to można przepisać w dużo prostszej formie, która będzie bardzo użyteczna dla naszych celów:

$$\psi(t) = U(t)\psi_0.$$

Symbol ψ_0 reprezentuje stan układu w pewnej ustalonej chwili początkowej, $\psi(t)$ jest oczywiście stanem w późniejszej chwili t, a $U(t)$ to pewien operator liniowy, który zależy od hamiltonianu H i także od czasu t. Operator ten należy do klasy tzw. operatorów unitarnych, które nie zmieniają długości danego wektora. Sens powyższego równania jest dość jasny. Implikuje ono, że jeśli dany jest stan wyjściowy ψ_0 oraz całkowita energia układu (hamiltonian), to stan tego układu w dowolnej chwili późniejszej t jest jednoznacznie wyznaczony poprzez działanie operatora ewolucji $U(t)$.

Mamy zatem do czynienia ze ścisłym i bezwyjątkowym determinizmem – równanie Schrödingera jednoznacznie określa ewolucję stanu każdego układu fizycznego. Może to być

sporym zaskoczeniem. Czy nie popełniliśmy tutaj jakiegoś błędu? Jak pogodzić domniemany deterministyczny charakter ewolucji obiektów kwantowych z indeterminizmem obecnym w probabilistycznych regułach mechaniki kwantowej? Rozwiązanie tego pozornego paradoksu leży w szczególnej formie stanów kwantowych. Jak pamiętamy, stan układu w danej chwili nie określa jednoznacznie wszystkich możliwych do zmierzenia parametrów, a tylko prawdopodobieństwa uzyskania poszczególnych wartości. Determinizm równania Schrödingera oznacza, że rozkład prawdopodobieństwa wyników pomiarowych w danej chwili jednoznacznie wyznacza rozkłady prawdopodobieństw w chwilach późniejszych. Natomiast determinizm ten nie pozwala na dokładne określenie, jakie rezultaty pomiarowe uzyskamy. Pokazaliśmy już, że standardowa mechanika kwantowa nie dysponuje parametrami ukrytymi, które wyznaczyłyby wartości wszystkich wielkości mierzalnych. Zatem determinizm mechaniki kwantowej implikuje jedynie, że „niekompletna" informacja na temat danego układu w pewnym momencie jednoznacznie wyznacza niekompletną informację w późniejszych chwilach. Jest to w pewnym sensie „determinizm indeterminizmu".

Jednakże pewien problem pozostaje. Skoro indeterminizm mechaniki kwantowej przejawia się na poziomie procesów pomiarowych, a ewolucja fizyczna układów kwantowych dana jest deterministycznym równaniem Schrödingera, można wnosić, że pomiary nie podlegają temu równaniu. Wydaje się więc, że w mechanice kwantowej mamy do czynienia z dwoma typami procesów: spokojnym, deterministycznym rozwojem funkcji falowej, zgodnie z równaniem Schrödingera, oraz gwałtownym, nieprzewidywalnym i jedynie probabilistycznym procesem pomiarowym. Tak też przedstawiana jest ta kwestia w wielu podręcznikach i ujęciach mechaniki kwantowej. Budzi to jednak uzasadnione wątpliwości. Czy ograniczenie stosowalności równania Schrödingera nie pozbawia go statusu uniwersalnego prawa przyrody? Do tego należy zapytać, gdzie dokładnie przebiega granica między procesami „Schrödingerowskimi" (nazywanymi również „unitarnymi") a indeterministycznymi procesami pomiarowymi. Co fizycznie odróżnia pomiary od innych procesów? Co jest takiego specjalnego w pomiarach? Na to pytanie udzielano różnych odpowiedzi, szczególnie w różnych interpretacjach mechaniki kwantowej. My jednak spróbujemy potraktować pomiary tak jak wszystkie inne procesy fizyczne. Zasadniczo powinny się one dać opisać w kategoriach ewolucji unitarnej. Zobaczymy, dokąd doprowadzi nas taka idea.

Posłużmy się znów przykładem pomiaru spinu elektronu. Potraktujmy urządzenie pomiarowe jako „czarną skrzynkę", której wewnętrzna zasada działania jest dla nas nieznana. Wiemy tylko, że kiedy na wejściu naszej skrzynki pojawia się elektron, na wyjściu otrzymujemy jeden z dwóch rezultatów: albo spin do góry, albo na dół. Zatem skrzynka do pomiaru spinu musi mieć dwa możliwe stany odpowiadające tym rezultatom. Oznaczmy je jako S_+ i S_-. Do tego dodajmy jeszcze trzeci stan S_0 przed pomiarem, który sygnalizuje gotowość urządzenia. Załóżmy teraz, że wchłonięty elektron posiadał spin z skierowany do góry. Zatem początkowy stan całego układu elektron plus urządzenie pomiarowe będzie dany w postaci iloczynu $z_+ \otimes S_0$. Jeśli urządzenie działa poprawnie, to na wyjściu powinien pojawić się rezultat S_+. Czyli operator ewolucji U, opisujący proces pomiarowy, musi zadziałać na wejściowy stan w następujący sposób:

$$U(z_+ \otimes S_0) = z_+ \otimes S_+ \,.$$

Analogicznie powinna wyglądać sytuacja ze spinem elektronu w dół. Jak natomiast zachowa się urządzenie pomiarowe, kiedy wejściowy elektron będzie w stanie superpozycji spinów w górę i w dół? Odpowiedź zawarta jest w sposobie, w jaki operatory liniowe działają

na wektory. Jak sama nazwa wskazuje, spełniają one warunek liniowości, czyli zadziałanie operatora na sumie wektorowej jest równe sumie działania tego operatora na każdy wektor z osobna. Można to zapisać następująco:

$$U(z_+ \otimes S_0 + z_- \otimes S_0) = U(z_+ \otimes S_0) + U(z_- \otimes S_0) = z_+ \otimes S_+ + z_- \otimes S_- \, .$$

Zatem stan końcowy układu złożonego z elektronu i urządzenia pomiarowego będzie superpozycją stanów z rezultatem „spin do góry" i rezultatem „spin w dół". Jest to nic innego jak powtórzenie problemu kota Schrödingera. Zgodnie z obowiązującą regułą ewolucji układów kwantowych, makroskopowe urządzenie pomiarowe powinno znaleźć się w superpozycji. Jednakże w rzeczywistości obserwujemy dokładnie jeden z dwóch rezultatów, a nie tajemnicze „nałożenie" dwóch różnych wyników.

Powyższe rozumowanie, prowadzące do konfliktu z doświadczeniem, opiera się na trzech fundamentalnych założeniach. Wypiszmy je poniżej w punktach, jako że będą one punktami wyjścia do różnych alternatywnych interpretacji mechaniki kwantowej.

1. Równanie Schrödingera jest bezwyjątkowym prawem, obowiązującym dla wszystkich układów fizycznych.
2. Superpozycja stanów z dobrze zdefiniowanymi, lecz różnymi wartościami danej wielkości jest stanem, w którym wielkość ta nie jest dobrze określona.
3. Każdy pomiar fizyczny kończy się dokładnie jednym rezultatem.

Jak wykazaliśmy, te trzy założenia prowadzą do sprzeczności, a więc któreś z nich musi zostać odrzucone. W następnym paragrafie omówimy najważniejsze próby uniknięcia powstałego paradoksu.

7.7. Interpretacje mechaniki kwantowej

Wspomnieliśmy wcześniej, że najbardziej rozpowszechnioną metodą poradzenia sobie z problemem pomiaru jest odrzucenie przesłanki numer 1. Pomiary wprowadzają nową jakość, gdyż nie poddają się unitarnej ewolucji Schrödingerowskiej. Stanowią one przykład procesów zasadniczo niedeterministycznych, w przeciwieństwie do deterministycznej ewolucji opisanej równaniem Schrödingera. Ponieważ gwałtowna zmiana stanu wywołana pomiarem określana jest często mianem kolapsu paczki falowej, proponowane rozwiązanie określa się jako teorię kolapsu. Zasadniczym wyzwaniem jest tutaj ścisłe wyznaczenie granicy między kwantowymi procesami deterministycznymi a indeterministycznymi pomiarami. Twórcy mechaniki kwantowej, przede wszystkim Bohr, preferowali rozwiązanie oparte na ostrym podziale między mikro- i makroobiektami. Cechą charakterystyczną procesów pomiarowych, odróżniającą je od „zwykłych" zjawisk mikroświata, miałaby być interakcja makroskopowych urządzeń pomiarowych, składających się z ogromnej liczby cząstek, z niewielkimi, mikroskopowymi układami poddanymi pomiarom. To właśnie ta interakcja byłaby odpowiedzialna za złamanie reguły rządzącej samymi zjawiskami mikroświata. Bohr twierdził, że wszelkie problemy interpretacyjne mechaniki kwantowej ujawniają się, kiedy chcemy zastosować do opisu mikroświata klasyczne pojęcia stworzone przez kontakt z makroskopowymi obiektami. Teza ta stała się podstawą tzw. ortodoksyjnej interpretacji mechaniki kwantowej, znanej również pod nazwą interpretacji kopenhaskiej.

Ostry podział między makro- a mikroświatem natrafia jednak na zasadnicze trudności. Przede wszystkim nie jest jasne, gdzie dokładnie miałaby przebiegać granica tego podziału.

Czy molekuła składająca się z dwudziestu atomów jest mikro- czy makroobiektem? A tysiąc takich molekuł? Jeszcze poważniejszy jest inny problem. Wyjaśnienie unikalności kolapsu pomiarowego przez odwołanie do makroskopowości nie tłumaczy, dlaczego makroskopowe obiekty nie podlegają w pełni Schrödingerowskiej dynamice. Co jest takiego szczególnego w systemach składających się z wielkiej liczby elementarnych składników, że nie mogą one znajdować się w superponowanych stanach? Doświadczenie mówi nam, że tak jest istotnie, ale nasza teoria powinna być w stanie to uzasadnić, a nie po prostu przyjmować jako dane. Powrócimy do tego problemu w dodatku w ramce.

Istnieją alternatywne ujęcia kolapsu pomiarowego. Przez pewien czas popularnością cieszyła się np. koncepcja, zgodnie z którą kolaps jest rezultatem interakcji świadomości obserwatora z rzeczywistością fizyczną. Dopóki obserwator nie ujrzy kota w pudełku, stan zwierzęcia jest nieokreślony. Rozwiązanie to jest oczywiście narażone na przeróżne zarzuty. Nie wiadomo, co ze świadomością kota czy innych zwierząt – czy one też nie powinny mieć destrukcyjnego wpływu na kwantowe superpozycje? Poza tym oczywiście konieczne jest przyjęcie dualistycznego podziału na rzeczywistością fizyczną i umysłową, a także założenie tajemniczego oddziaływania umysłu na świat fizyczny. Obecnie niewielu fizyków decyduje się na tak radykalne stanowisko (do wyjątków należy fizyk amerykański Henry Stapp).

Najciekawsza interpretacja mechaniki kwantowej, odrzucająca pierwszą przesłankę paradoksu pomiaru, została zaproponowana pod koniec ubiegłego stulecia przez grupę fizyków włoskich pod kierunkiem nieżyjącego już GianCarla Ghirardiego (pozostali dwaj to Alberto Rimini i Tulio Weber). Interpretacja ta nosi nazwę teorii spontanicznej lokalizacji lub też teorii GRW. Jej zasadnicza idea jest bardzo prosta. Postuluje się, że dla każdego obiektu kwantowego, którego położenie charakteryzuje się „rozmytą" funkcją falową, istnieje pewne bardzo niewielkie prawdopodobieństwo, że funkcja ta spontanicznie zlokalizuje się wokół jednego punktu. Proces lokalizacji jest zupełnie indeterministyczny i nie podlega równaniu Schrödingera. Jest on również niezmiernie rzadki, tak że średni czas oczekiwania na nastąpienie owej spontanicznej lokalizacji dla jednej cząstki jest porównywalny z czasem życia wszechświata. Jednakże jeśli weźmiemy pod uwagę makroskopowe ciało, które zawiera ogromną liczbę cząstek (rzędu 10^{20}), to oczywiście jest niemal pewne, że któraś z cząstek ulegnie takiej lokalizacji. Ponieważ cząstki kwantowe należące do danego systemu pozostają ze sobą w skomplikowanych relacjach splątania, lokalizacja nawet jednej z nich powoduje, że pozostałe natychmiast idą jej śladem i także ulegają lokalizacji (jest to efekt analogiczny do „uzgadniania" spinów w doświadczeniu z elektronami w splątanym stanie singletowym). Zatem ciała makroskopowe, nawet jeśli przez drobny ułamek sekundy znajdują się w superpozycji stanów z różnymi położeniami, niemal natychmiast zlokalizują się w jednym dobrze określonym miejscu. To samo oczywiście dotyczy urządzeń pomiarowych, czyli problem jednoznaczności rezultatów pomiarowych został rozwiązany.

Teoria GRW w bardzo elegancki sposób tłumaczy różnicę w zachowaniu pojedynczych obiektów kwantowych i ich ogromnych kolekcji. Niestety, jej główną trudnością jest, że nie istnieją żadne dane doświadczalne, potwierdzające istnienie takiego całkowicie stochastycznego efektu lokalizacji. Co więcej, przeprowadzone eksperymenty sugerują, że jeśli taki efekt istnieje, średni czas oczekiwania na lokalizację będzie zbyt długi, aby wytłumaczyć obserwowane zachowanie ciał makroskopowych. Musimy zatem szukać dalszych rozwiązań paradoksu pomiaru. Inna grupa interpretacji kwestionuje drugie założenie, czyli tezę, że układy kwantowe w superpozycji nie posiadają dobrze określonych cech. Jest to oczywiście stara dobra koncepcja parametrów ukrytych, która choć mocno nadwyrężona przez twierdzenie

Bella, nie została zupełnie obalona. Najlepiej rozwiniętą wersją tej koncepcji, posiadającą sporą grupę zagorzałych zwolenników zarówno wśród fizyków, jak i filozofów, jest mechanika, której twórcą był David Bohm.[10] Interpretacja Bohmowska opiera się na założeniu, że wszystkie obiekty kwantowe mają dobrze określone lokalizacje i trajektorie – czyli że zasadniczo wyglądają jak klasyczne obiekty z teorii newtonowskiej. Jednakże ich zachowanie nie podlega prawom mechaniki Newtona, a nowym prawom kwantowym. Teoria Bohma wprowadza pewne dodatkowe równanie, które uzależnia prędkość danego ciała (a nie przyspieszenie, jak u Newtona) od kwantowej funkcji falowej. Zatem funkcja falowa pełni tutaj rolę pola sił, oddziałując na cząstki i „popychając" je przez przestrzeń. Jednocześnie funkcja falowa, podobnie jak w standardowej mechanice kwantowej, zawiera probabilistyczną informację, jakie są szanse znalezienia cząstki w odpowiednim obszarze. Oczywiście prawdopodobieństwo to charakteryzuje teraz jedynie naszą niewiedzę, gdyż naprawdę cząstki mają dobrze określoną lokalizację, której jednak z fundamentalnych powodów nie znamy.

W interpretacji Bohmowskiej rezultaty pomiarów są jednoznacznie wyznaczone przez położenie cząstki w obrębie jej funkcji falowej. Zasadniczo wszystkie pomiary dla Bohma sprowadzają się do pomiaru położenia w przestrzeni (np. mierzymy spin w doświadczeniu Sterna-Gerlacha przez lokalizację elektronu albo w górnej, albo w dolnej wiązce). Superpozycja stanu na wejściu oznacza, że funkcja falowa składa się z części reprezentującej spin do góry i z części spinu w dół. Jeśli elektron na początku oddziaływania znajdzie się w obszarze, w którym dominuje część funkcji „spin do góry", to w wyniku oddziaływania z magnesem elektron ten poleci w górę i zarejestrujemy go w odpowiednim miejscu. Rezultat ten jest jednoznacznie wyznaczony przez uprzednie położenie elektronu, którego jednak nie jesteśmy w stanie określić.

Twierdzenie Bella poucza nas, że każda teoria zakładająca zdeterminowanie wszystkich rezultatów pomiarowych musi być mocno nielokalna. Tak jest w istocie w wypadku teorii Bohma. Nielokalność tej teorii wynika z faktu, że na ewolucję danej cząstki ma wpływ cała funkcja falowa rozciągnięta w przestrzeni, która z kolei zależy od rozkładu innych, odległych obiektów. Na przykład w wypadku doświadczenia z dwiema szczelinami to, czy druga szczelina jest zamknięta czy otwarta, ma wpływ na kształt funkcji falowej, a zatem także na zachowanie elektronu przechodzącego przez pierwszą szczelinę. Zatem w pewnym sensie elektron „wie" bez żadnego opóźnienia, czy druga szczelina jest otwarta i odpowiednio do tego dostosowuje swoją trajektorię.

Wreszcie pojawia się możliwość odrzucenia trzeciej przesłanki. Na pozór może wydawać się to niedorzecznością. Jak można kwestionować oczywisty fakt, że po wejściu do laboratorium fizyk widzi, czy elektrony po przejściu przez urządzenie Sterna-Gerlacha zostały zarejestrowane na górnej czy dolnej części kliszy fotograficznej? I co by to miało znaczyć, że pomiary nie mają jednoznacznych rezultatów? Jednakże istnieje bardzo popularna koncepcja, która wyjaśnia, w jaki sposób jest to możliwe. Interpretacja ta nazywa się teorią wielu światów lub też interpretacją Everetta. Hugh Everett III w swojej rozprawie doktorskiej zasugerował, że równanie Schrödingera powinno być spełnione przez wszystkie fizyczne procesy, niezależnie od tego, czy nazwiemy je pomiarami czy nie. Everett był zdecydowany nie wprowadzać do teorii kwantowej żadnych dodatkowych równań czy zależności, jak np. w wypadku teorii Bohma. Jego interpretacja jest w pewnym sensie „najczystszą" interpretacją for-

[10] Podobne idee formułował również wspominany wcześniej de Broglie, dlatego tez teoria ta jest czasem nazywana teorią de Broglie'a-Bohma.

malizmu kwantowo-mechanicznego. Natomiast rezultaty pomiarowe skorelowane z poszczególnymi składnikami superpozycji Everett zinterpretował jako istniejące „relatywnie" do wyboru tego a nie innego składnika. Co to jednak znaczy „istnieć relatywnie"? Brian de Witt zasugerował rozwiązanie, które jest do dziś uważane za standardowe: każdy rezultat pomiarowy istnieje w innej rzeczywistości zwanej światem. Zatem w wyniku pomiaru nasz świat rozpada się na tyle światów, ile jest możliwych do uzyskania rezultatów. W każdym z tych światów jest obserwator, który widzi jeden rezultat, ale naprawdę wszystkie rezultaty istnieją „na raz". Jeśli na przykład mierzymy spin danego elektronu, to pomiar taki skutkuje rozszczepieniem na dwa światy, z których jeden zawiera rezultat „spin do góry", a drugi „spin w dół".

Trzeba przyznać, że jest to śmiała hipoteza, chociaż na wielu filozofach nie robi ona wielkiego wrażenia (np. filozof David Lewis nie miał oporów przed postulowaniem realnego istnienia ogromnej liczby tzw. światów możliwych dla wyjaśnienia znaczenia modalnych pojęć możliwości i konieczności). Natomiast od razu pojawiają się trudności, które muszą zostać rozwiązane. Po pierwsze, istnieje formalny problem arbitralności w wyborze odpowiednich składników superpozycji danego stanu, które mają odpowiadać odrębnym światom. Jak pamiętamy, w wyniku procesu pomiarowego system poddany pomiarowi i aparat pomiarowy znajdą się w superpozycji ortogonalnych stanów odpowiadających różnym rezultatom. Okazuje się jednak, że matematycznie rzecz biorąc istnieje wiele (nawet nieskończenie wiele) alternatywnych rozkładów tego samego stanu na inne ortogonalne składowe, które wcale nie odpowiadają otrzymanym rezultatom pomiarowym, a tylko ich superpozycjom.[11] Dlaczego w takim razie światy możliwe odpowiadają tylko wyróżnionym rozkładom danego stanu kwantowego? Problem ten jest znany pod nazwą problemu wyróżnionej bazy. Jego typowe rozwiązanie odwołuje się do pewnego procesu, zwanego dekoherencją. Więcej informacji na ten temat znajdziecie w ramce.

Przy analizie oddziaływania pojedynczego układu fizycznego (np. cząstki) z makroskopowym obiektem ważną rolę odgrywa proces dekoherencji. Dekoherencja polega z grubsza na tym, że stany mikroskopowego układu, odpowiadające dobrze określonym wartościom pewnej wielkości mierzalnej (głównie położenia) zostają skorelowane z pewnymi stanami układu makroskopowego. Innymi słowy, mikroskopowe stany własne zostają „zarejestrowane" w globalnym stanie układu złożonego z bardzo wielkiej liczby mikroukładów. W rezultacie tego procesu stan układu poddanego takiej interakcji zmienia się z superpozycji na stan mieszany (por. wpis w ramce na s. 247). Oznacza to zanik typowych dla superpozycji efektów interferencyjnych (stąd nazwa „dekoherencja", czyli zanik koherencyjności). Istnieją modele teoretyczne pokazujące, że najbardziej prawdopodobnym rodzajem dekoherencji dla typowych układów pomiarowych jest dekoherencja ze względu na stany własne położenia. Zatem można przyjąć, że w wyniku pomiaru stan cząstki „rozpada się" na stany z dobrze określonym położeniem (alternatywne „światy"), które ze sobą nie interferują.

Inny problem dotykający interpretacji wieloświatowej ma charakter bardziej filozoficzny. Jak wiadomo, podstawowym pojęciem w mechanice kwantowej jest pojęcie prawdopodobieństwa. Przed pomiarem np. spinu możemy stwierdzić jedynie, że prawdopodobień-

[11] Łatwo to wytłumaczyć geometrycznie: jak wiadomo, dla każdego wektora istnieje nieskończenie wiele sposobów rozłożenia go na prostopadłe do siebie wektory składowe.

stwo rezultatu „do góry" jest takie a takie, gdyż nie wiemy, jaki będzie rezultat. Natomiast w teorii wieloświatowej ta zasadnicza niewiedza znika. Ponieważ dla każdego możliwego rezultatu istnieje świat, w którym ten rezultat się pojawi, nie widać potrzeby wprowadzania obiektywnego pojęcia prawdopodobieństwa. Ewolucja układu jest jednoznacznie określona – wynika to z faktu, że w interpretacji Everetta równanie Schrödingera obowiązuje bezwyjątkowo, a przecież, jak już widzieliśmy, jest ono całkowicie deterministycznie. Niektórzy próbują rozwiązać ten problem przez wprowadzenie „miary realności": dla każdego świata istnieje pewna miara, dana przez kwadrat amplitudy składnika superpozycji odpowiadającego temu światu, która określa, jak bardzo realny jest dany świat. Jednak pojęcie miary realności jest samo w sobie bardzo niejasne. Istnieje inne podejście, ciekawsze z punktu widzenia filozofa. Odwołuje się ono do interpretacji prawdopodobieństwa jako miary niewiedzy dotyczącej indywidualnego osobowego doświadczenia. Rozbicie uniwersum na szereg alternatywnych możliwych światów jest związane z tym, że w przyszłości obserwator będzie miał szereg kopii czy też kontynuacji. Mimo to pozostaje niewiadomą, którą z kontynuacji ujrzy dany obserwator. Jest to analogiczne do następującego problemu: podajemy obserwatorowi środek usypiający i wywozimy go do jednego z dwóch miejsc: Gdańska lub Wrocławia. Po obudzeniu pytamy go: „Gdzie się znajdujesz?". Obserwator może jedynie powiedzieć, że z pewnym prawdopodobieństwem będzie to Gdańsk lub Wrocław. Jest to tzw. problem samo-lokacji. W przypadku kwantowo-mechanicznym każdy z obserwatorów po pomiarze, ale przed zobaczeniem jego rezultatu powinien stwierdzić, że prawdopodobieństwo ujrzenia danego rezultatu jest takie, jak to przewiduje formalizm. Jednakże pozostaje wątpliwość, czy prawdopodobieństwa te mają również sens przed pomiarem, kiedy jeszcze jest tylko jeden obserwator.

7.8. Statystyki kwantowe i nieodróżnialność

Ostatnim zagadnieniem, które omówimy w niniejszym rozdziale, będzie kwestia zachowania dużych grup cząstek kwantowych. Jak wiadomo, gdy mamy do czynienia ze znaczną liczbą obiektów, ich opis musi być podany w języku statystyki czy też prawdopodobieństwa. Dotyczy to w równym stopniu ciał podlegających fizyce klasycznej, jak i obiektów kwantowych. Jednakże istnieje zasadnicza różnica między obiektami klasycznymi a kwantowymi w szczegółach takiego opisu. Jest to związane z fundamentalnym problemem nieodróżnialności. Zacznijmy może od rozpatrzenia przypadku klasycznych cząstek, takich jak atomy czy molekuły fizyki statystycznej. W fizyce klasycznej każda indywidualna cząstka jest obdarzona dobrze określonym położeniem i trajektorią, co zasadniczo (choć nie w praktyce) umożliwia odróżnienie i „ponumerowanie" każdej cząstki z osobna. Jeśli chcemy teraz opisać stan bardzo dużej liczby takich cząstek, możemy to zrobić za pomocą przypisania każdemu numerowi poszczególnej cząstki jego indywidualnego stanu. Pamiętamy z rozdziału poświęconego mechanice statystycznej, że takie przypisanie nazywa się aranżacją. Poszczególne aranżacje prowadzą do tzw. dystrybucji, czyli do określenia, ile cząstek zajmuje ile stanów, jednakże bez podania, która cząstka zajmuje który stan. Liczba aranżacji realizująca daną dystrybucję daje nam miarę prawdopodobieństwa, z jakim możemy oczekiwać pojawienia się danej dystrybucji.

Zilustrujmy te rozważania bardzo prostym przykładem. Załóżmy, że mamy do dyspozycji dwie cząstki 1 i 2 oraz dwa stany A i B. Istnieją cztery możliwe rozłożenia tych cząstek: obie mogą trafić do stanu A, obie do B, pierwsza do A i druga do B oraz pierwsza do B

i druga do A. Dwie ostatnie aranżacje realizują tę samą dystrybucję: jedna cząstka w A, jedna w B. Jeśli przyjmiemy założenie, że każda aranżacja jest jednakowo prawdopodobna, to w rezultacie otrzymamy następujące prawdopodobieństwa dystrybucji: dystrybucja „dwie cząstki w A" – ¼, dystrybucja „dwie cząstki w B" – ¼, dystrybucja „jedna cząstka w A, jedna w B" – ½. Jest to przykład klasycznego rozkładu prawdopodobieństwa, noszącego nazwę statystyki Maxwella-Boltzmanna (tab. 7.1).

Aranżacja	Statystyka Maxwella-Boltzmanna	Statystyka Bosego-Einsteina	Statystyka Fermiego-Diraca
A(1, 2)	¼	⅓	0
B(1, 2)	¼	⅓	0
A(1), B(2)	¼	⅓	1
A(2), B(1)	¼		

Tab. 7.1. Prawdopodobieństwa aranżacji według trzech statystyk

Jednakże, jak się okazuje, cząstki kwantowe nie podlegają tej statystyce. Ich obserwowane statystyczne zachowanie jest inne. Cząstki takie jak np. fotony obsadzają stany A i B w taki sposób, że prawdopodobieństwo dystrybucji, w której trafią do różnych stanów, jest takie samo jak prawdopodobieństwo, że będą one w tym samym stanie. Wszystkie trzy prawdopodobieństwa wynoszą 1/3. Można to wytłumaczyć tym, że aranżacje „cząstka 1 w A, cząstka 2 w B" i zamieniona aranżacja „cząstka 1 w B cząstka 2 w A" są w istocie jedną i tą samą aranżacją. Sytuacje, które różnią się jedynie zamianą, czyli permutacją cząstek, są identyczne. Nie ma sposobu dowiedzieć się, która cząstka trafiła do którego stanu, gdyż są one nieodróżnialne (np. nie mają identyfikujących je trajektorii).

Istnieje grupa obiektów kwantowych, których zachowanie jest jeszcze inne. Należą do nich wszystkie elementarne cząstki, z których składa się materia – elektrony, protony, neutrony, a także mniejsze obiekty jak kwarki. Dopuszczalna dla nich jest tylko jedna dystrybucja: ta, w której jedna cząstka znajdzie się w A, a druga w B. Z niewiadomych powodów cząstki te „nie chcą" obsadzać tych samych stanów. Jeśli jeden stan jest już zajęty, druga cząstka musi poszukać sobie innego stanu. Zachowanie takie opisane jest przy pomocy reguły zwanej „zakazem Pauliego", która wyklucza zajmowanie danego stanu przez więcej niż jedną cząstkę. Takie cząstki nazywa się fermionami, a statystykę, której one podlegają, statystyką Fermiego-Diraca. Z kolei fotony i inne zachowujące się podobnie cząstki określa się mianem bozonów. Ich zachowanie opisuje statystyka Bosego-Einsteina.

Powyższe statystyki są wytłumaczone szczególną formą „łącznych" stanów, jakie mogą zajmować bozony lub fermiony. Jak pamiętamy, stan całego układu dwóch lub więcej cząstek kwantowych jest opisywalny iloczynem tensorowym stanów poszczególnych cząstek lub też dowolną kombinacją liniową (superpozycją) takich iloczynów. Rozważmy jako przykład stan, w którym jedna cząstka posiada cechę A, a druga B. Zasadniczo można by opisać ten stan jako iloczyn $A \otimes B$. Jednakże taki iloczyn jest matematycznie różny od „zamienionego" iloczynu $B \otimes A$, podczas gdy cząstki kwantowe nie odróżniają stanów poddanych permutacji. Aby rozwiązać ten problem, wprowadza się pewne ograniczenia na możliwe stany

układów wielu cząstek. Podstawowym wymogiem jest, aby stan fizyczny układu wielu cząstek pozostał taki sam przy zamianie, czyli permutacji cząstek. Postulat ten może być zrealizowany na dwa sposoby. Po pierwsze, reprezentacja stanu cząstek po permutacji może być dokładnie identyczna z reprezentacją stanu wyjściowego. Dopuszczalne jest także, że permutacja zmieni znak wektora stanu na przeciwny. Nie ma to żadnego wpływu na stan fizyczny, gdyż wektory można przemnożyć przez dowolną liczbę, a nadal reprezentują one ten sam stan.[12]

Wektory stanu, które nie ulegają zmianie przy permutacji cząstek, nazywa się symetrycznymi. Charakteryzują one bozony. Z kolei stany fermionowe są nazwane antysymetrycznymi. To terminologia trochę myląca, bo może sugerować, że antysymetryczność jest brakiem (przeciwieństwem) symetrii. W istocie jest to symetria, tylko innego rodzaju. Polega na tym, że permutacja cząstek może „co najwyżej" zmienić znak danego wektora stanu. Proste permutacje dwóch cząstek (zwane w matematyce transpozycjami) zawsze zmieniają znak wektora antysymetrycznego, natomiast niektóre bardziej skomplikowane (np. tzw. cykliczne permutacje, zamieniające ustawienie trzech cząstek 123 na 231) pozostawiają znak niezmieniony.

Zastosujmy wprowadzone wyżej pojęcia do analizy sytuacji z dwiema cząstkami dystrybuowanymi pomiędzy dwoma stanami A i B. W przypadku bozonów możliwe są trzy łączne stany symetryczne skonstruowane przy pomocy takich pojedynczych stanów. Są to:

$$A \otimes A,$$

$$B \otimes B,$$

$$A \otimes B + B \otimes A.$$

Jak zatem widać, mamy trzy, a nie cztery możliwości realizacji obsadzenia w przypadku dwóch cząstek i dwóch stanów. Odtwarza to statystykę Bosego-Einsteina, omówioną powyżej.

Z kolei w przypadku fermionów istnieje tylko jeden stan złożony z dwóch stanów A i B, który spełnia warunek antysymetryczności:

$$A \otimes B - B \otimes A.$$

Łatwo zauważyć, że zamiana A na B zmieni znak całości na przeciwny. Przy okazji widzimy, że nie może istnieć antysymetryczny stan całości złożony z tych samych stanów jednocząstkowych. Prowadzi to od razu do zakazu Pauliego. Jeden stan może być obsadzony tylko przez jeden fermion.

Ograniczenie stanów wielu cząstek tego samego rodzaju do symetrycznych bądź antysymetrycznych funkcji pojedynczych stanów ma konsekwencje wykraczające poza staty-

[12] Można postawić pytanie, dlaczego nie wprowadzić możliwości permutacji, która mnoży wyjściowy stan przez dowolną liczbę zespoloną, a nie tylko +1 lub –1. Zwykle formułowana jest odpowiedź, że dwukrotne zastosowanie tej samej permutacji dwóch cząstek powinno nam dać stan wyjściowy, a tylko liczby 1 i –1 podniesione do kwadratu dają jeden. Jednakże teoretycznie możliwe są sytuacje, w których permutacja stanu trzech lub więcej cząstek powoduje przemnożenie wektora przez liczbę zespoloną o module 1 (pierwiastek n-tego stopnia z 1). Złożenie kilku takich permutacji powinno w rezultacie „wrócić" do stanu wyjściowego. Cząstki opisane tego typu symetriami nazywa się „paracząstkami". Jak dotąd nie znaleziono empirycznego świadectwa na rzecz istnienia paracząstek – wszystkie znane cząstki dzielą się bez wyjątku na bozony i fermiony.

styczne zachowania dużych grup. Istnieją argumenty, że postulat symetryzacji implikuje, iż cząstki tego samego typu (np. elektrony czy fotony) stają się całkowicie nieodróżnialne ze względu na wszystkie mierzalne własności. Jest dość oczywiste, że np. dwa elektrony są nieodróżnialne, jeśli ograniczymy się do ich identyfikujących, charakterystycznych własności, takich jak masa czy ładunek elektryczny (a także spin „całkowity", który wynosi $\frac{\sqrt{3}}{2}\hbar$). Kategoria elektronów nie rozbija się na mniejsze podkategorie o odmiennych cechach charakterystycznych. Jak się jednak wydaje, pojedyncze elektrony mogą się od siebie różnić cechami „nabytymi" (przygodnymi), niestanowiącymi ich istoty, jak np. położenie, pęd czy energia. Mamy przecież do czynienia z elektronami zlokalizowanymi w odpowiednim urządzeniu pomiarowym w laboratorium, które na pewno różnią się lokalizacją od elektronów na Marsie czy w galaktyce Andromedy.

Okazuje się, że formalizm stanów antysymetrycznych (elektrony są fermionami), a także symetrycznych prowadzi do odmiennego i zaskakującego wniosku. Rozważmy dwie cząstki tego samego rodzaju, ponumerowane odpowiednio 1 i 2. Ich stan, jak już wiemy, opisywany jest pewnym wektorem $\psi(1, 2)$, odpowiednio zachowującym się przy zastosowaniu operacji zamiany etykiet 1 i 2:

$$\psi(2, 1) = \pm\psi(1, 2),$$

gdzie znak zależy od tego, czy mamy do czynienia z fermionami czy bozonami. Jeśli przyjmiemy, że cała informacja na temat fizycznych własności cząstek 1 i 2 wyczerpuje się w wektorze ψ, to taka symetryczność niedwuznacznie wskazuje, że wszystko, co jest fizycznie prawdą o cząstce 1 musi być także prawdą o cząstce 2. Etykiety cząstek 1 i 2 zajmują dokładnie takie same miejsca w wektorze stanu, a zatem każde prawdziwe zdanie o cząstce 1 da się „przetransponować" na identyczne zdanie o cząstce 2. Mówiąc jeszcze inaczej, zdania przypisujące własności poszczególnym cząstkom powinny być inwariantne (niezmiennicze) względem permutacji. Oznacza to, że jeśli prawdziwe jest zdanie „Cząstka 1 posiada własność *P*", to prawdziwe musi być również zdanie „Cząstka 2 posiada własność *P*".

Mamy więc zaskakującą konsekwencję, że cząstki tego samego rodzaju nie mogą się różnić żadnymi własnościami. Jest to jawne złamanie zasady tożsamości przedmiotów nieodróżnialnych, o której mówiliśmy przy okazji argumentu Leibniza przeciwko substancjalnemu traktowaniu przestrzeni. Zgodnie z tą zasadą, każde dwa przedmioty powinny się różnić przynajmniej jedną własnością. Może to być własność „wewnętrzna", charakteryzująca tylko sam przedmiot, lub też „zewnętrzna", odnosząca dany przedmiot do innych obiektów. W przypadku symetrycznych czy antysymetrycznych stanów kwantowych ani wewnętrzne, ani zewnętrzne własności nie mogą odróżnić cząstek. Co zatem powoduje, że mamy do czynienia z dwiema, a nie z jedną cząstką?

Wielu filozofów i fizyków uważa, że w takiej sytuacji powinniśmy odrzucić założenie, że cząstki kwantowe są indywidualnymi obiektami, odrębnymi od reszty świata numerycznie i jakościowo. Jako ilustrację takiego podejścia często przytacza się przykład pieniędzy na koncie bankowym. Jednego dnia Jan wpłaca na swoje konto złotówkę, następnego kolejną złotówkę, a po kilku dniach wypłaca kwotę jednego złotego. Którą złotówkę wypłacił Jan? Pytanie jest oczywiście absurdalne. Złotówki na koncie niczym się od siebie nie różnią – są wirtualnymi przedmiotami o takich samych charakterystykach. Podobnie zachowują się elektrony – ich tożsamość nie jest jednoznacznie określona. Niektórzy twierdzą wręcz, że do obiektów kwantowych nie stosuje się zwykłe prawo logiki znane jako zasada tożsamości,

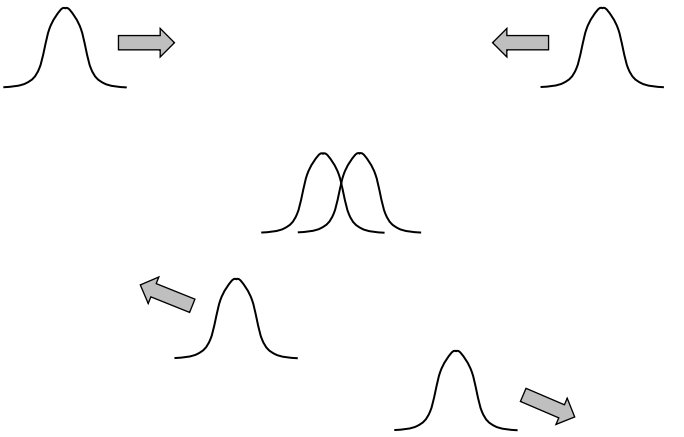
które głosi, że każdy przedmiot jest tożsamy sam ze sobą. Zwolennicy tego podejścia rozwijają nieklasyczną teorię nie-indywiduów, zgodnie z którą zbiorowiska takich nie-indywidualnych obiektów charakteryzują się licznością, ale nie mogą być uporządkowane. Jest sens pytać, ile złotówek znajduje się na koncie, ale nie można wysłać do banku zlecenia, żeby pierwsza i trzecia złotówka na koncie zostały zamienione ze sobą miejscami.

Nie wszyscy jednak sądzą, że należy odrzucić podstawowe prawa logiki w odniesieniu do obiektów kwantowych. Istnieje podejście, zgodnie z którym do kwantów można zastosować pojęcie tzw. słabej odróżnialności – opiera się ono na relacjach, a nie własnościach. Przykładem jest relacja posiadania przeciwnie skierowanych spinów w stanie singletowym EPR, jaka może łączyć przedmioty odrębne numerycznie, ale nigdy nie stosuje się do porównania przedmiotu z samym sobą (żaden obiekt nie może mieć spinu przeciwnego do swojego spinu).[13] Jeszcze bardziej „konserwatywne" podejście podkreśla, że nawet w symetrycznym języku mechaniki kwantowej możliwe jest wyrażenie absolutnej odróżnialności za pomocą zwykłych własności. Zilustrujmy tę obserwację prostym przykładem. Zdanie „Cząstka numer 1 ma spin do góry, a cząstka 2 ma spin w dół" nie jest niezmiennicze względem permutacji, gdyż zamiana etykiet 1 i 2 zmienia to zdanie na inne, które nie będzie prawdziwe, jeśli to pierwsze jest prawdziwe. Rozważmy jednak następujące, trochę bardziej skomplikowane zdanie: „Albo cząstka numer 1 na spin do góry, a cząstka numer 2 ma spin do dołu, albo cząstka numer 1 ma spin do dołu, a cząstka numer 2 ma spin do góry". Zdanie to jest w pełni symetryczne względem permutacji – zamiana etykiet 1 i 2 zachowuje jego sens. Jednocześnie jednak wyraża ono całkiem jasno ideę, że cząstki są odróżnialne za pomocą spinów. Okazuje się, że w formalizmie mechaniki kwantowej można sformułować analogiczne zdania, które okazują się prawdziwe, zatem można utrzymywać – wbrew wcześniejszym sugestiom – że cząstki tego samego rodzaju są jednak odróżnialne.

Powyższe rozważania dotyczyły kwestii odróżnialności i tożsamości obiektów kwantowych w pewnej chwili – mówimy w takim wypadku o tzw. identyczności i odróżnialności synchronicznej. Innym, choć powiązanym problemem jest tożsamość diachroniczna, czyli tożsamość w czasie. Jest to znane z filozofii pytanie o to, jak określić, czy dany przedmiot w chwili t_1 jest tym samym przedmiotem co w innej chwili t_2. W przypadku obiektów kwantowych odpowiedź może okazać się z zasadniczych powodów niemożliwa. Jest tak np. w sytuacji, kiedy dwie cząstki tego samego rodzaju oddziałują ze sobą w taki sposób, że ich funkcje falowe się najpierw przekrywają, a później rozchodzą (rys. 7.5). Wtedy formalizm kwantowy nakazuje wziąć pod uwagę dwie możliwe „drogi" dojścia do stanu końcowego, przy czym dla obliczenia prawdopodobieństwa znalezienia cząstek w określonych obszarach sumuje się te drogi, a nie prawdopodobieństwa. Jest to analogiczne do doświadczenia z dwiema szczelinami – powstaje efekt interferencyjny, który jest wskaźnikiem utraty dia-

[13] Uważny czytelnik może w tym momencie zaprotestować (jeśli jesteś tym czytelnikiem – gratulacje!). Jak to możliwe, aby cząstki miały przeciwne spiny, skoro przed chwilą stwierdziliśmy, że cząstki tego samego rodzaju nie mogą posiadać różnych własności? Sprawa jest nieco skomplikowana. W istocie w stanie singletowym (EPR) cząstki nie posiadają żadnych dobrze określonych spinów – spin w każdym kierunku jest określony tylko z prawdopodobieństwem ½. Kiedy mówimy o antykorelacji spinów, nie chcemy powiedzieć, że jeden elektron ma naprawdę spin do góry, a drugi spin w dół, a tylko że ich suma jest równa zeru. Jest oczywiście przedziwne, że możemy dodawać spiny, które nie są dobrze określone, a rezultat będzie mimo to dobrze określony. To jest właśnie splątanie kwantowe, z którym zetknęliśmy się już przy okazji argumentu EPR.

chronicznej tożsamości przez cząstki. Cząstki w stanie końcowym nie mogą być jednoznacznie powiązane z cząstkami przed oddziaływaniem.



Rys. 7.5. Oddziaływanie dwóch nieodróżnialnych cząstek (funkcji falowych)

7.9.* Elementy formalizmu mechaniki kwantowej

W warstwie matematyczno-formalnej mechanika kwantowa opiera się na bardzo eleganckich podstawach, które są godne uwagi szerszego grona filozofów. Zasadniczą teorią matematyczną wykorzystywaną w opisie kwantowym jest algebra liniowa, podparta analizą matematyczną, szczególnie przy zastosowaniu do konkretnych problemów. My skoncentrujemy się głównie na tej pierwszej teorii i jej wykorzystaniu w teorii kwantów. Dodatkowo wprowadzimy rozpowszechnioną w mechanice kwantowej notację Diraca, która jest niezmiernie wygodnym i intuicyjnym, choć nie niezbędnym sposobem zapisu formuł (są podręczniki, które jej nie stosują). Podstawowe pojęcie teorii kwantów to pojęcie przestrzeni wektorowej. Składa się ona oczywiście z wektorów, w notacji Diraca zapisywanych jako $|u\rangle$, $|v\rangle$ itd. W żargonie kwantowym wektory te nazywa się również „ketami" (źródło tej terminologii wyjaśnimy później). Na wektorach zdefiniowane są dwie operacje: dodawanie oraz mnożenie przez liczbę. W przypadku przestrzeni używanych w mechanice kwantowej są to niestety liczby zespolone, czyli liczby o postaci $a + ib$, gdzie a, b – liczby rzeczywiste, a $i = \sqrt{-1}$. Obie operacje na wektorach spełniają szereg intuicyjnych warunków, takich jak łączność czy rozdzielność mnożenia względem dodawania. (Zakłada się również istnienie wektora zerowego.) Nie będziemy tutaj wypisywać wszystkich aksjomatów; można je znaleźć w każdym podręczniku. W każdym razie wprowadzone w tych aksjomatach operacje umożliwiają nam tworzenie liniowych kombinacji wektorów, takich jak np. $a|u\rangle + b|v\rangle$.

Bardzo ważną operacją definiowaną w przestrzeniach wektorowych jest iloczyn skalarny wektorów, który poznaliśmy już wcześniej. Iloczyn skalarny wektora $|u\rangle$ z $|v\rangle$ oznacza się w notacji Diraca jako $\langle u|v\rangle$. Jest to w ogólności liczba zespolona. Iloczyn skalarny spełnia warunek liniowości w drugim argumencie, co oznacza następującą prawidłowość:

$$\langle u|av + bw\rangle = a\langle u|v\rangle + b\langle u|w\rangle. \tag{7.2}$$

Zwykle zakłada się, że iloczyn skalarny jest przemienny, tj. kolejność czynników nie ma znaczenia. Jednakże w wypadku przestrzeni wektorowych nad liczbami zespolonymi zamiana kolejności w iloczynie zmienia w niewielkim stopniu rezultat mnożenia. Zamiast oryginalnej liczby dostajemy jej sprzężenie zespolone. Sprzężeniem liczby zespolonej $a + ib$ jest $a - ib$. Operację sprzężenia oznaczamy przy pomocy gwiazdki *. Zatem mamy:

$$\langle u|v\rangle = \langle v|u\rangle^*.$$

Wynika z tego, że iloczyn skalarny będzie antyliniowy w pierwszym argumencie, co oznacza podobną formułę do (7.2), z tym że współczynniki a i b będą sprzężone (rozpiszcie to sobie). Zauważmy jeszcze ważną własność sprzężenia zespolonego, która przyda się nam za chwilę: sprzężenie zespolone liczby rzeczywistej jest tożsame z tą liczbą, co daje definicję liczby rzeczywistej w postaci warunku $c^* = c$. Ponadto iloczyn liczby zespolonej $a + ib$ z jej sprzężeniem tworzy również liczbę rzeczywistą o postaci $a^2 + b^2$. Nazywamy tę liczbę kwadratem modułu i oznaczamy następująco: $|c|^2 = cc^*$.

Iloczyn skalarny umożliwia zdefiniowanie pojęcia prostopadłości, zwanego ortogonalnością, a także długości wektora. Dwa wektory są ortogonalne, gdy ich iloczyn jest równy zeru. Z kolei długość wektora to (dodatni) pierwiastek z jego iloczynu z samym sobą: $\sqrt{\langle u|u\rangle}$. Oczywiście będzie to liczba rzeczywista, gdyż iloczyn pod pierwiastkiem spełnia warunek $\langle u|u\rangle^* = \langle u|u\rangle$. Dla wektorów o długości jednostkowej $\langle u|u\rangle = 1$.

Każda przestrzeń wektorowa ma określony wymiar. Wymiarem danej przestrzeni nazywamy maksymalną liczbę niezależnych liniowo wektorów, czyli takich, które nie mogą być wyrażone przez swoje wzajemne kombinacje liniowe. W przestrzeni skończenie wymiarowej (a tylko takimi będziemy się tutaj zajmować) można wyróżnić N niezależnych wektorów, takich że każdy inny wektor da się przedstawić za pomocą ich kombinacji. Takie wektory nazywamy bazowymi. Jeśli dodatkowo wektory bazowe są do siebie ortogonalne i mają jednostkową długość, powstałą bazę nazywamy ortonormalną. Wektory bazowe często symbolizuje się jako $|e_1\rangle$, $|e_2\rangle$, …, $|e_N\rangle$. Oczywiście liczba wektorów bazowych jest równa wymiarowi przestrzeni.

Wprowadźmy teraz kluczowe pojęcie operatora liniowego na przestrzeni wektorowej. Operator liniowy jest to funkcja, która przekształca wektory na wektory, spełniając przy tym warunek liniowości:

$$A(a|u\rangle + b|v\rangle) = aA|u\rangle + bA|v\rangle.$$

Niezmiernie ważną kategorię operatorów liniowych stanowią tzw. operatory hermitowskie (termin pochodzi od nazwiska francuskiego matematyka Charlesa Hermite'a), zwane również samosprzężonymi. Ich swobodna definicja jest następująca: są to operatory, które działają w taki sam sposób na pierwszy i drugi element w iloczynie skalarnym. Dokładniej, zachodzi następująca równość dla dowolnych dwóch wektorów $|u\rangle$ i $|v\rangle$:

$$\langle u|Av\rangle = \langle Au|v\rangle.$$

Definicję tę można przeformułować, zamieniając kolejność czynników mnożenia po prawej stronie:

$$\langle u|Av\rangle = \langle v|Au\rangle^*.$$

Wprowadźmy teraz pojęcie sprzężenia dla operatorów. Operatorem sprzężonym do danego operatora A nazwiemy taki operator $A^\dagger$, który spełnia następującą równość:

$$\langle u|A^{\dagger}v\rangle = \langle v|Au\rangle^{*}.$$

Widzimy zatem, że warunek bycia operatorem hermitowskim można intuicyjnie przepisać w postaci równania $A = A^{\dagger}$, co uzasadnia nazwę „operator samosprzężony", czyli taki, którego sprzężenie jest z nim tożsame.

Następne pojęcia to pojęcia wektora własnego i wartości własnej dla danego operatora. Dany wektor nazywamy wektorem własnym dla operatora A, gdy działanie A na tym wektorze jest równoważne przemnożeniu go przez liczbę. Liczbę tę nazwiemy odpowiadającą temu wektorowi wartością własną:

$$A|u_a\rangle = a|u_a\rangle.$$

Nie wszystkie operatory posiadają wektory własne (np. operatory obrotów ich nie mają), ale jeśli operator je posiada, to ma ich więcej niż jeden. Bardzo ważnym faktem charakteryzującym operatory hermitowskie jest to, że wektory własne odpowiadające różnym wartościom własnym muszą być do siebie ortogonalne. Niech będą dane wektory $|u_a\rangle$ i $|u_b\rangle$, takie że zachodzą następujące równości dla $a \neq b$:

$$\begin{aligned} A|u_a\rangle &= a|u_a\rangle, \\ A|u_b\rangle &= b|u_b\rangle. \end{aligned} \tag{7.3}$$

Przemnóżmy obie strony pierwszego równania przez wektor $|u_b\rangle$, a drugiego przez $|u_a\rangle$:

$$\begin{aligned} \langle u_b|A|u_a\rangle &= a\langle u_b|u_a\rangle \\ \langle u_a|A|u_b\rangle &= b\langle u_a|u_b\rangle. \end{aligned}$$

Biorąc sprzężenie zespolone obu stron pierwszego równania i korzystając z definicji operatora hermitowskiego, przekształcamy to równanie do postaci:

$$\langle u_a|A|u_b\rangle = a^{*}\langle u_a|u_b\rangle,$$

skąd mamy:

$$a^{*}\langle u_a|u_b\rangle = b\langle u_a|u_b\rangle.$$

Równanie to może być spełnione, tylko gdy $\langle u_a|u_b\rangle = 0$, co oznacza ortogonalność wektorów.

W podobny sposób można udowodnić, że wartości własne dla operatorów hermitowskich muszą być liczbami rzeczywistymi. Pozostawiam to jako ćwiczenie do wykonania (wskazówka: przemnóżcie pierwsze z równań (7.3) przez wektor $|u_a\rangle$ i weźcie sprzężenie zespolone obu stron). Ta cecha operatorów hermitowskich będzie bardzo przydatna przy ich interpretacji fizycznej. Zakończymy natomiast ten przegląd podstawowych faktów na temat operatorów, przytaczając bez dowodu niezmiernie ważne twierdzenie spektralne. Głosi ono, z grubsza rzecz biorąc, że wektory własne operatora hermitowskiego tworzą ortogonalną bazę danej przestrzeni. Oznacza to, że działanie takiego operatora na dowolnym wektorze można przedstawić jako rozłożenie tego wektora na składowe w tej bazie wektorów własnych, przemnożenie każdej składowej przez odpowiadającą jej wartość własną i zsumowanie wyników.

Interpretacja fizyczna wprowadzonych elementów formalizmu matematycznego jest następująca. Wektory przestrzeni reprezentują stany fizyczne danego układu, a ich suma – superpozycję stanów. Z kolei operatory hermitowskie są formalną reprezentacją wielkości mie-

rzalnych (zwanych obserwablami). Wektor własny operatora reprezentującego obserwablę wyznacza stan, w którym układ ma dobrze określoną wartość tej obserwabli – jest ona równa wartości własnej tego wektora. Widać teraz, jak istotny jest matematyczny fakt, że operatory hermitowskie mają tylko rzeczywiste wartości własne, gdyż rezultaty pomiarowe są dane w postaci takich liczb, a nie ogólnie liczb zespolonych. Co natomiast ze stanami, które są nietrywialną kombinacją wektorów własnych? W takiej sytuacji, jak wiemy, istnieją tylko prawdopodobieństwa, że pomiar ujawni jedną z możliwych wartości własnych. Reguła obliczania tych prawdopodobieństw jest prosta: są one dane przez kwadraty współczynników rozkładu wektora na poszczególne składowe wzdłuż wektorów własnych. Dokładniej można to przedstawić tak: niech $|u\rangle$ będzie stanem danego układu fizycznego, a $|v_a\rangle$ wektorem własnym operatora-obserwabli A odpowiadającym wartości własnej a. Zakładamy ponadto, że wektory $|u\rangle$ i $|v_a\rangle$ są znormalizowane, czyli ich długość wynosi 1. W takiej sytuacji prawdopodobieństwo, że pomiar układu w stanie $|u\rangle$ ujawni wartość a, wynosi (jest to tzw. reguła Borna):

$$P_u(A = a) = |\langle u|v_a\rangle|^2.$$

Często jesteśmy zainteresowani nie prawdopodobieństwami poszczególnych rezultatów, a „globalną" informacją na temat średniej z rezultatów pomiarów w dużej ich serii. Taką średnią wartość (zwaną również wartością oczekiwaną, co jest jednak dość mylące[14]) można łatwo obliczyć, biorąc prawdopodobieństwo każdego wyniku z osobna, mnożąc go przez ten wynik i sumując. Nietrudno pokazać, że w rezultacie otrzymamy następującą formułę reprezentującą uśrednioną wartość wielkości A w stanie $|u\rangle$:

$$\langle u|A|u\rangle. \tag{7.4}$$

Zatrzymajmy się jeszcze przy kwestii notacji Diraca i jej użyteczności, na przykładzie dodatkowego pojęcia formalnego, które może oddać pewne usługi, choć nie jest niezbędne do opisu podstawowych cech świata kwantowego. W zapisie Diracowskim można potraktować wyrażenie na iloczyn skalarny wektorów $\langle u|v\rangle$ jako składające się z dwóch niezależnych składników: $\langle u||v\rangle$. Drugi element rozpoznajemy jako ket $|v\rangle$. Czym jest natomiast pierwszy składnik $\langle u|$, będący niejako zwierciadlanym odbiciem keta? Okazuje się, że można go zinterpretować jako pewien nowy obiekt, zwany „bra". Nazwa ta jest pochodna względem angielskiego słowa „bracket", oznaczającego nawias. Jak widać, wektory bra są pierwszym składnikiem nawiasu oznaczającego iloczyn skalarny, a kety drugim.

Jaką rolę pełni nowy obiekt bra? Formalna definicja jest następująca: dla każdego keta $|u\rangle$ odpowiadający mu wektor bra $\langle u|$ jest funkcjonałem liniowym, który każdemu wektorowi z przestrzeni wektorowej (czyli każdemu ketowi) przypisuje liczbę, będącą rezultatem iloczynu tego keta i keta $|u\rangle$. W zapisie matematycznym:

$$\langle u|(|v\rangle) = \langle u|v\rangle.$$

Po lewej stronie tej równości mamy obiekt $\langle u|$ „działający" na $|v\rangle$, a po prawej liczbę. Ci z czytelników, którzy odważnie przeczytali paragrafy z gwiazdką z poprzedniego rozdziału,

[14] Dlaczego jest to mylące? Odpowiedź jest prosta – bardzo często „wartość oczekiwana" danej wielkości jest wartością, której w ogóle nie powinniśmy oczekiwać w eksperymencie. Na przykład wartością oczekiwaną spinu w stanie singletowym jest 0, choć wiadomo, że spin fermionu nie może przyjmować wartości zerowej. Jednakże użycie terminu „wartość oczekiwana" przyjęło się powszechnie na określenie średniej z wielu powtórzeń eksperymentu.

powinni rozpoznać, że tak zdefiniowane obiekty bra tworzą tzw. przestrzeń dualną do wyjściowej przestrzeni, o której mówiliśmy przy okazji definicji przestrzeni stycznej i pojęcia wektorów kowariantnych oraz kontrawariantnych. Przestrzeń dualna do danej jest w pewnym sensie jej odbiciem zwierciadlanym. Aby „przejść" z jednej przestrzeni do drugiej, należy po pierwsze zamienić każdy ket $|u\rangle$ na odpowiadający mu wektor bra $\langle u|$. Po drugie, musimy również zamienić wszystkie liczby w odpowiednich kombinacjach ketów na ich sprzężenia zespolone. Na przykład odpowiednikiem keta $a|u\rangle + b|v\rangle$ będzie następujący wektor bra: $a^*\langle u| + b^*\langle v|$. I wreszcie operatorowi A w przestrzeni ketów odpowiada operator sprzężony $A^\dagger$ w przestrzeni dualnej wektorów bra.

Na tym nie kończy się niezwykła elastyczność notacji Diraca. Rozważmy na przykład następujący dziwny twór:

$$|u\rangle\langle v|.$$

Spróbujmy się domyślić, co może oznaczać taki zapis. Po prawej stronie mamy wektor bra, który jest funkcjonałem liniowym, a zatem przy zastosowaniu do dowolnego wektora daje nam liczbę. Natomiast lewa część powyższej formuły to ket $|u\rangle$. Pomnożenie keta przez liczbę (kolejność nie gra roli) daje nam nowy ket. Zatem cała powyższa formuła oznacza operator, przekształcający kety w kety. Można łatwo sprawdzić, że będzie to operator liniowy. Jest to tzw. operator rzutowy na prostą, wzdłuż której leży wektor $|u\rangle$, w kierunku prostopadłym do kierunku wektora $|v\rangle$. Łatwo sprawdzić, że każdy wektor transformowany jest na wektor, który jest wielokrotnością $|u\rangle$, a do tego wektory prostopadłe do $|v\rangle$ są transformowane na wektor zerowy.

Szczególnym przypadkiem operatora rzutowego jest tzw. ortogonalny jednowymiarowy operator rzutowy. Ma on postać:

$$|u\rangle\langle u|.$$

Można pokazać, że jest to operator hermitowski, a zatem reprezentuje wielkość mierzalną.[15] Przyjmuje ona tylko dwie dopuszczalne wartości: 0 i 1. Wartość 1 jest związana z wektorem $|u\rangle$, a 0 z wszystkimi wektorami do niego ortogonalnymi. Oznacza to, że operator ten „mierzy", czy układ znajduje się w stanie $|u\rangle$ (wartość 1), czy w stanie do niego ortogonalnym (wartość 0). W szczególności, dla dowolnej obserwabli A, operator rzutowy na jeden z jej wektorów własnych mierzy, czy układ posiada wartość skorelowaną z tym wektorem, czy nie.[16]

Pamiętamy, że niektóre wielkości mierzalne w mechanice kwantowej nie mogą być jednocześnie dokładnie określone. Reprezentujące te wielkości operatory nazywa się niekompatybilnymi. Dwa operatory A i B są niekompatybilne, gdy jest pewien wektor własny jed-

[15] Dowód tego faktu pokazuje, jak niezmiernie użyteczna i intuicyjna jest notacja Diraca. Przytoczę go tutaj dla ilustracji. Weźmy dowolne dwa wektory $|a\rangle$ i $|b\rangle$. Jeśli P jest naszym operatorem rzutowym, to wyrażenie $\langle a|Pb\rangle$ można zapisać jako $\langle a|u\rangle\langle u|b\rangle$. Zauważmy, że oba elementy symbolu $|u\rangle\langle u|$ zostały „rozdzielone" i „doczepione" każdy do jednego z dwóch wektorów $|a\rangle$ i $|b\rangle$ tworząc dwa iloczyny skalarne! Jest to zupełnie poprawne matematycznie. W celu obliczenia $\langle Pa|b\rangle$ zamieńmy to wyrażenie na $\langle b|Pa\rangle^*$, a następnie znów wstawmy $|u\rangle\langle u|$ w miejsce P. Otrzymamy $\langle b|u\rangle^*\langle u|a\rangle^*$, czyli $\langle u|b\rangle\langle a|u\rangle$, co jest identyczne z wcześniejszym rezultatem. Zatem P jest hermitowski.

[16] Wprowadza się też operatory rzutowe na przestrzenie wielowymiarowe. Można je przedstawić w postaci sumy operatorów $|u_i\rangle\langle u_i|$ gdzie $|u_i\rangle$ są do siebie wzajemnie ortogonalne.

nego z nich, który nie jest wektorem własnym drugiego. Alternatywnie, A i B są kompatybilne, gdy mają dokładnie identyczne wektory własne. Warunek ten można przedstawić równoważnie w następujący sposób: kompatybilne operatory komutują ze sobą (są przemienne, czyli $AB = BA$). Weźmy dowolny wektor własny $|u\rangle$ operatora A, który z założenia jest również wektorem własnym dla B. Zatem $AB|u\rangle = ab|u\rangle$ oraz $BA|u\rangle = ba|u\rangle = ab|u\rangle$, gdzie a i b są odpowiednimi wartościami własnymi. Ponieważ każdy wektor da się przedstawić jako liniowa kombinacja wektorów własnych, działanie AB na dowolnym wektorze jest takie samo jak BA, czyli operatory komutują.

Zilustrujmy powyższe abstrakcyjne rozważania jakimś konkretnym przykładem. Najlepszy do tego celu będzie spin połówkowy, który służył nam w różnych sytuacjach do objaśniania złożoności zachowania świata kwantowego. Wektory w skończenie-wymiarowej przestrzeni można reprezentować za pomocą kolumn liczb, które jak pamiętamy z poprzednich rozdziałów, reprezentują ich składowe w pewnym układzie współrzędnych – czyli w pewnej bazie. Jeśli np. wektor $|u\rangle$ w dwuwymiarowej przestrzeni ma postać sumy $2|e_1\rangle + 3|e_2\rangle$, to jego reprezentacja kolumnowa w tej bazie będzie następująca: $\binom{2}{3}$. Dwuwymiarowa przestrzeń Hilberta wystarcza do opisania spinów połówkowych, które mają dwie wartości. Naszym zadaniem będzie wyznaczenie postaci operatorów dla spinów w trzech kierunkach w przestrzeni fizycznej: s_x, s_y i s_z. Jako bazę przestrzeni stanów (nie należy jej mylić z przestrzenią fizyczną) wybierzmy dwa wektory własne dla operatora s_z. Oczywiście reprezentacje kolumnowe wektorów bazy są bardzo proste: to $\binom{1}{0}$ i $\binom{0}{1}$. Z kolei operatory linowe będą miały postać macierzy o dwóch wierszach i dwóch kolumnach, a ich działanie na wektor jest dane w postaci znanej zasady mnożenia kolumny przez macierz. Zatem szukamy macierzy, która w działaniu na wektor $\binom{1}{0}$ odtwarza ten sam wektor przemnożony przez jego wartość własną – i podobnie dla drugiego wektora. Jako wartości własne przyjmujemy liczby ½ i –½ (dla uproszczenia pomijamy stałą Plancka $\hbar$). Równania określające operator s_z są więc następujące:

$$s_z \binom{1}{0} = \frac{1}{2}\binom{1}{0},$$

$$s_z \binom{0}{1} = -\frac{1}{2}\binom{0}{1}.$$

Nietrudno sprawdzić, że macierz spełniająca powyższe dwa warunki ma następującą postać:

$$s_z = \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Można się o tym przekonać, mnożąc liczby z pierwszego wiersza przez odpowiednio pierwszy i drugi element danej kolumny, co da nam pierwszy element nowego wektora, a następnie powtarzając to dla drugiego wiersza. Powyższa macierz jest jedną z trzech tzw. macierzy Pauliego, reprezentujących spin. Zauważmy, że ma ona tzw. zdiagonalizowaną postać, w której jedyne niezerowe elementy znajdują się na przekątnej. Jest to zawsze prawdą, jeśli wektory własne danej macierzy są wybrane jako wektory bazy. Liczby na przekątnej są wtedy wartościami własnymi.

W celu wyznaczenia pozostałych dwóch macierzy odwołamy się do empirycznego faktu, że prawdopodobieństwa obu wartości dla spinów w kierunkach x i y są równe ½ w stanach

z dobrze określonym spinem *z*. Oznacza to, że wektory własne dla operatorów s_x i s_y powinny być liniowymi kombinacjami wektorów własnych dla s_z ze współczynnikami, których kwadraty modułów są sobie równe. Przyjmijmy, że wektor własny dla operatora s_x odpowiadający rezultatowi ½ („do góry") ma postać:

$$\frac{1}{\sqrt{2}}\begin{pmatrix}1\\0\end{pmatrix}+\frac{1}{\sqrt{2}}\begin{pmatrix}0\\1\end{pmatrix}=\begin{pmatrix}\frac{1}{\sqrt{2}}\\\frac{1}{\sqrt{2}}\end{pmatrix}.$$

Drugi wektor własny odpowiadający rezultatowi „w dół" zapiszemy następująco:

$$\frac{1}{\sqrt{2}}\begin{pmatrix}1\\0\end{pmatrix}-\frac{1}{\sqrt{2}}\begin{pmatrix}0\\1\end{pmatrix}=\begin{pmatrix}\frac{1}{\sqrt{2}}\\-\frac{1}{\sqrt{2}}\end{pmatrix}.$$

Łatwo sprawdzić, że powyższe wektory są ortogonalne do siebie. Ich iloczyn, liczony standardowo jako suma iloczynów pierwszych składowych i drugich składowych, jest równy zeru.[17] Możemy teraz zapisać równania analogiczne do tych dla spinu s_z, charakteryzujące działanie poszukiwanej macierzy s_x, na powyższe wektory:

$$s_x\begin{pmatrix}\frac{1}{\sqrt{2}}\\\frac{1}{\sqrt{2}}\end{pmatrix}=\frac{1}{2}\begin{pmatrix}\frac{1}{\sqrt{2}}\\\frac{1}{\sqrt{2}}\end{pmatrix},$$

$$s_x\begin{pmatrix}\frac{1}{\sqrt{2}}\\-\frac{1}{\sqrt{2}}\end{pmatrix}=-\frac{1}{2}\begin{pmatrix}\frac{1}{\sqrt{2}}\\-\frac{1}{\sqrt{2}}\end{pmatrix}.$$

Rozwiązanie prostych równań liniowych dostarcza nam informacji na temat formy poszukiwanej macierzy:

$$s_x=\frac{1}{2}\begin{pmatrix}0&1\\1&0\end{pmatrix}.$$

W przypadku trzeciej macierzy, reprezentującej spin w kierunku *y*, powinniśmy nałożyć analogiczne warunki na jej wektory własne – tj. powinny one być liniowymi kombinacjami wektorów własnych dla spinów s_z oraz s_x ze współczynnikami o tych samych modułach. Tu jednak pojawia się problem. Okazuje się, że dla liczb rzeczywistych nie istnieją rozwiązania odpowiednich równań reprezentujących te nałożone warunki. Zatem dwuwymiarowa przestrzeń Hilberta nad liczbami rzeczywistymi nie wystarcza do opisu wszystkich składowych spinu połówkowego. Potrzebne w tym celu są liczby zespolone. Oto postać kolumnowa wektorów własnych poszukiwanego operatora s_y, które pozostają w odpowiednich relacjach do wektorów własnych s_z i s_x:[18]

$$\frac{1}{2}\begin{pmatrix}1-i\\1+i\end{pmatrix},\frac{1}{2}\begin{pmatrix}1-i\\-1+i\end{pmatrix}.$$

[17] Pamiętajmy jednak, że jeśli składowe pierwszego wektora zawierają liczby zespolone, to przy obliczaniu iloczynu skalarnego należy wziąć ich sprzężenia zespolone.

[18] Sprawdźcie, że wektory te rozkładają się w odpowiedni sposób na wektory własne operatorów s_z i s_x. Najlepiej to zrobić obliczając odpowiednie iloczyny skalarne, które dadzą nam współczynniki rozkładów.

Są to wektory własne następującej macierzy (warto to sprawdzić):

$$s_y = \frac{1}{2}\begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}.$$

Mamy już macierzową postać trzech operatorów s_x, s_y i s_z, reprezentujących spiny w trzech prostopadłych kierunkach.[19] Widzimy, że operatory te mają różne wektory własne, zatem cząstka nie może znajdować się w stanie, w którym spiny w różnych kierunkach są dobrze określone. Można to wyrazić jeszcze inaczej, pokazując, że macierze s_x, s_y i s_z nie komutują (nie są przemienne). Obliczmy na przykład iloczyn macierzy $s_x s_y$, a następnie $s_y s_x$:

$$s_x s_y = \frac{1}{4}\begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix},$$

$$s_y s_x = \frac{1}{4}\begin{pmatrix} -i & 0 \\ 0 & i \end{pmatrix}.$$

Jak widać, wyniki różnią się od siebie. Jeszcze inny sposób na wyrażenie tego faktu wykorzystuje pojęcie komutatora, czyli kombinacji $s_x s_y - s_y s_x$. Komutator dla nieprzemiennych operacji jest zawsze niezerowy. W naszym wypadku mamy ciekawą prawidłowość, która łączy komutator dwóch macierzy Pauliego z trzecią macierzą:

$$s_x s_y - s_y s_x = i s_z.$$

Okazuje się, że ta relacja zachodzi „cyklicznie” dla każdego komutatora dwóch macierzy Pauliego – jest on równy trzeciej macierzy przemnożonej przez *i*.

Pokażmy jeszcze, jaki będzie rezultat następującej kombinacji macierzy Pauliego:

$$s_x^2 + s_y^2 + s_z^2.$$

Łatwo sprawdzić, że podniesienie do kwadratu każdej z trzech macierzy Pauliego daje następujący rezultat:

$$\frac{1}{4}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

a zatem suma takich trzech elementów to:

$$\frac{3}{4}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Jest to macierz jednostkowa przemnożona przez liczbę $\frac{3}{4}$. Zatem, jak łatwo się przekonać, każdy wektor jest jej wektorem własnym z wartością własną równą $\frac{3}{4}$. Oznacza to, że kwadrat całkowitego wektora spinu ma zawsze taką samą dobrze określoną wartość, niezależną od stanu cząstki.

Przejdźmy teraz do podstawowego równania nierelatywistycznej mechaniki kwantowej, czyli do równania Schrödingera. Jak już wcześniej pisaliśmy, równanie to ma następującą ogólną postać (w zapisie Diraca):

[19] Można jeszcze pokazać, że operatory reprezentowane przez macierze Pauliego są hermitowskie, czyli samosprzężone. Istnieje proste kryterium samosprzężoności macierzy: jest tak wtedy, gdy liczby, które są swoimi „odbiciami” względem przekątnej, są swoimi wzajemnymi sprzężeniami. Widać, że macierze Pauliego spełniają ten warunek.

$$i\hbar\frac{\partial}{\partial t}|\psi\rangle = H|\psi\rangle.$$

Przyjrzyjmy się nieco dokładniej postaci operatora *H*, czyli hamiltonianu. W fizyce klasycznej *H* może być zapisany jako suma energii kinetycznej i potencjalnej:

$$H = \frac{p^2}{2m} + V.$$

W mechanice kwantowej wielkości takie jak pęd są reprezentowane nie przez funkcje, a operatory. Istnieją argumenty oparte na analizie zjawisk falowych, które pokazują, że operator reprezentujący pęd w kierunku osi *x* powinien mieć następującą formę:

$$-i\hbar\frac{\partial}{\partial x}.$$

Wstawiając tę formułę w miejsce pędu, otrzymujemy następującą rozwiniętą formę równania Schrödingera:

$$i\hbar\frac{\partial}{\partial t}|\psi\rangle = \left[-\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)+V(x,y,z)\right]|\psi\rangle. \tag{7.5}$$

Załóżmy, że funkcja falowa $|\psi\rangle$ jest separowalna, tj. może być przedstawiona jako iloczyn części zależnej od czasu i części zależnej od przestrzeni: $|\psi\rangle = U(t)\left|\psi_0(x,y,z)\right\rangle$. Wstawiając tę formułę do powyższego, otrzymamy:

$$i\hbar\left|\psi_0\right\rangle\frac{d}{dt}U(t) = U(t)H\left|\psi_0\right\rangle. \tag{7.6}$$

Jeśli dodatkowo założymy, że hamiltonian danego układu nie zależy od czasu, to możemy z powyższego równania wyznaczyć postać funkcji (operatora) *U*(*t*). Będzie to mianowicie funkcja eksponencjalna:

$$U(t) = e^{-\frac{iH}{\hbar}t}.$$

Ci z czytelników, którzy znają reguły różniczkowania typowych funkcji, być może wiedzą, że pochodna funkcji eksponencjalnej e^x jest dokładnie tą samą funkcją (bardzo zgrabny dowód tego faktu odwołuje się do rozwinięcia tej funkcji w nieskończony szereg potęgowy). Zatem zróżniczkowanie powyższej funkcji po czasie i wstawienie wyniku do lewej strony równania Schrödingera daje nam dokładnie prawą stronę.

Pamiętajmy, że *U*(*t*) jest operatorem, jako że jest on definiowany jako funkcja innego operatora, czyli hamiltonianu *H*. Operator *U*(*t*) jest nazywany operatorem ewolucji, ponieważ zadziałanie nim na wektor stanu początkowego daje stan w chwili *t*. Dodatkowo, *U*(*t*) posiada ciekawą własność zwaną unitarnością. Weźmy iloczyn skalarny dwóch wektorów $\langle u|v\rangle$ i zadziałajmy na każdy z nich operatorem *U*(*t*):

$$\langle U(t)u|U(t)v\rangle = \left\langle e^{-\frac{iH}{\hbar}t}u\middle|e^{-\frac{iH}{\hbar}t}v\right\rangle.$$

Z własności iloczynu skalarnego, a także z definicji sprzężenia hermitowskiego dla operatorów wiemy, że operator działający na pierwszy składnik można „przesunąć” do drugiego

składnika, pod warunkiem, że zostanie on sprzężony. Ale sprzężenie operatora $e^{-\frac{iH}{\hbar}t}$ daje po prostu operator $e^{\frac{iH}{\hbar}t}$. Zatem dostajemy z powyższego:

$$\left\langle u \middle| e^{\frac{iH}{\hbar}t} e^{-\frac{iH}{\hbar}t} v \right\rangle = \langle u|v\rangle.$$

Jak widać, operator $U(t)$ zachowuje iloczyn skalarny, a zatem także długość wektorów. Własność tę nazywa się unitarnością, a sam operator – unitarnym. Można zatem powiedzieć, że ewolucja Schrödingerowska danego wektora stanu jest czymś w rodzaju obrotu – wektor nie zmienia długości, lecz tylko swój kierunek. Wynika stąd także, że jeśli stan początkowy jest wektorem własnym danego operatora, to ewolucja Schrödingerowska transformująca go na inny wektor własny musi przejść przez stany pośrednie, będące superpozycją tych dwóch. Czyli układy kwantowe mają „skłonność" do tracenia dobrze określonych wartości danych wielkości.

Sformułujmy jeszcze jedną ciekawą obserwację dotyczącą równania Schrödingera z hamiltonianem niezależnym od czasu. Proste przekształcenie równania (7.6) daje nam:

$$H\left|\psi_0\right\rangle = i\hbar \frac{1}{U(t)} \frac{d}{dt} U(t) \left|\psi_0\right\rangle .$$

Operator po lewej stronie równania zależy tylko od położenia w przestrzeni, a po prawej stronie od czasu. Jeśli to równanie ma być spełnione dla każdej wartości x, y, z i t, obie strony muszą być równe pewnej stałej liczbie. Zapiszmy to skrótowo:

$$H\left|\psi_0\right\rangle = E\left|\psi_0\right\rangle .$$

Jest to nic innego jak równanie własne dla operatora H, który jest operatorem energii. Rozwiązując to równanie, otrzymamy wektory własne operatora energii oraz jego wartości własne, czyli dozwolone wartości energii, jakie może przyjmować dany układ. W ten sposób możemy się dowiedzieć, np. że cząstka w tzw. studni potencjału, czyli umieszczona w potencjale, który przyjmuje wartość skończoną w pewnym obszarze, a poza tym obszarem jest nieskończony, będzie miała skwantowane poziomy energetyczne. Podobnie wygląda analiza poziomów energetycznych atomu wodoru z potencjałem Coulombowskim $k\frac{q}{r}$. Dodatkowo otrzymamy matematyczną postać funkcji falowych elektronu w atomie wodoru dla kolejnych poziomów energetycznych, czyli kształt tzw. orbitali.

Na zakończenie przyjrzyjmy się matematycznej stronie opisu układów złożonych z wielu cząstek, w tym niezmiernie ważnych stanów splątanych. W celu opisu stanów dwóch układów fizycznych wprowadza się iloczyn tensorowy przestrzeni stanów dla każdego układu z osobna: $\mathcal{H}_1 \otimes \mathcal{H}_2$. Jest to przestrzeń złożona z obiektów o postaci $|u\rangle_1 \otimes |v\rangle_2$, gdzie $|u\rangle_1 \in \mathcal{H}_1$ i $|v\rangle_2 \in \mathcal{H}_2$, oraz wszystkich liniowych kombinacji takich wektorów. Na szczególną uwagę zasługują kombinacje, które nie mogą być przedstawione w postaci prostego iloczynu, takie jak np.:

$$|u\rangle_1 \otimes |v\rangle_2 + |v\rangle_1 \otimes |u\rangle_2.$$

Stany takie nazywamy splątanymi. Przykładem rozważanym wcześniej w niniejszym rozdziale jest stan singletowy spinu dwóch fermionów, np. elektronów (zwany również stanem EPR):

$$\frac{1}{\sqrt{2}}(|z_+\rangle\otimes|z_-\rangle - |z_-\rangle\otimes|z_+\rangle). \tag{7.7}$$

Spróbujmy obliczyć, jakie będą prawdopodobieństwa uzyskania rezultatów pomiarowych spinów dla obu cząstek w powyższym stanie. W tym celu musimy formalnie przedstawić w przestrzeni $\mathcal{H}_1 \otimes \mathcal{H}_2$ operatory reprezentujące własności każdej cząstek z osobna. Generalnie: jeśli A jest operatorem działającym w przestrzeni $\mathcal{H}_1$ i reprezentującym pewną wielkość pierwszej cząstki, to iloczyn $A \otimes I$, gdzie I jest operatorem identycznościowym w przestrzeni $\mathcal{H}_2$ reprezentuje dokładnie tę samą wielkość cząstki pierwszej w przestrzeni $\mathcal{H}_1 \otimes \mathcal{H}_2$. Zasada działania iloczynów operatorów $A \otimes B$ w przestrzeni $\mathcal{H}_1 \otimes \mathcal{H}_2$ jest prosta: każdy z operatorów A i B działa w swojej przestrzeni, a rezultaty ich działania są następnie mnożone tensorowo. Zatem rezultat działania $A \otimes B$ na wektor $|u\rangle \otimes |v\rangle$ to po prostu $A|u\rangle \otimes B|v\rangle$.

Będziemy chcieli policzyć prawdopodobieństwa uzyskania danej wartości spinu dla obu cząstek oddzielnie, jak również prawdopodobieństwo uzyskania dwóch wartości naraz. Najwygodniej to zrobić, używając odpowiednich operatorów rzutowych. Operator $|z_+\rangle\langle z_+|$ reprezentuje własność „spin do góry", a zatem iloczyn $|z_+\rangle\langle z_+| \otimes I$ będzie przedstawiał tę samą własność w odniesieniu do pierwszej cząstki, a iloczyn $I \otimes |z_+\rangle\langle z_+|$ w odniesieniu do drugiej. Z kolei iloczyn dwóch operatorów rzutowych spinu $|z_+\rangle\langle z_+| \otimes |z_+\rangle\langle z_+|$ wyraża tę samą własność „spin do góry" przypisaną obu cząstkom. Możemy teraz policzyć wartości oczekiwane w stanie (7.7) dla wszystkich trzech operatorów rzutowych, zgodnie z przedstawioną wcześniej formułą (7.4). Ponieważ operator rzutowy może przyjmować tylko wartości 1 (kiedy układ posiada daną własność) lub 0 (kiedy jej nie posiada), jego wartość oczekiwana jest automatycznie równa prawdopodobieństwu, że układ istotnie ma daną własność.

Działając operatorem $|z_+\rangle\langle z_+| \otimes I$ na wektor (7.7), otrzymamy:

$$\frac{1}{\sqrt{2}}(|z_+\rangle\langle z_+|z_+\rangle \otimes |z_-\rangle - |z_+\rangle\langle z_+|z_-\rangle \otimes |z_+\rangle) = \frac{1}{\sqrt{2}}|z_+\rangle\otimes|z_-\rangle$$

(korzystamy z faktu, że $\langle z_+|z_-\rangle = 0$ ze względu na ortogonalność oraz $\langle z_+|z_+\rangle = 1$ z założenia jednostkowej długości). Obliczając iloczyn skalarny wektora (7.7) z powyższym, otrzymujemy:[20]

$$\frac{1}{2}(\langle z_+|z_+\rangle\langle z_-|z_-\rangle - \langle z_+|z_-\rangle\langle z_-|z_+\rangle) = \frac{1}{2}.$$

Taki sam rezultat dostaniemy dla drugiego operatora rzutowego $I \otimes |z_+\rangle\langle z_+|$. Zatem dla obu cząstek prawdopodobieństwo uzyskania rezultatu „spin do góry" jest takie samo co „spin do dołu". Policzmy teraz rezultat działania trzeciego operatora $|z_+\rangle\langle z_+| \otimes |z_+\rangle\langle z_+|$ na wektor (7.7):

$$\frac{1}{\sqrt{2}}(|z_+\rangle\langle z_+|z_+\rangle \otimes |z_+\rangle\langle z_+|z_-\rangle - |z_+\rangle\langle z_+|z_-\rangle \otimes |z_+\rangle\langle z_+|z_+\rangle) = \mathbf{0}.$$

Skoro wynik okazał się wektorem zerowym, jego iloczyn z każdym innym wektorem jest również równy zeru. Zatem prawdopodobieństwo tego, że obie cząstki będą miały spin do góry jest zerowe. Mamy więc przykład kwantowej nielokalności. Chociaż każda cząstka

[20] Należy jeszcze pamiętać, że iloczyn skalarny dwóch wektorów $|u\rangle \otimes |v\rangle$ i $|z\rangle \otimes |w\rangle$ liczymy mnożąc iloczyny pierwszego składnika z pierwszym i drugiego z drugim: $\langle u|z\rangle\langle v|w\rangle$.

wzięta z osobna może ujawnić obie wartości spinu (góra lub dół), to jednak muszą one być zawsze przeciwne dla obu cząstek. Jeśli jedna z nich pokaże wynik „do góry", druga musi pokazać „w dół".

Pytania i problemy

1. Jakie są eksperymentalne podstawy dla hipotezy kwantyzacji światła (fal elektromagnetycznych)?

2. Omów doświadczenie Sterna-Gerlacha i doświadczenie z interferencją elektronów (doświadczenie z dwiema szczelinami). Jakie płyną z nich wnioski dotyczące posiadania własności przez obiekty kwantowe?

3. Omów pojęcie superpozycji stanów oraz jego probabilistyczną interpretację.

4. Przedstaw kwantową interpretację własności mierzalnych (obserwabli), pojęcie stanu własnego oraz pojęcie niekompatybilności dla obserwabli. Jakie konsekwencje empiryczne ma niekompatybilność obserwabli, np. pędu i położenia?

5. Jak definiujemy stany splątane w mechanice kwantowej?

6. Przedstaw argument EPR za niekompletnością probabilistycznego opisu mechaniki kwantowej. Jaką rolę w tym argumencie odgrywa zasada lokalności?

7. Sformułuj dwie wersje zasady lokalności: zasadę niezależności od parametru i niezależności od rezultatu. Która z tych wersji jest założona w argumencie EPR?

8. Dokonaj analizy argumentu prowadzącego do nierówności Bella (w wersji CHSH). Czy możliwe jest zastąpienie mechaniki kwantowej nie-probabilistyczną i lokalną teorią, która daje takie same przewidywania empiryczne co standardowa teoria?

9. Jak pogodzić deterministyczny charakter równania Schrödingera z występowaniem w mechanice kwantowej nieredukowalnego pojęcia prawdopodobieństwa?

10. Omów problem pomiaru, korzystając z abstrakcyjnego ujęcia procesu pomiarowego za pomocą unitarnego operatora ewolucji. Przedstaw trzy tezy dotyczące pomiaru, które łącznie prowadzą do sprzeczności.

11. Porównaj standardową interpretację mechaniki kwantowej opartą na pojęciu kolapsu z interpretacją GRW. W jaki sposób ta ostatnia wyjaśnia fakt redukcji funkcji falowej mierzonego obiektu w zetknięciu z makroskopowym urządzeniem pomiarowym?

12. Jakie są główne założenia i trudności wieloświatowej interpretacji mechaniki kwantowej?

13. Omów statystyczne zachowanie dwóch rodzajów cząstek kwantowych (fermionów i bozonów) i porównaj je z zachowaniem cząstek klasycznych. Jakie konsekwencje dla kwestii odróżnialności cząstek ma istnienie statystyk kwantowych?

Literatura uzupełniająca

Istnieje wiele popularnych i quasi-popularnych wprowadzeń do mechaniki kwantowej. Osobiście dla początkujących polecałbym książkę (niestety tylko po angielsku): D. Albert, *Quantum Mechanics and Experience*, Harvard University Press, Cambridge, Mass 1992.

Bardzo użyteczne wprowadzenie do pojęć teorii kwantowej znajdziemy w rozdziale 6. cytowanej już książki słynnego angielskiego fizyka: R. Penrose, *Nowy umysł cesarza*, PWN, Warszawa 1995.

Dostępny jest polski przekład artykułu wprowadzającego do metafizycznych konsekwencji teorii kwantowej: T. Maudlin „Metafizyczne implikacje fizyki kwantowej", *Roczniki Filozoficzne*, Tom LXI , numer 4, 2021.

Następująca praca zbiorowa obejmuje prawie wszystkie aspekty filozoficznych problemów mechaniki kwantowej: C. Friebe , M. Kuhlmann, H. Lyre, P.M. Näger, O. Passon, M. Stöckler, *The Philosophy of Quantum Physics*, Springer , Cham 2018.

Problem nieodróżnialności obiektów kwantowych jest szczegółowo analizowany w monografii: T. Bigaj, *Identity and Indiscernibility in Quantum Mechanics*, Palgrave-Macmillan, Cham 2022.

ZAMIAST ZAKOŃCZENIA

Nasza podróż po fizyce i jej filozoficznych aspektach dobiegła kresu. Przebyliśmy długą drogę, mierzoną zarówno wiekami, jak i stopniem komplikacji teorii fizycznych i ich aparatu formalno-pojęciowego. Zamiast podejmować próbę syntezy czy formułować ogólne konkluzje zwróćmy uwagę, że w istocie znaleźliśmy się dopiero na początku drogi. Niniejsze wprowadzenie nie objęło wielu niezwykle ważnych obszarów fizyki współczesnej, mających kapitalne znaczenie dla filozoficznych dyskusji prowadzonych w ostatnich latach. Wspomnijmy może na zakończenie o tych „wielkich nieobecnych", do których mam nadzieję, będzie jeszcze okazja powrócić. Jedną z dziedzin, która od zawsze leżała na styku filozofii i nauki (a także religii) jest kosmologia, czyli nauka o wszechświecie jako całości. Pierwsze pytania o charakterze kosmologicznym dotyczyły takich kwestii jak pytanie, czy wszechświat jest czasowo i przestrzennie ograniczony, czy też rozciąga się w nieskończoność. Były to zagadnienia głównie o charakterze spekulatywnym, jako że nie istniały możliwości poddania ich badaniom empirycznym. Dopiero rozwój astronomicznych technik obserwacyjnych od końca dziewiętnastego wieku pozwolił na bliższe przyjrzenie się megaskopowej strukturze wszechświata.

Najsłynniejszym odkryciem dwudziestowiecznej kosmologii było zaobserwowanie zjawiska rozszerzania wszechświata, a następnie zdobycie empirycznych danych w postaci tzw. promieniowania reliktowego (tła) sugerującego istnienie początkowej osobliwości zwanej Wielkim Wybuchem (*Big Bang*, co dosłownie należałoby raczej tłumaczyć jako „wielki huk"). Zgodnie z obecnym stanem wiedzy, nasz wszechświat liczy sobie nieco poniżej czternastu miliardów lat, podczas których przechodził przez wiele faz ekspansji, co trwa do chwili obecnej. Ważne jest, aby pamiętać, że rozszerzanie wszechświata nie polega na ucieczce gwiazd i galaktyk, ale na „rozdęciu" samej przestrzeni w taki sposób, że żaden punkt we wszechświecie nie jest wyróżniony jako centrum, z którego wybiegają wszystkie obiekty, jak w wypadku zwykłych eksplozji. Teoretyczną podstawą współczesnej kosmologii jest ogólna teoria względności, umożliwiająca wprowadzenie pojęcia globalnej krzywizny. W standardowych modelach wszechświata (tzw. modele FLRW, Friedmana-Lemaitre'a-Robertsona-Walkera) można opisać wszechświat jako przestrzennie ograniczony z dodatnią globalną krzywizną (jak w wypadku sfery) lub też jako nieograniczony z krzywizną zerową bądź ujemną (jak na powierzchni siodła). W modelu o dodatniej krzywiźnie wszechświat nie rozszerza się w nieskończoność: po etapie rozszerzania następuje etap „kurczenia", prowadzący

do „Wielkiej Zapaści" (*Big Crunch*). Pozostałe modele umożliwiają nieograniczoną ekspansję (sprawa się nieco komplikuje, jeśli do równań dodamy tzw. stałą kosmologiczną). Podstawowa teoria Wielkiego Wybuchu zetknęła się z wieloma teoretycznymi i empirycznymi problemami, co doprowadziło m.in. do sformułowania nowej hipotezy tzw. wszechświata inflacyjnego czy też wprowadzenia zagadkowych i kontrowersyjnych pojęć ciemnej materii i ciemnej energii.

Przechodząc z poziomu megazjawisk na poziom mikroskopowy, musimy przede wszystkim zauważyć, że standardowa mechanika kwantowa (zwana również nierelatywistyczną) nie jest ostatnim słowem fizyki. Już w momencie jej tworzenia zauważono poważne ograniczenia związane z faktem, że fundamentalne dla tej teorii równanie Schrödingera nie jest relatywistycznie niezmiennicze. Próby „urelatywistycznienia" mechaniki kwantowej doprowadziły do powstania tzw. kwantowej teorii pola, która jest jeszcze bardziej zdumiewająca niż jej poprzedniczka. Na przykład umożliwia ona opisanie sytuacji, w których liczba cząstek określonego rodzaju nie jest dobrze określona (stan jest superpozycją stanów o różnych liczbach cząstek). Wprowadza się tutaj również zagadkowe pojęcie cząstek wirtualnych, których wytworzenie jest zasadniczo wykluczone prawami zachowania masy i energii, a jednak mogą się one pojawić na „mgnienie oka". Kwantowa teoria pola opisuje znane z fizyki klasycznej pola, takie jak pole elektromagnetyczne, za pomocą dyskretnych kwantów – cząstek przenoszących dane oddziaływanie. W wypadku oddziaływań elektromagnetycznych takimi cząstkami są fotony, natomiast typowe dla mikroświata oddziaływania silne i słabe przenoszone są odpowiednimi bozonami (takimi jak gluony, mezony czy bozony pośredniczące).

Zasadniczym problemem filozoficznym związanym z kwantową teorią pola jest pytanie, co w tej teorii jest bardziej fundamentalne: rozciągłe pole czy dyskretne cząstki. Jest to szczególnie ważne w kontekście zagadnienia indywidualności i tożsamości obiektów kwantowych, gdyż kryteria tożsamości dla pól i cząstek są odmienne. Inny problem wiąże się z notoryczną matematyczną trudnością pojawiającą się przy próbie dokładnego opisu odpowiednich oddziaływań. Okazuje się, że pewne wielkości charakteryzujące te oddziaływania „uciekają" do nieskończoności, w wyniku czego niektóre równania matematyczne tracą sens. Fizycy wymyślili kilka sposobów radzenia sobie z tą trudnością, zwanych „procedurami renormalizacyjnymi", co przyprawia matematyków o ból głowy. Niestety nie zawsze takie procedury są dostępne (podstawową teorią, dla której udało się dokonać renormalizacji, jest elektrodynamika kwantowa QED). Powstaje oczywiście pytanie, czy renormalizacja nie jest przypadkiem czysto formalnym zabiegiem *ad hoc*, bez głębszego teoretycznego uzasadnienia. Istnieje na ten temat wiele argumentów za i przeciw.

Szczegółowe badanie świata cząstek elementarnych odkryło ogromną różnorodność występujących w nim obiektów, daleko wykraczającą poza znane elektrony, protony czy neutrony. Tak zwany model standardowy klasyfikuje cząstki w odpowiednie grupy. Na przykład cząstki zwane leptonami obejmują elektrony, miony i taony oraz odpowiadające im neutrina. Z kolei cięższe cząstki klasyfikujemy jako hadrony, wśród których występują bariony (w tym nukleony, czyli protony i neutrony) oraz mezony. Ciężkie cząstki nie są w istocie elementarne, lecz składają się z mniejszych obiektów, zwanych kwarkami. Choć własności kwarków można badać przez odpowiednie oddziaływania z zawierającymi je cząstkami, to jednak wyodrębnienie poszczególnych kwarków jest z zasadniczych powodów niemożliwe. Warto również zwrócić uwagę na fakt, że każdej istniejącej cząstce odpowiada jej antycząstka, która w wypadku cząstek naładowanych elektrycznie będzie miała zawsze przeciwny znak (dla

niektórych cząstek, takich jak np. foton, sama cząstka jest swoją antycząstką). Aby wprowadzić nieco porządku do chaosu przeróżnych rodzajów cząstek, stosuje się matematyczną teorię grup, dzięki której odkryto przedziwne i nie zawsze łatwo wytłumaczalne symetrie między cząstkami (jak np. tworzenie przez pewne grupy barionów i mezonów tzw. oktetów zawierających po osiem cząstek). Można postawić pytanie, czy istnienie takich symetrii (a nawet „supersymetrii", jak w tzw. teorii strun) ma jakieś głębsze filozoficzne czy też ontologiczne uzasadnienie.

Najpoważniejszym wyzwaniem fizyki współczesnej jest pogodzenie dwóch fundamentalnych, lecz zasadniczo niezgodnych teorii: ogólnej teorii względności z mechaniką kwantową. Rezultatem takiej syntezy powinna być nowa teoria zwana grawitacją kwantową. Niestety, jak na razie próby skwantowania oddziaływań grawitacyjnych nie zakończyły się powodzeniem, ze względu na mnożące się trudności matematyczne i pojęciowe. Niektórzy fizycy uważają, że do prac nad stworzeniem takiej teorii powinni także włączyć się filozofowie. Jest to oczywiście zadanie, któremu mogłaby podołać jedynie wąska grupa filozofów, mających również doskonałe przygotowanie fizyczno-matematyczne. W każdym razie należy zauważyć, że wiele prób stworzenia grawitacji kwantowej dotyka fundamentalnych problemów o charakterze filozoficznym, takich jak natura czasu i przestrzeni. Mówi się o tym, że zgodnie z kwantową grawitacją czas i przestrzeń byłyby własnościami emergentnymi, „pojawiającymi" się dopiero na pewnym etapie rozwoju materii. Oczywiście pozostaje kwestią otwartą, czy można nadać wyrażeniom takim jak „pojawiać się" sens inny od czasowego czy czasoprzestrzennego, aby uniknąć oczywistego błędnego koła.

Przerwijmy jednak te rozważania, które można by jeszcze długo ciągnąć. Być może w kolejnej książce podejmę próbę przybliżenia niektórych z naszkicowanych tutaj zagadnień filozoficznie zorientowanym czytelnikom. Na razie mogę polecić monumentalną pracę słynnego fizyka brytyjskiego, która choć opublikowana prawie dwadzieścia lat temu, nie straciła wiele na aktualności. Oto jej tłumaczenie na język polski: R. Penrose, *Droga do rzeczywistości*, Prószyński i S-ka, Warszawa 2020. Jest to praca niezwykle wymagająca dla czytelnika, ale jej uważna lektura, nawet we fragmentach, będzie warta wysiłku.

INDEKS

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

Z

Książka zawiera poglądowe wprowadzenie do fizyki współczesnej z podkreśleniem jej aspektów filozoficznych. Przedstawia rozwój najważniejszych pojęć i zagadnień fizycznych od starożytnych teorii astronomicznych, przez mechanikę newtonowską, termodynamikę, teorię elektromagnetyzmu do obu teorii względności i mechaniki kwantowej. Szczegółowo opisuje powiązane problemy filozoficzne, takie jak zagadnienie determinizmu i przewidywalności, spór o status czasu i przestrzeni, ontologiczny status pól fizycznych, testowanie i akceptacja teorii empirycznych. Książka może służyć studentom filozofii zainteresowanym filozoficznymi aspektami nauk szczegółowych, a także studentom kierunków przyrodniczych, pragnącym uzupełnić swoją specjalistyczną wiedzę o zagadnienia filozoficzne.

Tomasz Bigaj jest profesorem na Wydziale Filozofii Uniwersytetu Warszawskiego. Ukończył studia fizyczne i filozoficzne. Zajmuje się filozoficznymi problemami fizyki, ontologią analityczną oraz logiką filozoficzną. Opublikował ponad sześćdziesiąt artykułów naukowych, głównie w międzynarodowych czasopismach (*Synthese*, *Erkenntnis*, *Foundations of Science, Studies in History and Philosophy of Modern Physics, Foundations of Physics*, *Quantum Reports*). Jest ponadto autorem książek: *Identity and Indiscernibility in Quantum Mechanics*, Palgrave-Macmillan 2022; *Nonlocality and Possible Worlds*, Ontos Verlag/de Gruyter 2006; *Kwanty, liczby, abstrakty*, WN Semper 2002.

ISBN 978-83-971439-0-6