LABORATORIUM MENTIS

2023 · VOL 1 NO 1

Wydawnictwo Academicon

Labolatorium Mentis, 2023

KONTAKT / CONTACT

Redakcja *Laboratorium Mentis* / *Laboratorium Mentis* editorial office
20-810 Lublin, Poland, ul. Heleny Modrzejewskiej 13
e-mail: info@labmentis.eu
www: ojs.academicon.pl/lm

WYDAWCA / PUBLISHER

Wydawnictwo Academicon / Academicon Press
20-810 Lublin, Poland, ul. Heleny Modrzejewskiej 13
e-mail: wydawnictwo@academicon.pl
www: omp.academicon.pl
www: academicon.pl/wydawnictwo

OPRACOWANIE WYDAWNICZE / PUBLISHING WORK

Studio DTP Academicon / Academicon DTP Studio
e-mail: dtp@academicon.pl
www: dtp.academicon.pl

Wersją pierwotną czasopisma jest wersja elektroniczna.
The electronic version of the journal is original version.

Spis treści

Wydawnictwo Academicon
LUBLIN 2023

LABORATORIUM MENTIS | 2023, T. 1, NR. 1

DOI: 10.52097/lm.8146

List od redakcji

Szanowni Państwo!

Mimo ogromnej liczby już istniejących czasopism filozoficznych uważamy, że istnieje zapotrzebowanie na nowy, ściśle określony typ periodyku filozoficznego.

Po pierwsze, w przypadku większości czasopism zarówno napisanie artykułu, który byłby w nich opublikowany, jak i jego lektura są (często niepotrzebnie) czasochłonne. Wymogi, jakie stawia się autorom, obejmują konieczność uwzględnienia bogatej, przede wszystkim najnowszej literatury, wykazania znajomości stanu badań w zakresie poruszanego tematu, co w dość znacznym stopniu oznacza konieczność poświęcenia sporej części publikacji odtwórczemu sprawozdaniu z tego, co napisali inni. Średnia długość artykułu filozoficznego to ponad dwadzieścia stron. Znaczna jego część, obok głównej tezy i argumentacji, zawiera szczegółową prezentację poglądów i argumentacji innych autorów lub rozważania dotyczące wątków pobocznych. Czytelnik, szczególnie czytający w celu przeprowadzenia badań na wąsko określony temat, często znajduje się w sytuacji, gdzie na jedną godzinę lektury materiału, którego szukał, przypada kilka godzin lektury materiału, który nie zawiera poszukiwanych treści. Proponowane przez autora *novum*, najbardziej wartościowa część artykułu to zazwyczaj jedynie ułamek jego objętości.

Po drugie i ważniejsze, żadne istniejące dotychczas czasopismo (przynajmniej wedle naszej wiedzy) nie jest poświęcone przede wszystkim eksperymentom myślowym. Owszem, eksperymenty tego rodzaju

stanowią znaczącą część filozofowania, co pokazują publikacje, niemniej rzadko są one głównym składnikiem tekstu.

Kierując się powyższymi obserwacjami oraz przekonaniem, że dobry tekst filozoficzny powinien wyrażać jak najwięcej treści w jak najmniejszej ilości słów, doszliśmy do wniosku, że warto powołać do życia czasopismo, poświęcone nowatorskim eksperymentem myślowym, którego teksty będą charakteryzowały się efektywnością oraz rozsądną objętością. Stąd nazwa naszego periodyku – *Laboratorium Mentis*.

Chcemy, aby nasze czasopismo stanowiło miejsce, w którym filozofowie mogliby dzielić się nowatorskimi pomysłami, argumentami, najlepiej w formie eksperymentu myślowego, nawet jeśli są one jeszcze w fazie dopracowywania. Chcemy też, aby te eksperymenty stanowiły okazję do dyskusji, stąd będziemy otwarci również na artykuły polemiczne nawiązujące wprost do publikowanych eksperymentów myślowych.

Dodatkową racją przemawiającą na rzecz powstania czasopisma filozoficznego o tak określonym profilu jest to, że być może już w niedalekiej przyszłości teksty filozoficzne, w których prezentuje się i omawia poglądy innych osób, w dużej mierze będą pisane przez sztuczną inteligencję, która zapewne będzie to robiła lepiej i szybciej niż ludzie. Mamy jednak nadzieję, że to, co najbardziej twórcze w ludzkim filozofowaniu, w tym formułowanie nowatorskich eksperymentów myślowych, nadal pozostanie domeną ludzi, a nie maszyn.

Aby zrealizować założone przez nas cele, proponujemy następujący format tekstów, które byłyby przez nas publikowane. Po pierwsze, objętość tekstu powinna być jak najmniejsza, tylko taka, jakiej domaga się skuteczne zrealizowanie celu artykułu. Sugerujemy, aby teksty miały następującą strukturę. Pierwsza część ma stanowić zaprezentowanie problemu oraz celu artykułu. W części tej można, dla jaśniejszego zarysowania problemu, zwięźle przedstawić aktualny stan dyskusji, co oznacza także konieczność wskazania na kilka najistotniejszych dla przedmiotu rozważań publikacji, niemniej oczekujemy, że autorzy ograniczą się tu do niezbędnego minimum. Tekst powinien zawierać

wyodrębnioną część, w której zaprezentowany będzie eksperyment myślowy (na zasadzie wyjątku może to być po prostu nowatorski argument) oraz wniosek, na rzecz którego ten eksperyment ma przemawiać. Zakładamy, że eksperyment myślowy nie zawsze ma funkcję argumentacyjną, że może też zostać użyty w celu zilustrowania lub głębszego zrozumienia jakiejś tezy lub problemu. Oczekujemy także, że autorzy poświęcą kilka stron antycypowanym zarzutom oraz podejmą próbę obrony swojego pomysłu przed takowymi. Całość tekstu nie powinna przekraczać (poza dobrze uzasadnionymi wyjątkami) pół arkusza wydawniczego (20 000 znaków ze spacjami).

Kierując się przekonaniem, że filozofia jest dyscypliną bez granic, a dyskusja powinna obejmować wszystkich potencjalnie zainteresowanych danym tematem, będziemy otwarci na publikacje w języku angielskim. Niemniej uważamy, że kultywowanie rodzimego języka filozoficznego jest niezwykle ważne dla rozwoju kultury, stąd równie otwarci będziemy na teksty polskojęzyczne. Idealnie będziemy dążyli do dwujęzycznego publikowania artykułów.

Labolartorium Mentis będzie rocznikiem publikowanym online w otwartym dostępie. To, ile zeszytów w ciągu roku będzie publikowanych, będzie zależało od liczby i jakości nadsyłanych tekstów.

Mamy nadzieję, że nasze czasopismo będzie stanowiło godny uwagi, twórczy wkład w rozwój dyscypliny filozoficznej. Zapraszamy do nadsyłania tekstów.

Redakcja

LABORATORIUM MENTIS | 2023, VOL. 1, NO. 1

DOI: 10.52097/lm.8146

Editorial letter

Dear Colleagues!

Despite a huge number of already existing philosophical journals, we believe there is a demand for a new, narrowly defined type of philosophical periodical.

Firstly, in the case of most existing journals, both writing an article that would be published in them and reading it are (often unnecessarily) time-consuming. The requirements imposed on authors include the necessity of considering a rich, primarily the most recent literature, demonstrating knowledge of the state of research on the discussed topic, which to a considerable extent means the necessity of devoting a significant part of the publication to a summary of what others have written. The average length of a philosophical article is over twenty pages. A substantial part of it, alongside the main thesis and arguments, contains a detailed presentation of the views and arguments of other authors or considerations concerning side issues. The reader, especially one reading to research a narrowly defined topic, often needs to spend several hours reading material they are not looking for to have one hour of reading the sought content. The novelty proposed by the author, the most valuable part of the article, is usually only a fraction of its volume.

Secondly, no existing journal (at least to our knowledge) is primarily dedicated to thought experiments. Of course, experiments of this kind constitute a significant part of philosophy, which is reflected in publications, but rarely are they the main component of a concise text.

Guided by the above observations and the conviction that a good philosophical text should express as much content as possible in as

few words as possible, we have concluded that it is worth creating a journal primarily dedicated to innovative thought experiments whose texts will be characterized by efficiency and reasonable volume. Hence the name of our journal, *Laboratorium Mentis*.

We want our journal to be a place where philosophers can share innovative ideas and arguments, preferably in the form of a thought experiment, even if they are still in the process of refinement. We also want these experiments to be an opportunity for discussion, so we will be open to polemical articles directly related to the published thought experiments.

An additional reason in favour of creating a philosophical journal of such a profile is that perhaps in the near future, philosophical texts in which the views of other people are presented and discussed will largely be written by artificial intelligence, which will probably do it better and faster than humans. However, we hope that what is most creative in human philosophy, including formulating innovative thought experiments, will still remain the domain of humans rather than machines.

In order to achieve the goals set by us, we propose the following format for texts that would be published by *Labolatorium Mentis*. Firstly, the volume of the text should be as small as possible, only as much as is required to effectively achieve the article's goal. Although we will be flexible regarding the required size of published texts, we suggest they do not exceed twenty thousand characters, including spaces. We recommend that the articles have the following structure. The first part should be a presentation of the problem and the article's purpose. In this part, the authors can briefly present the current state of the discussion, which also means the need to indicate a few of the most important publications. Still, we expect the authors to limit themselves here to the necessary minimum. The text should contain a separate part in which a thought experiment will be presented (as an exception, this can be an innovative argument) and a conclusion for which this experiment should speak. We assume that a thought experiment does

not always have an argumentative function but can also be used to illustrate or gain a deeper understanding of a given issue or claim. We also expect the authors to spend a few pages anticipating objections and attempting to defend their ideas against them.

Sharing the belief that philosophy is a discipline without borders and that the discussion should involve all those potentially interested in a given topic, we will be open to publications in the English language. Nevertheless, we also believe that the development of the native language of Polish philosophy is extremely important for developing culture in Poland. For this reason, we will also be open to Polish-language texts. Ideally, we will strive for bilingual publications of articles.

Laboratorium Mentis will be an annual periodical published online in open access. The number of volumes published in a year will depend on the quantity and quality of the texts submitted.

We hope that our journal will be a worthwhile, creative contribution to the development of the philosophical discipline. We invite you to submit texts.

Editors

LABORATORIUM MENTIS | 2023, T. 1, NR. 1

DOI: 10.52097/lm.8148

PIOTR BIŁGORAJSKI
Katolicki Uniwersytet Lubelski Jana Pawła II (Polska)
0000-0001-5139-3455
piotr.bilgorajski@kul.pl

Czy eksperymenty myślowe są poważną metodą uprawiania filozofii?[1]

Received: 24.11.2023 / Revised: 04.12.2023 / Accepted: 12.12.2023 / Available: 20.12.2023

Abstrakt:
W artykule bronię praktyki stosowania eksperymentów myślowych przed zarzutem, że nie jest to poważny sposób filozoficznej argumentacji. Osią krytyki wymierzonej w eksperymenty myślowe jest założenie, że wytwory wyobraźni, ze względu na ich brak osadzenia w rzeczywistości, są z zasady niewiarygodne. Przyjmując istnienie analogii między eksperymentami myślowymi a eksperymentami rzeczywistymi, wskazuję na istnienie kryteriów wyznaczających ramy udanego eksperymentu myślowego.

Słowa kluczowe:
eksperymenty myślowe, metafilozofia, wyobraźnia, mentalne symulacje

Jak cytować:
Biłgorajski, P. (2023). Czy eksperymenty myślowe są poważną metodą uprawiania filozofii? [polski oryginał]. *Laboratorium Mentis*, *1*(1), 10–21. https://doi.org/10.52097/lm.8148

[1] Praca została sfinansowana ze środków Narodowego Centrum Nauki w ramach projektu „Eksperymenty myślowe w filozofii – geneza, struktura, funkcje", no. UMO-2017/25/N/HS1/03019.

Zacznijmy od historyjki. Wyobraźmy sobie profesora filozofii, który dużą część swojej kariery naukowej zbudował na przedstawieniu eksperymentów myślowych, które na lata nadały ton wielu dyskusjom filozoficznym. Jednak pewnego dnia grupa badaczy publicznie ogłasza, że – mimo wielu prób – nie udało się zreprodukować wyników eksperymentów myślowych zaproponowanych przez tego filozofa. Na tej podstawie wyciągają wniosek, że te eksperymenty myślowe zostały prawdopodobnie sfałszowane. Znany filozof, w obliczu przytłaczających dowodów, przyznaje się do oszustwa i w atmosferze skandalu odchodzi z uniwersytetu i już nigdy więcej nie zajmuje się filozofią.

Czy taka historyjka jest możliwa? Można podejrzewać, że wiele osób uzna to za żart. W powszechnym mniemaniu jest przecież czymś niewiarygodnym, aby sfałszować dane otrzymane w rezultacie przeprowadzenia eksperymentu myślowego, czyli czynności wykonywanej wyłącznie w wyobraźni. Ale chyba każdy się także zgodzi, że gdyby podobna historyjka dotyczyła jakiegoś znanego fizyka lub biologa, którego oskarżono by o sfałszowanie wyników jego eksperymentów, ale tym razem rzeczywistych, to nikt nie uznałby tego za dobry żart, a wręcz przeciwnie – bylibyśmy świadkami powszechnego i całkowicie uzasadnionego oburzenia.

Chciałbym poważnie potraktować pytanie, dlaczego pierwsza historyjka, dotycząca fałszowania eksperymentów myślowych, wydaje się zabawna, natomiast przypadki fałszowania rzeczywistych eksperymentów – już mniej. Co jest w niej śmiesznego? Wyjaśnieniem może być tutaj popularna teoria humoru, która głosi, że humor wynika z dostrzeżonej niespójności. W tym kontekście historyjka wydaje się śmieszna, ponieważ koncepcja sfałszowanego eksperymentu myślowego zawiera wewnętrzną sprzeczność. Niespójność polega na tym, że nie można sfałszować wytworów wyobraźni, stąd eksperymenty myślowe, jako odbywające się wyłącznie wyobraźni, nie mogą być „oszukane". W eksperymentach myślowych wszystko jest dozwolone.

Taki wniosek łatwo oddaje pole krytykom eksperymentów myślowych, takim jak na przykład Kathleen Wilkes (2003), która

porównywała eksperymenty myślowe do opowieści *fantasy*. Jej zdaniem miejsce dla takich historyjek powinno być w beletrystyce, a nie w pracach naukowych. Z drugiej strony praktyka uprawiania filozofii zdaje się świadczyć o tym, że taki pogląd jest rzadki. Eksperymenty myślowe są przecież powszechnie stosowane nie tylko w oryginalnych pracach filozoficznych, ale także jako narzędzie popularyzacji filozofii. Czyżby filozofowie stosowali narzędzie, którego rezultaty uważają za niewiarygodne?

W artykule spróbuję obronić powagę eksperymentów myślowych. Zacznę od przedstawienia popularnych typologii eksperymentów myślowych, a następnie zaproponuję definicję eksperymentu myślowego zbudowaną na zasadzie analogii do eksperymentów rzeczywistych. Przyjmę założenie, że sfałszowany eksperyment myślowy jest szczególnym przypadkiem nieudanego eksperymentu, i wskażę kryteria nieudanego eksperymentu myślowego. Jeśli bowiem eksperyment myślowy może się nie udać, to istnieją kryteria narzucające ograniczenia na udany eksperyment myślowy.

Po co filozofom eksperymenty myślowe?

Jedna z pierwszych typologii eksperymentów myślowych pochodzi od Karla Poppera, który wyróżnił eksperymenty myślowe tworzone z intencją heurystyczną, krytyczną (destruktywną) i apologetyczną (konstruktywną) (Popper, 2002).

Heurystyczne eksperymenty myślowe w atrakcyjny dla odbiorcy sposób prezentują pewną teorię, przez co ułatwiają jej szerszą recepcję. Tego typu eksperymenty mogą służyć za ilustrację uznanych już teorii lub w sposób uproszczony prezentować wyniki danej teorii w celach popularyzatorskich. Eksperymenty krytyczne tworzy się przeciwko jakiejś teorii lub w celu podważenia założeń i wniosków innych eksperymentów myślowych. Eksperymenty myślowe w tej funkcji najczęściej prezentowane są jako kontrprzykłady dla jakiegoś ogólnego twierdzenia.

Natomiast eksperymenty apologetyczne służą dostarczaniu przykładów potwierdzających określoną teorię (Popper, 2002, s. 243).

Typologię Poppera warto uzupełnić o propozycję Tamary Gendler, która dzieli eksperymenty myślowe na trzy kategorie: (i) faktualne, (ii) konceptualne i (iii) ewaluacyjne. Faktualne eksperymenty myślowe dotyczą eksperymentów myślowych obecnych w naukach empirycznych. Na przykład w eksperymencie myślowym „kamienie Galileusza" zadajemy pytanie, co się wydarzy, jeśli zrzucimy połączone ze sobą dwa kamienie o różnej masie. Konceptualne eksperymenty myślowe można spotkać w metafizyce i epistemologii, gdzie służą sprawdzeniu, czy dane pojęcie stosuje się do opisanej sytuacji. Na przykład, czy pojęcie wiedzy stosuje się do tzw. problemu Gettiera. Ewaluacyjne eksperymenty myślowe występują w etyce i estetyce. Tutaj odbiorca konfrontowany jest na przykład z dylematem wagonika i jego zadaniem jest dokonać oceny moralnej przedstawionej sytuacji.

Powyższe typologie wskazują na funkcje eksperymentów myślowych. W jaki jednak sposób eksperymenty myślowe realizują te funkcje? Chris Daly zauważa, że w literaturze można spotkać eksperymenty myślowe, które funkcjonują jako: (1) „wyzwalacze", (2) wglądy w świat platońskich idei, (3) argumenty, (4) odmiana realnych eksperymentów oraz (5) mentalne modele (Daly, 2010).

Z koncepcją eksperymentów myślowych jako wyzwalaczy związany jest Thomas Kuhn. Według niego eksperymenty myślowe służą ujmowaniu dostępnych danych w nowe schematy pojęciowe i dzięki temu ułatwiają (wyzwalają) wykrywanie nagromadzonych sprzeczności i anomalii (Kuhn, 1985). Przedstawicielem platońskiego podejścia do eksperymentów myślowych jest James Brown, który uważa, że eksperymenty myślowe – poprzez jakiś rodzaj intuicji – umożliwiają dostęp do platońskiego świata koniecznych i niezmiennych prawd (Brown, 1991).

John Norton (2004) broni stanowiska, że eksperymenty myślowe w zasadzie nie różnią się od argumentów, dlatego też ich zasadniczą funkcją jest przekonywanie. Podobnie uważa Daniel Dennett (2013), nazywając eksperymenty myślowe „pompami intuicji". W tym ujęciu

eksperymenty myślowe służą jedynie perswazji, ale jeśli podobnie jest z innymi argumentami, to dobry eksperyment myślowy jest dobrym argumentem, a zły eksperyment myślowy jest argumentem złym. Z drugiej strony Roy Sorensen (1997), dowartościowując eksperymenty myślowe, twierdzi, że właściwie nie powinno się dodawać przydawki „myślowy" do eksperymentów, ponieważ te eksperymenty, które w nauce powszechnie uważa się za „myślowe", są na tyle podobne do „rzeczywistych" eksperymentów, że zanika potrzeba ich odróżniania.

Timothy Williamson (2007) uważa, że tym, co szczególnie wyróżnia filozoficzne eksperymenty myślowe, jest ich modalny charakter. Ten aspekt podkreślany jest w koncepcji eksperymentów myślowych traktowanych jako mentalne modele (Nersessian, 2018). W tym ujęciu eksperymenty myślowe prezentują sytuacje, które odpowiadają lub skłaniają do odpowiedzi na pytanie: „Co by było, gdyby?". Wyobrażona sytuacja stanowi model, czyli reprezentację pewnego możliwego stanu rzeczy, zaś eksperyment myślowy polega na przeprowadzeniu symulacji zachowania danego stanu rzeczy. Zdaniem Nancy Nersessian proces symulacji przebiega w trzech krokach. Pierwszy etap to skonstruowanie mentalnego modelu, który jest reprezentacją wybranego aspektu rzeczywistości. Następnie dokonuje się określonych manipulacji na przedstawionym modelu. Na końcu wykorzystuje się otrzymane w wyniku manipulacji rezultaty do wnioskowania na temat modelowanego aspektu rzeczywistości.

Koncepcja symulacji zakłada, że proces wyobraźni odbywa się *offline* i chociaż jego celem jest odwzorowanie rzeczywistości, to nie jest on, w odróżnieniu od percepcji, „kontrolowany" przez rzeczywistość. Można więc postawić zarzut, że w takim ujęciu brakuje kryterium odróżniania wyobrażeń wartościowych od produktów czystej fantazji. W odpowiedzi zwolennicy koncepcji symulacji kładą nacisk na to, że mechanizm wyobraźni w przypadku odbierania fikcji nie różni się od tego, jak wykorzystywana jest ona do zrozumienia otoczenia w sytuacjach codziennych. W wyobraźni często tworzymy fikcyjne scenariusze i testujemy różne rozwiązania bez podejmowania wysiłku

i ryzyka ich realizacji w rzeczywistości. Takie kontrolowane korzystanie z wyobraźni jest na tyle powszechne, że niektórzy badacze wskazują, że zdolność do tworzenia symulacji posiada uzasadnienie ewolucyjne (Williamson, 2016).

Tym, co łączy te koncepcje, jest założenie, że istnieją pewne strukturalne podobieństwa między eksperymentem myślowym a eksperymentem rzeczywistym (naukowym, realnym). Eksperyment rzeczywisty to procedura, która polega na wpływaniu na pewien stan rzeczy, aby zaobserwować, co się wydarzy w celu potwierdzenia lub obalenia jakiejś naukowej hipotezy. Eksperyment myślowy, podobnie jak rzeczywisty, wykonywany jest w określonym celu poznawczym oraz polega na intencjonalnym zmienianiu jakiegoś stanu rzeczy. Chociaż jednak w eksperymencie rzeczywistym materiał poddawany zmianie jest empiryczny i faktualny (aktualnie istniejący), to w eksperymencie myślowym – wyobrażeniowy i kontrfaktyczny. Innymi słowy, odpowiednikiem wywołania pewnego stanu rzeczy w eksperymencie rzeczywistym jest w eksperymencie myślowym wyobrażenie sobie zajścia pewnego stanu rzeczy, natomiast odpowiednikiem obserwacji rezultatu eksperymentu jest odpowiednie rozumowanie, odbywające się niejako w „laboratorium umysłu".

Czy eksperymenty myślowe zasługują na miano eksperymentów?

Roy Sorensen uważa, że tym, co łączy eksperymenty myślowe i rzeczywiste, jest „majsterkowanie". Oznacza to, że projektowanie zarówno eksperymentu rzeczywistego, jak i myślowego polega na tworzeniu pewnych sztucznych warunków, w których uwzględnia się tylko czynniki istotne dla przebiegu procedury (tzw. warunki *ceteris paribus*). I jak zazwyczaj celem eksperymentu jest ujawnienie jakiejś anomalii, czyli zjawiska, które nie poddaje się wyjaśnieniu przez testowaną teorię, tak eksperymenty myślowe najczęściej dostarczają kontrprzykładu dla jakiejś filozoficznej koncepcji.

Sorensen zauważa, że eksperymenty myślowe i rzeczywiste funkcjonują w podobny sposób: jako punkty odniesienia w toczonych dyskusjach. Znane eksperymenty myślowe, takie jak np. problem Gettiera lub dylematy wagonika, stają się wzorcowym sposobem prowadzenia analiz (w tym przypadku – analiz dotyczących pojęcia wiedzy lub zakresu stosowalności określonych teorii etycznych). Eksperymenty myślowe podatne są też na krytykę i korektę. Dynamika sporów w filozofii pokazuje, że najczęstszą odpowiedzią na zaproponowany eksperyment myślowy jest jakiś kontreksperyment myślowy.

Sorensen wskazuje także na różnice między eksperymentami rzeczywistymi a myślowymi, ale jego zdaniem nie są one na tyle istotne, by unieważnić podobieństwa. Omawia pewne oczywiste cechy eksperymentów rzeczywistych – fakt, że eksperymenty rzeczywiste są zwykle wykonywane przez zespoły badawcze, w których można odróżnić osoby odpowiedzialne za projektowanie eksperymentu od osób, które te eksperymenty wykonują. W przypadku eksperymentów filozoficznych projekt i wykonanie nie wymagają takiego podziału pracy, co może być uznane za ich zaletę. Wydaje się także, że wyniki w eksperymentach myślowych nie są uzyskiwane przez przypadek, co zdarzało się przy okazji niektórych eksperymentów w fizyce. Trudno oczekiwać, by wynik eksperymentu myślowego mógł być dla osoby go przeprowadzającej zaskakujący. Z drugiej jednak strony eksperymenty myślowe byłyby znaczniej mniej podatne na takie zdarzenia losowe, jak np. awaria sprzętu. Z oczywistych względów myślowy eksperymentator ma większą kontrolę nad przebiegiem swojego rozumowania.

Eksperymenty myślowe – w odróżnieniu do eksperymentów rzeczywistych – nie wymagają także drogiej i skomplikowanej aparatury badawczej. W tym kontekście metaforycznym odpowiednikiem fizycznego laboratorium w filozofii może być umysł filozofa, zaś jakość takiego „laboratorium" zależałaby od odpowiedniego poziomu edukacji i inteligencji właściwych osobie przeprowadzającej eksperyment filozoficzny.

Kiedy eksperyment myślowy może się nie udać?

Eksperyment myślowy rozpoczyna się od przedstawienia pewnej możliwej, fikcyjnej sytuacji, która przebiega w sposób, który – jeśli eksperyment zakończy się sukcesem – skłoni odbiorcę do wyciągnięcia określonego wniosku. Eksperyment myślowy jest więc rozumowaniem, w którym można wyróżnić trzy etapy: (1) opisana jest wyobrażeniowa (możliwa) sytuacja („Wyobraź sobie kobietę o imieniu Maria, która nie zna kolorów, ale zna całą fizykę barw…"), (2) przedstawiona sytuacja posiada narracyjny charakter („Maria widzi czerwoną różę i dowiaduje się czegoś nowego o świecie…"), (3) rezultat przedstawionej narracji potwierdza lub obala tezę filozoficzną („A więc… Fizykalizm jest fałszywy!"). Takie ujęcie struktury eksperymentu myślowego pozwala w klarowny sposób wskazać, kiedy eksperyment myślowy nie będzie udany:

1. Niewyobrażalność: eksperyment myślowy można podważyć, wskazując na to, że przedstawiona w nim sytuacja nie jest możliwa do wyobrażenia (ponieważ jest niespójna lub opisana jest zbyt ogólnie).

2. Niekonkluzywność: nawet jeśli sytuacja jest wyobrażalna, to nie ma dobrych racji, aby przyjąć określony rezultat tej sytuacji.

3. Brak odniesienia do naszego świata: nawet jeśli sytuacja jest wyobrażalna i istnieją dobre racje, aby przyjąć określony rezultat tej sytuacji, to nie stanowi ona podstawy dla twierdzenia dotyczącego naszego świata (Gendler, 2000, s. 22).

W pierwszym sensie eksperyment myślowy zawodzi, jeśli przedstawiony w nim stan rzeczy jest niewyobrażalny. Zarzut niewyobrażalności jest często spotykaną krytyką niektórych, szczególnie ekstrawaganckich eksperymentów myślowych, takich jak np. dotyczących możliwości istnienia filozoficznych zombie (Chalmers, 2010). Wskazanie, że zombie, czyli istota pod fizycznym względem identyczna ze mną, ale pozbawiona pierwszoosobowego punktu widzenia, jest niepojmowalne, to popularny sposób na osłabienie tego argumentu. Jeśli więc filozoficzne

zombie jest niewyobrażalne, to taki eksperyment myślowy można byłoby uznać za nieudany.

W drugim sensie nieudany eksperyment myślowy polegałby na tym, że chociaż eksperymentator jest w stanie wyobrazić sobie przedstawioną sytuację, to niewłaściwie wyobraża sobie jej przebieg. W znanym eksperymencie myślowym Franka Jacksona Maria, gdy po raz pierwszy widzi czerwoną różę, dowiaduje się czegoś nowego o świecie, skąd wyciąga się wniosek, że fizykalizm jest fałszywy (Jackson, 1986). Krytyk tego eksperymentu myślowego mógłby jednak upierać się, że przebieg tego eksperymentu powinien wyglądać inaczej. Na przykład w taki sposób, że Maria na widok czerwonej róży wykrzykuje, że róża wygląda dokładnie tak, jak Maria sądziła – wszak Maria posiada całą wiedzę na temat fizyki barw i kolor róży nie powinien być dla niej niczym nowym ani zaskakującym. Analogiczna sytuacja może wystąpić w eksperymencie rzeczywistym, jeśli na przykład eksperyment przebiega bez zakłóceń, ale generuje błędne dane (z powodu uszkodzenia sprzętu lub błędu na etapie projektowania eksperymentu).

Nieudany eksperyment myślowy w trzecim sensie to eksperyment, który prowadzi do poprawnego wniosku, jednak nie dostarcza odpowiedzi na pytanie, które sprowokowało jego przeprowadzenie. W eksperymencie rzeczywistym byłaby to na przykład sytuacja, w której eksperyment generuje poprawny wynik, jednak kończy się fiaskiem, ponieważ nie dostarcza tych danych, które interesowały badaczy. Przykładem nieudanego eksperymentu myślowego być może „Młyn" Gottfrieda Leibniza. Formułując argument przeciwko stanowisku, że ludzki umysł ma charakter mechaniczny, zapraszał on czytelnika, aby ten wyobraził sobie wnętrze mechanicznego umysłu, w którym – niczym we wnętrzu młyna – w żadnym miejscu nie bylibyśmy w stanie zaobserwować zjawisk mentalnych, a jedynie „części, które popychają się wzajemnie, nigdy jednak nic, co tłumaczyłoby postrzeżenie" (Leibniz, wyd. 1969, s. 300). Możemy wyobrazić sobie młyn i zgodzić się z Leibnizem, że nie zobaczymy w nim niczego, co tłumaczyłoby

postrzeżenie. Jeśli natomiast zakwestionujemy analogię młyn–umysł, to odrzucimy też wniosek Leibniza, że mechanicyzm jest fałszywy.

Konkluzja

Artykuł zacząłem od historyjki i pytania, co jest w niej zabawnego. Jeśli uznamy eksperymenty myślowe za poważną metodę uprawiania filozofii, to odpowiedź brzmi: nic. Humor bierze się z przyjęcia pewnej szczególnej koncepcji eksperymentów myślowych, wedle której eksperymenty myślowe to procedura, w której wszystko jest dozwolone. Wydaje się jednak, że nie jest to koncepcja trafna, bowiem na każdym etapie eksperymentu myślowego możemy zapytać, czy został on przeprowadzony poprawnie.

Sorensen porównywał eksperymenty myślowe do kompasu (Sorensen, 1992, s. 288). Kompas jest prostym, chociaż użytecznym narzędziem wskazywania kierunku. Nie jest to jednak urządzenie niezawodne – wskazania kompasu są na przykład niewiarygodne w okolicach bieguna północnego. Niewiele osób wie, jak naprawdę działa kompas, chociaż nie stanowi to przeszkody w skutecznym posługiwaniu się tym urządzeniem. Analogicznie, filozoficzne eksperymenty myślowe, korzystając z bogactwa wyobraźni, wskazują na możliwe stany rzeczy. Jednak łatwość tworzenia eksperymentów myślowych – wszak opowiadanie historyjek typu „co by było, gdyby…" nie wymaga specjalnych technicznych umiejętności – nie przekłada się na wiarygodność uzyskanych w ten sposób rezultatów. Dlatego świadomość ograniczeń wytworów wyobraźni pozwala na korzystanie z eksperymentów myślowych z większą ostrożnością.

Eksperymenty myślowe mogą pod pewnymi względami przypominać kompas, pod innymi – lupę. Celem bowiem fantastycznych historii podawanych przez filozofów jest drobiazgowe testowanie teorii filozoficznych w celu sprawdzenia, czy moc wyjaśniająca tych teorii obejmuje wszystkie możliwe sytuacje. Jeśli ktoś twierdzi, że z konieczności wiedza to prawdziwe i uzasadnione przekonanie, to ten pogląd

można podważyć, podając przykład sytuacji, kiedy ktoś dysponuje wiedzą, chociaż nie spełnia wszystkich warunków podanych w definicji wiedzy. Jeśli ktoś twierdzi, że fizykalizm jest prawdziwy, więc wszystko, cokolwiek istnieje, można opisać w kategoriach fizycznych, to ten pogląd można zakwestionować, podając przykład przedmiotu, który nie daje się opisać w kategoriach fizycznych. Podanie kontrprzykładu w filozofii polega na przedstawieniu pewnej wyobrażonej, możliwej sytuacji, czyli zaproponowaniu eksperymentu myślowego.

Bibliografia

Brown, J. (1991). *The Laboratory of Mind. Thought Experiments in the Natural Sciences*. London-New York: Routledge.

Chalmers, D. (2010). *Świadomy umysł*, przeł. M. Miłkowski. Warszawa: PWN.

Daly, Ch. (2010). *Introduction to Philosophical Methods*. Peterborough: Broadview Press.

Dennett, D. (2013). *Intuition Pumps and Other Tools for Thinking*. New York: W.W. Norton & Company.

Gendler, T. (2000). *Thought Experiment. On the Power and Limits of Imaginary Cases*. New York: Garland Publishing.

Jackson, F. (1986). What Mary Didn't Know. *Journal of Philosophy*, *83*(5), 291–295.

Kuhn, Th. (1985). *Dwa bieguny*, tłum. S. Amsterdamski. Warszawa: PIW.

Leibniz, G. (1969). *Zasady filozofii, czyli monadologia*, tłum. S. Cichowicz. Warszawa: PWN. Oryginalne dzieło wydane w 1720 r.

Nersessian, N. (2018). Cognitive science, mental modeling, and thought experiments. In M. Stuart, Y. Fehige, J. Brown (Eds.), *The Routledge Companion to Thought Experiments* (pp. 309–326). London: Routledge.

Norton, J. (2004). On Thought Experiments: Is There More to the Argument?. *Philosophy of Science*, *71*(5), 1139–1151. https://doi.org/10.1086/425238

Sorensen, R. (1992). *Thought Experiments*. Oxford: Oxford University Press.

Wilkes, K. (2003). *Real People: Personal Identity without Thought Experiments*. Oxford: Clarendon Press.

Williamson, T. (2007). *The Philosophy of Philosophy*. Malden: Blackwell.

Williamson, T. (2016). Knowing by Imagining. In A. Kind, P. Kung (Eds.), *Knowledge Through Imagination* (pp. 113–123). Oxford: Oxford University Press.

LABORATORIUM MENTIS | 2023, VOL. 1, NO. 1

DOI: 10.52097/lm.8149

PIOTR BIŁGORAJSKI
John Paul II Catholic University of Lublin (Poland)
0000-0001-5139-3455
piotr.bilgorajski@kul.pl

Are thought experiments a reliable method of doing philosophy?[1]

Received: 24.11.2023 / Revised: 04.12.2023 / Accepted: 12.12.2023 / Available: 20.12.2023

Abstract:
In the paper I defend the practice of using thought experiments against the claim that it is not a serious way of philosophical argumentation. At the heart of the criticism leveled against thought experiments is the assumption that the products of imagination, due to their lack of grounding in reality, are fundamentally unreliable. Assuming the existence of an analogy between thought experiments and real experiments, I point out that there are criteria that define the framework of a good thought experiment.

Key words:
thought experiments, metaphilosophy, imagination, mental simulations

How to cite:
Biłgorajski, P. (2023). Are thought experiments a reliable method of doing philosophy? [English translation]. *Laboratorium Mentis*, *1*(1), 22–32. https://doi.org/10.52097/lm.8149

[1] The work was supported by the National Science Centre, Poland, under research project "Thought Experiments in Philosophy: Origin, Structure, Functions," no. UMO-2017/25/N/HS1/03019.

Let us start with a story. Imagine a professor of philosophy who built a significant portion of her academic career on presenting thought experiments that set the tone for many philosophical discussions over the years. However, one day, a group of researchers publicly announced that despite many attempts, they failed to reproduce the results of the thought experiments proposed by this philosopher. Based on this, they conclude that these thought experiments were likely fabricated. The renowned philosopher, faced with overwhelming evidence, admits to the fraud and, amid scandal, leaves the university, never to engage in philosophy again.

Is such a story possible? One might suspect that many would consider it a joke. It is commonly deemed implausible to fabricate data resulting from the execution of a thought experiment, an activity conducted solely in the realm of imagination. Yet, I believe everyone would also agree that if a similar story involved a well-known physicist or biologist accused of fabricating the results of their experiments, but this time real ones, no one would consider it a good joke. On the contrary, we would witness widespread and entirely justified outrage.

I would like to seriously address the question of why the first story, concerning the fabrication of thought experiments, seems amusing, while cases of fabricating real experiments are less so. What is funny about it? An explanation might lie in the popular theory of humor, which suggests that humor arises from perceived inconsistency. In this context, the story seems funny because the concept of a fabricated thought experiment contains an internal contradiction. The inconsistency lies in the fact that products of imagination cannot be forged; hence, thought experiments, occurring solely within the imagination, cannot be "cheated". In thought experiments, anything is allowed.

Such a conclusion aligns with the views of critics of thought experiments, such as Kathleen Wilkes (2003), who compared thought experiments to fantasy stories. In her view, such narratives belong to fiction rather than scientific work. On the other hand, the practice of engaging in philosophy seems to suggest otherwise. Thought

experiments are widely used not only in original philosophical works but also as a tool for popularizing philosophy. Are philosophers using a tool which produces outcomes they consider unreliable?

I will attempt to defend the reliability of thought experiments. I will start by presenting popular typologies of thought experiments and then propose a definition of a thought experiment built on the principle of analogy to real experiments. I will argue that a fabricated thought experiment is a specific case of an unsuccessful experiment and will identify the criteria for an unsuccessful thought experiment. If indeed a thought experiment can fail, there must be criteria which determine when a thought experiment is successful.

Why do philosophers use thought experiments?

One of the earliest typologies of thought experiments comes from Karl Popper, who distinguished thought experiments created with heuristic, critical (destructive), and apologetic (constructive) intentions (Popper, 2002). Heuristic thought experiments present a certain theory in an appealing way, making it easier for the audience to grasp. Such thought experiments can serve as illustrations of established theories or simplify a presentation of a theory's results for popularization purposes. Critical experiments are devised either against a particular theory or to challenge the assumptions and conclusions of other thought experiments. Thought experiments in this function are often presented as counterexamples to some general assertion. Apologetic experiments provide examples that confirm a given theory (Popper, 2002, p. 243).

Popper's typology can be complemented by Tamara Gendler's proposal, which categorizes thought experiments into three categories: (i) factual, (ii) conceptual, and (iii) evaluative. Factual thought experiments are those present in empirical sciences. For example, in Galileo's thought experiment, one might ask what would happen if two stones of different masses were dropped together. Conceptual

thought experiments can be found in metaphysics and epistemology, serving to verify whether a given concept applies to a described state of affairs. For instance, whether the concept of knowledge applies to the so-called Gettier problem. Evaluative thought experiments appear in ethics and aesthetics. Here, the audience is confronted, for example, with the trolley dilemma, and their task is to make a moral judgment on the presented situation.

The above typologies indicate the functions of thought experiments. But how do thought experiments carry out these functions? Chris Daly suggests that thought experiments can function as (1) "triggers," (2) insights into the world of Platonic ideas, (3) arguments, (4) variations of real experiments, and (5) mental models (Daly, 2010).

Thomas Kuhn is associated with the concept of thought experiments as triggers. According to Kuhn, thought experiments help to fit available data into new conceptual schemes, facilitating the detection of accumulated contradictions and anomalies (Kuhn, 1977). James Brown represents the Platonic approach to thought experiments, suggesting that thought experiments, through some form of intuition, provide access to a Platonic realm of necessary truths (Brown, 1991).

John Norton (2004) argues that thought experiments do not fundamentally differ from arguments; their main function is persuasion. Similarly, Daniel Dennett (2013) refers to thought experiments as "intuition pumps". In this context, thought experiments serve solely as tools of persuasion. However, if this holds for other arguments, a good thought experiment would be a good argument, and a bad thought experiment would be a bad argument. On the other hand, Roy Sorensen (1997) claims that there should not be a need to add the qualifier "thought" to experiments because the experiments commonly regarded as "thought experiments" are so similar to "real" experiments that the distinction becomes negligible.

Timothy Williamson (2007) believes that an essential feature of philosophical thought experiments is their modal character. This aspect is emphasized in the concept of thought experiments treated

as mental models (Nersessian, 2018). In this approach, thought experiments present situations that correspond to or prompt responses to the question: "What if?". The imagined situation serves as a model, a representation of a possible state of affairs, and the thought experiment involves simulating the behavior of that state of affairs. According to Nancy Nersessian, the process of simulation occurs in three steps. The first stage involves constructing a mental model representing a selected aspect of reality. Then, specific manipulations are performed on the presented model. Finally, the results obtained from the manipulations are used to infer about the modeled aspect of reality.

The simulation theory assumes that imagination is similar to perception in a way, but although its purpose is to represent reality, it is not, unlike perception, "controlled" by reality. Thus, it can be argued that such a view lacks a criterion for distinguishing valuable imaginings from the products of pure fantasy. In response, proponents of the simulation theory emphasize that the mechanism of imagination when perceiving fiction is no different from how it is used in everyday situations. In imagination, we can create different scenarios and test various solutions without taking the effort and risk of implementing them in reality. For example, before I do something, I can imagine the possible consequences of an action and decide based on that. Such a controlled use of imagination is so common that some researchers indicate that the ability to create mental simulations has evolutionary justification (Williamson, 2016).

These concepts have in common the assumption that there are certain structural similarities between a thought experiment and a real (scientific, empirical) experiment. A real experiment is a procedure that involves influencing a certain state of affairs to observe what will happen with the aim of confirming or refuting a scientific hypothesis. Similarly, a thought experiment, like a real one, is conducted for cognitive purposes and involves intentionally changing a state of affairs. However, in a real experiment, the material undergoing change is empirical and factual (currently existing), while in a thought experiment, it

is imaginative and counterfactual. In other words, in a real experiment, bringing about a certain state of affairs is equivalent to imagining the occurrence of that state of affairs in a thought experiment, and the counterpart of observing the result of an experiment is appropriate reasoning taking place in a "laboratory of the mind".

Do thought experiments deserve to be called experiments?

Roy Sorensen believes that what thought experiments and real ones have in common is "tinkering." This means that designing both real and thought experiments involves creating specific conditions, considering only the essential factors for the procedure (so-called ceteris paribus conditions). And just as the usual goal of an experiment is to reveal some anomaly, that is, a phenomenon that does not submit to explanation by the tested theory, thought experiments most often provide counterexamples to certain philosophical concepts.

Sorensen notes that thought and real experiments function similarly as reference points in ongoing discussions. Well-known thought experiments, such as the Gettier problem or trolley dilemmas, become the standard method for conducting analyses (in these cases, analyses concerning the concept of knowledge or the scope of applicability of certain ethical theories). Thought experiments are also subject to criticism and correction. The dynamics of disputes in philosophy show that the most common response to a proposed thought experiment is some counter-thought experiment.

Sorensen also points out differences between real and thought experiments, but in his opinion, they are not significant enough to nullify the similarities. Sorensen discusses some obvious characteristics of real experiments—like the fact they are usually carried out by research teams, where individuals responsible for designing the experiment differ from those executing them. In the case of philosophical experiments, there is no division of labor into design and execution stages, which

might be seen as an advantage. It also seems that results in thought experiments are not obtained randomly, as in some physics experiments. It would be difficult to expect the outcome of a thought experiment to be surprising to the person conducting it. However, thought experiments would be much less susceptible to chance events, such as equipment failure. For obvious reasons, the thought experimenter has greater control over the course of their reasoning.

Thought experiments—unlike real experiments—also do not require expensive and complicated research equipment. The philosophical equivalent of a physical laboratory could be the philosopher's mind, and the quality of such a "laboratory" would depend on the appropriate level of education and intelligence of the person conducting the thought experiment.

When can a thought experiment fail?

A thought experiment begins by presenting a possible, fictional situation and, if done correctly, will lead the recipient to a specific conclusion. A thought experiment proceeds in three stages: (1) presenting an imaginative (possible) situation ("Imagine a woman named Mary who does not know colors but knows all about the physics of colors..."), (2) the presented situation has a narrative character ("Mary sees a red rose and learns something new about the world..."), (3) the result of the presented narrative confirms or refutes a philosophical thesis ("So... Physicalism is false!"). This structured view of a thought experiment allows us to indicate how it can fail:

1. Unimaginability: A thought experiment can be challenged by pointing out that the situation presented is inconceivable (because it is inconsistent or described too generally).

2. Inconclusiveness: Even if the situation is imaginable, there are no good reasons to accept its outcome.

3. Lack of reference to the actual world: Even if the situation is imaginable, and there are good reasons to accept its outcome, it does not provide a basis for a claim about our world (Gendler, 2000, p. 22).

In the first sense, a thought experiment fails if the depicted state of affairs is unimaginable. The argument from unimaginability or inconceivability is a common criticism of some particularly extravagant thought experiments, such as those concerning the possibility of philosophical zombies (Chalmers, 1997). Indicating that zombies, beings physically identical to me but devoid of a first-person point of view, are inconceivable is a popular way of weakening such an argument. Therefore, if philosophical zombies are inconceivable, such a thought experiment could be considered unsuccessful.

In the second sense, an unsuccessful thought experiment would involve the experimenter being able to imagine the situation but inaccurately envisaging its course. In the well-known thought experiment by Frank Jackson in which, upon seeing a red rose for the first time, Mary learns something new about the world, leads to the conclusion that physicalism is false (Jackson, 1986). However, critics of this thought experiment might argue that the course of this experiment should be different. For instance, they might argue that, upon seeing the red rose, Mary exclaims that the rose looks exactly the way she thought it would—after all, Mary has all the knowledge about the physics of colors, and the color of the rose should not be new or surprising to her. A similar situation could occur in a real experiment if, for example, the experiment proceeded without disruptions but generated incorrect data (due to equipment damage or an error in the experiment's design).

An unsuccessful thought experiment in the third sense is one leading to a correct conclusion but failing to provide an answer to the question that prompted its conduct. In a real experiment, this might be a situation where the experiment yields correct results but fails to provide the data searched for by researchers. An example of an unsuccessful thought experiment might be Gottfried Leibniz's "Mill." Leibniz, formulating an argument against the idea that the human mind has a mechanical

nature, invites the reader to imagine the interior of a mechanical mind, where, just like in a mill, we would not be able to observe any mental phenomena but only "parts pushing one another, and never anything which would explain a perception" (Leibniz, 1720/2014, p. 17). We can imagine a mill, and we can agree with Leibniz that we won't see anything which would explain perceptions in it, whereas if we reject the mill-mind analogy, we will not agree with Leibniz's conclusion that mechanicism is false.

Conclusion

I started the paper with a story and the question of what is funny about it. If we consider thought experiments a reliable method of doing philosophy, the answer is nothing. The humor comes from adopting a particular conception of thought experiments, according to which thought experiments are a procedure in which everything is allowed. However, this conception seems inaccurate since, at each stage of a thought experiment, we can ask whether it was conducted correctly.

Roy Sorensen compared thought experiments to a compass (1992, p. 288). A compass is a simple, albeit useful tool for indicating direction. However, it is not a reliable device—for example, compass indications are unreliable around the North Pole. Few people know how the compass really works, although this is not an obstacle to the effective use of the device. Similarly, philosophical thought experiments conducted in imagination point to possible states of affairs. However, the ease of creating thought experiments—after all, telling "what if..." stories does not require special technical skills—does not translate into the reliability of the results obtained in this way. Therefore, being aware of the limitations of the products of imagination allows us to use thought experiments with more caution.

Thought experiments can resemble a compass in some respects and a magnifying glass in others. The fantastic stories given by philosophers are to meticulously test philosophical theories to see if the explanatory

power of these theories covers all possible situations. Suppose someone claims that knowledge is a true and justified belief. In that case, this view can be undermined by giving an example of someone having knowledge despite not meeting all the conditions given in the definition of knowledge. If someone claims that physicalism is true, so that everything, whatever exists, can be described in physical terms, this view can be challenged by giving an example of an object that cannot be described in physical terms. To provide a counter-example in philosophy is to present some imagined possible situation, that is, to propose a thought experiment.

Bibliography

Brown, J. (1991). *The Laboratory of Mind: Thought Experiments in the Natural Sciences*. London-New York: Routledge.

Chalmers, D. (1997). *The Conscious Mind*. Oxford University Press.

Daly, C. (2010). *Introduction to Philosophical Methods*. Peterborough: Broadview Press.

Dennett, D. (2013). *Intuition Pumps and Other Tools for Thinking*. New York: W.W. Norton & Company.

Gendler, T. (2000). *Thought Experiment: On the Power and Limits of Imaginary Cases*. New York: Garland Publishing.

Jackson, F. (1986). What Mary Didn't Know. *Journal of Philosophy, 83*(5), 291–295.

Kuhn, T. (1977). *The Essential Tension. Selected Studies in Scientific Tradition and Change*. Chicago: University of Chicago Press.

Leibniz, G. (2014). In L. Strickland (Ed.), *Leibniz's Monadology: A New Translation and Guide*. Edinburgh, UK: Edinburgh University Press. Original work published in 1720.

Nersessian, N. (2018). Cognitive science, mental modeling, and thought experiments. In M. Stuart, Y. Fehige, J. Brown (Eds.), *The Routledge Companion to Thought Experiments* (pp. 309–326). London: Routledge.

Norton, J. (2004). On Thought Experiments: Is There More to the Argument?. *Philosophy of Science*, *71*(5), 1139–1151. https://doi.org/10.1086/425238

Sorensen, R. (1992). *Thought Experiments*. Oxford: Oxford University Press.

Wilkes, K. (2003). *Real People: Personal Identity without Thought Experiments*. Oxford: Clarendon Press.

Williamson, T. (2007). *The Philosophy of Philosophy*. Malden: Blackwell.

Williamson, T. (2016). Knowing by Imagining. In A. Kind, P. Kung (Eds.), *Knowledge Through Imagination* (pp. 113–123). Oxford: Oxford University Press.

LABORATORIUM MENTIS | 2023, T. 1, NR. 1

DOI: 10.52097/lm.8152

James Tartaglia
Uniwersytet w Keele (Wielka Brytania)
0000-0002-5251-1908
j.tartaglia@keele.ac.uk

Wolna wola a wiara w determinizm[1]

Received: 4.11.2023 / Revised: 24.11.2023 / Accepted: 1.12.2023 / Available: 20.12.2023

Abstrakt:
Artykuł porusza kwestię wolnej woli i determinizmu za pomocą dyskusji nad paradoksem Newcomba, przedstawionej w postaci dialogu między duchami – paniami Traf i Przeznaczenie. Argumentuję, że przyjęcie determinizmu, sugerowanego przez metafizykę materialistyczną, stoi w sprzeczności z naszym doświadczeniem wolności wyboru. Paradoks Newcomba opisuje dylemat polegający na wyborze między jednym lub dwoma pudełkami w celu maksymalizacji zawartości tych pudełek, którą z góry określiła maszyna przewidująca wybór. Postać Heather, stawiając czoła temu dylematowi, symbolizuje ludzkość borykającą się z problemem wolnej woli w obliczu determinizmu. Twierdzę, że nasze doświadczenie wolności nie pozwala nam wierzyć, że determinizm jest prawdziwy, gdy odpowiednio się nad tą kwestią zastanowimy, podobnie jak Heather nie była w stanie uwierzyć w determinizm, kiedy musiała dokonać wyboru w sytuacji opisanej w paradoksie.

Słowa kluczowe:
wolna wola, determinizm, paradoks Newcomba, Robert Nozick

Jak cytować:
Tartaglia, J. (2023). Wolna wola a wiara w determinizm [polskie tłumaczenie]. *Laboratorium Mentis*, *1*(1), 33–39. https://doi.org/10.52097/lm.8152

1 Tekst jest adaptacją rodz. 3, cz. 4 książki autora pt. *Inner Space Philosophy: Why the next stage of human development should be philosophical, explained radically (suitable for wolves)*, która ma się ukazać 28 czerwca 2024 w wydawnictwie Iff Books (Winchester, UK).

Panie Traf i Przeznaczenie to duchy rozprawiające o wolnej woli i determinizmie. Starają się nie przyciągać uwagi ducha Filozofii, ponieważ uważają ją zarówno za przerażającą, jak i nudną. Paradoks Newcomba po raz pierwszy został przedstawiony filozofom w artykule Roberta Nozicka z 1969 r. (Nozick, 1969). Paradoks ten jest zazwyczaj omawiany w kontekście teorii decyzji, ale jeden z moich nauczycieli, J.J. Valberg, użył go do wyciągnięcia wniosków na temat wolnej woli i determinizmu w swojej nieopublikowanej książce zatytułowanej *Wola* (2011). W przeciwieństwie do Valberga, który przyjmuje determinizm i jest kompatybilistą, dochodzę do wniosku, że determinizm musi być fałszywy.

Pani Traf: Jeśli obiecasz, że zachowasz ciszę, opowiem ci, skąd wiem, że determinizm nie jest prawdziwy i dlaczego prawdziwy być nie może. Był to dla mnie wstrząs niemały, gdy się o tym dowiedziałam, gdyż już od czasów rewolucji naukowej w XVII w. zawsze zachęcano mnie do myślenia o sobie jako działającej na podstawie deterministycznych praw fizyki, właściwie z nimi tożsamej. Wówczas był to dla mnie obraz zupełnie nowy, dotrzymałam jednak kroku czasom – nie każdy ma zdolność do takiej przemiany. To dlatego przetrwałam tak długo. Mimo to zajęło mi to trochę czasu. W jednej chwili myślisz, że jesteś mistycznym duchem damy grającym ludzkimi losami, jakby to były pionki na szachownicy, w drugiej zaś okazujesz się nieuchronnym rozwojem sił natury. Cóż, tak naprawdę nigdy nie porzuciłam do końca pierwszego wizerunku siebie. Tak bardzo do mnie pasował, niemniej dałam się nakłonić na to „rozwijanie natury". Dlaczego nie? Jeśli nie możesz ich pokonać, dołącz do nich! Ten obraz jednak musi być chybiony. Przekonał mnie o tym fizyk o imieniu William Newcomb... Fizyk, eh, myślałbyś, że to byłby (raczej)... Nie! Nie wolno wymawiać tego słowa.

Newcomb wymyślił mnie jako posiadającą doskonałą wiedzę o wszechświecie fizycznym, zdolną dzięki temu przewidzieć decyzje pewnej osoby. Ta osoba – nazwijmy ją Heather – ma do wyboru dwie opcje, a ja mam przewidzieć, którą wybierze. Jest to dość proste, jeśli

wiesz wszystko o fizycznie zdeterminowanym wszechświecie, nieprawdaż? Zgodnie z moim nowym obrazem cała ta ogromna wiedza miała być jedynie moją samoświadomością... Poznaj samego siebie i tym podobne. Nic wielkiego. Zatem, znając wszelkie zmiany w mózgu Heather w trakcie podejmowania decyzji, mogę obliczyć, jaki to będzie miało wpływ na ruchy jej ust i języka (czyli co powie, gdy ogłosi swoją decyzję). Zatem z łatwością mogę przewidzieć jej wszelką decyzję.

Decyzja Heather dotyczy tego, czy wybierze pieniądze z dwóch skrzyń (B1 i B2), czy tylko z jednej (B2). B1 zawsze zawiera tysiąc dolarów, ale to, co zawiera B2, zależy od tego, jakie będą moje przewidywania. Jeśli stwierdzę, że wybierze tylko skrzynię B2, nagrodzę ją, umieszczając w niej milion dolarów. Jeśli jednak uznam, że będzie chciwa i wybierze obie skrzynie, to nie włożę do skrzyni B2 żadnych pieniędzy, otrzyma tylko tysiąc dolarów z B1.

Zabawa zaczyna się, gdy już – dokonawszy predykcji – umieszczę (albo i nie umieszczę) milion w B2, w zależności od tego, co przewidziałam. Gdy to się stanie, Heather ma pięć minut, aby zdecydować, czy chce obie skrzynki, czy tylko jedną. Możesz sobie wyobrazić, całe to wahanie! Pierwszym odruchem zawsze będzie wybór samej skrzynki B2. Tak dyktuje zdrowy rozsądek. Wie ona wszystko, co ci opowiedziałem, zna całą tę historię, więc myśli, że lepiej wybrać tylko B2, aby zdobyć milion. Uważa, że jeśli wybierze obie skrzynki, to skoro przewidziałam, że tak zrobi, nie dostanie miliona. Jeśli jednak ma odrobinę rozsądku, następną rzeczą, jaka jej przyjdzie do głowy, będzie:

> Heather (wedle tego, jak zapamiętała Pani Traf): Zaraz, zaraz, przecież Pani Traf już osądziła, co uczynię, jej osąd jest już przeszłym wydarzeniem! Pani Traf albo umieściła milion w skrzynce, albo nie. To już się dokonało i teraz nic tego nie zmieni. Zatem mogę równie dobrze wybrać obie skrzynki, bo jeśli milion jest już w B2, to będzie tam bez względu na to, co teraz uczynię. Wiem, że te dodatkowe tysiąc dolarów z B1 to niewiele, ale nie wiąże się z tym żadne ryzyko. Mogłabym i po nie sięgnąć, skoro mam taką możliwość.

> Moja decyzja nie może zmienić tego, co ona już uczyniła, więc... wybieram obie skrzynki, proszę. To moja ostateczna odpowiedź.

Pani Traf: Mogłabyś pomyśleć, że to jest pewnego rodzaju paradoks – tak to widział Newcomb – ponieważ racja przemawiająca za wyborem tylko jednej skrzynki wydaje się równie silna co racja za wyborem obu. Doszłam jednak do wniosku, że samo już rozważanie przez Heather wyboru obu skrzynek oznacza, iż nie może tak naprawdę uwierzyć w trafność moich przewidywań. Już samo rozważanie wyboru obu skrzynek, przy pełnym rozumieniu sytuacji, może oznaczać, że mi nie ufa. I jest to oczywiście w pełni zrozumiałe, jeśli mi nie ufa , ponieważ jeśli wybierze tylko skrzynkę B2, może skończyć z niczym (w scenariuszu, w którym fałszywie przewidziałam, że wybierze obie skrzynki i dlatego nie umieściłam w B2 żadnych pieniędzy).

Tak więc wybór obu skrzynek ma sens tylko wtedy, jeśli się mi nie ufa, albo raczej – nie wierzy w determinizm. Dla Heather to oznacza to samo, co dla większości podlegających mi współcześnie ludzi, ponieważ uważają, że jestem nieuchronnym nurtem rzeczywistości fizycznej lub czymś wspieranym przez nią, czy jakoś tak. Jeśli jednak Heather mi ufa prawdziwie, to wybierze tylko jedną skrzynkę, ponieważ spodziewa się, że – wiedząc, iż takiego dokona wyboru – już umieściłam milion w skrzynce.

Podsumowując, jeśli Heather mi ufa, powinna wybrać jedną skrzynkę, ale jeśli mi nie ufa, powinna wybrać obie. Ponieważ powinna mi ufać, to powinna wybrać jedną skrzynkę.

No cóż, myślałam o tym już od jakiegoś czasu, kiedy wpadłam na pomysł, żeby to przetestować eksperymentalnie, tak dla pewności. Pracowałam wówczas z pewną bizneswoman o imieniu Heather. Zarobiła miliardy w firmie internetowej, niezbyt miła osoba. Powiedziała mi, że próbuje zmienić przyszłość przyjaźni. No więc zapewniłam ją, że jej przeznaczeniem jest sukces, że cierpienie, które powoduje, zostanie zrekompensowane w dłuższej perspektywie, że powinna zignorować wszelkie ryzyko... Wiesz, to samo, co mam zwyczaj szeptać im do ucha.

W każdym razie pewnej nocy zabrałam ją do krainy snów, żeby wypróbować mój eksperyment. Kiedy doszłyśmy do momentu, w którym miała pięć minut na podjęcie decyzji, to, co odkryłam, zadziwiło mnie. Okazało się, że nie była nawet w stanie mi zaufać!

Pani Przeznaczenie: To jest to, o czym ci wcześniej mówiłam.

Pani Traf: Zamilknij na chwilę, dobrze? Teraz dochodzimy do ciekawej części. Spójrz, Heather była bardzo bystrą kobietą, więc zrozumiała scenariusz wystarczająco dobrze, żeby pojąć, że w ostatecznym rozrachunku ufać mi po prostu nie ma sensu. Oto, jak to wyjaśniła. Powiedziała to tak ładnie, że zacytuję ją słowo w słowo.

> Heather (wedle słów Pani Traf): Jeśli mam ufać, że zawsze przewidujesz trafnie, to mam wybór między B2, oznaczającym zdobycie miliona dolarów, albo B1 i B2, dającym zaledwie tysiąc dolarów. Wybór, którego teraz dokonam, zdecyduje, ile pieniędzy otrzymam, ponieważ jakikolwiek by ten wybór nie był, będzie on tym, który już przewidziałaś. Ale to przecież niemożliwe, nie mogę podejmować decyzji dotyczących czegoś, co już wydarzyło się w przeszłości. Obawiam się zatem, że po prostu nie mogę Ci ufać. Jest czymś paradoksalnym doświadczać wolności, podczas gdy wierzy się w determinizm, a ponieważ nie mogę odrzucić doświadczenia wolności, nie mogę zaprzeczyć, że obecnie mam wybór do dokonania, ani zaprzeczyć faktowi, że tu jestem i go dokonuję. Nie zostawiasz mi żadnej alternatywy dla odrzucenia determinizmu.

Pani Traf: Wytłumaczyła to dobrze, nie sądzisz? Od razu zdałam sobie sprawę, że miała rację. Musiałam to przyznać przed samą sobą: nie jestem naprawdę wszechwiedzącym, deterministycznie rozwijającym się naturalnym światem, ani, jak lubią myśleć o tym ludzie, deterministycznym światem fizycznym, który zasadniczo (!) pozwala im (!) przewidzieć, co się wydarzy. Trochę to mnie zdeprymowało, przyznam,

ale miało to wiele sensu, ponieważ z pewnością nie zawsze miałam rację, kiedy byłam tajemniczą Panią-duchem. Czasem mówiłam dowódcy wojskowemu, że zmierza ku chwale, że jego armia jest niezwyciężona, a potem... Oj!

Pani Przeznaczenie: Czy rozmawiamy o tej samej Heather, którą mi przekazałaś, tej australijskiej potentatce internetowej?

Pani Traf: Tejże samej. Teraz wiesz, czemu ją porzuciłam.

Pani Przeznaczenie: Ze mną też nie miała zbyt dobrze. Ciągle przechodziła pod drabinami, nie pozdrawiała srok, nie zakładała bielizny przynoszącej szczęście. Nie było dnia, w którym nie stanęłaby mi na odcisk. Szybko pękłam i nakłoniłam ją do sprzedaży jej udziałów w firmie w najgorszym możliwym momencie. Nawet to nie zmieniło jej zachowania. Nie zaczęła ze mną rozmawiać więcej niż przedtem. Nie błagała mnie o lepszy los, więc, obawiam się, że stałam się nieco mściwa – namówiłam ją do podjęcia pracy jako wykładowca filozofii. Po tej decyzji jej upadek był już przesądzony.

Filozofia: Aha, zdało mi się, że ktoś wezwał moje imię. Ktoś mnie wzywał!!!

Pani Traf (wzdychając głęboko): O nie! Tylko nie ty!…

Komentarz

Twierdzę, że nie możemy konsekwentnie wierzyć w prawdziwość determinizmu, jeśli rozważymy praktyczne implikacje życia w świecie deterministycznym. Tego dotyczy odkrycie Heather w jej ostatniej mowie. Jeśli ma pełne zaufanie do maszyny przewidującej (Pani Traf / determinizm), to będzie całkowicie przekonana, że maszyna trafnie

przewidziała, czy wybierze obie skrzynki, czy tylko jedną. Podejmując swoją decyzję, powinna być przekonana, że jeśli wybierze tylko jedną skrzynkę, to w środku znajdzie milion dolarów, jeśli zaś wybierze dwie skrzynki, to w jednej z nich nie znajdzie nic. Jednak tego rodzaju przekonanie oznacza uznanie, że w chwili wyboru ma ona wpływ na to, czy pieniądze są w skrzynce. Z jej obecnej perspektywy epistemicznej wygląda to tak, jakby miała wybrać między naciśnięciem jednego albo drugiego przycisku. Jeden z nich daje jej milion dolarów, a drugi nie. Jednak jest to coś, w co tak naprawdę nie może wierzyć, ponieważ wie, że w chwili podejmowania decyzji pieniądze albo będą już w skrzynce, albo nie i że jej wybór nie może tego zmienić. A jednak, jeśli w pełni ufa przewidującemu, nie powinna mieć wrażenia, że jej wybór rzeczywiście to może zmienić, to znaczy, że w tej chwili to od niej zależy wybór między dwoma różnymi wynikami. Jeśli wybiera tylko jedną skrzynkę, to dlatego, że chce pieniędzy i uważa, że musi coś zrobić, aby je zdobyć, tj. podjąć właściwą decyzję.

Wniosek z tego argumentu jest taki, że doświadczając wolnej woli, jak wszyscy, nie możemy w pełni zaufać maszynie przewidującej, a zatem nie możemy konsekwentnie wierzyć w determinizm. Obraz determinizmu zaprezentowany przez Panią Traf odzwierciedla mój pogląd, że teza determinizmu, która zazwyczaj jest przedstawiana raczej jako naukowa, jest w rzeczywistości zakorzeniona w starożytnych przesądach takich jak astrologia (Tartaglia, 2020, rozdz. 6, sekcja 4).

Bibliografia

Nozick, R. (1969). Problem Newcomba i dwie zasady wyboru. W: N. Rescher (red.), *Eseje ku czci Carla G. Hempela* (s. 114–146). Dordrecht: Springer.

Tartaglia, J. (2020). *Philosophy in a Technological World: Gods and Titans*. London: Bloomsbury.

LABORATORIUM MENTIS | 2023, VOL. 1, NO. 1

DOI: 10.52097/lm.8153

JAMES TARTAGLIA
Keele University (United Kingdom)
0000-0002-5251-1908
j.tartaglia@keele.ac.uk

Free will and believing in determinism[1]

Received: 4.11.2023 / Revised: 24.11.2023 / Accepted: 1.12.2023 / Available: 20.12.2023

Abstract:
The article addresses the issue of free will and determinism through a discussion of Newcomb's paradox, presented as a dialogue between the spirits of Lady Luck and Fate. I argue that commitment to determinism, which is suggested by materialist metaphysics, is in contradiction with our experience of freedom of choice. Newcomb's paradox describes the dilemma of choosing between either one or two boxes in order to maximise the quantity of money these boxes contain, which has been determined by the machine predicting what your decision will be. The character of Heather, faced with this dilemma, symbolizes humanity grappling with the issue of free will in the face of determinism. I claim that our experience of freedom prevents us from believing that determinism is true when we properly reflect on the issue, just as Heather could not believe in determinism when she had to make the choice in the situation described in the paradox.

Keywords:
free will, determinism, Newcomb's paradox, Robert Nozick

How to cite:
Tartaglia, J. (2023). Free will and believing in determinism [English original]. *Laboratorium Mentis*, *1*(1), 40–46. https://doi.org/10.52097/lm.8153

1 Adapted from Chapter 3, Part 4 of James Tartaglia's *Inner Space Philosophy: Why the next stage of human development should be philosophical, explained radically (suitable for wolves)*. Winchester, UK: Iff Books (publication date: 28 June 2024).

Lady Luck and Fate are spirits having a conversation about free will and determinism. They are trying not to attract the attention of the Philosophy spirit since they find her both frightening and boring. Newcomb's paradox was first brought to the attention of philosophers in a 1969 article by Robert Nozick (Nozick, 1969). The paradox is usually discussed in the context of decision theory, but one of my teachers, J.J. Valberg, used it to draw conclusions about free will and determinism in his unpublished book entitled *Will* (2011)—unlike Valberg, who accepts determinism and is a compatibilist, the conclusion I draw is that determinism must be false.

Lady Luck: If you promise to keep your voice down, I'll tell you how I know determinism isn't true, why it just can't be. It came as quite a shock to me when I found out because ever since the scientific revolution of the 17th century, I've been encouraged to think of myself as backed up by the deterministic laws of physics, as pretty much the same thing as them, really. It was a new image for me at the time, thoroughly up to date—not everyone's got the wherewithal to remake themselves like that; it's the reason for my longevity. Still, it took some getting used to. One minute, you think you're a mystical lady-spirit playing with human lives as if they were chessmen; the next, you're the inevitable temporal unfolding of natural forces. Well, I never really gave up on the first self-image; it's so much more *me*, but I did throw myself into the 'unfolding of nature' thing. Why not? If you can't beat 'em, join 'em! But it doesn't work; I found that out from a physicist called William Newcomb … a physicist, eh, you'd have thought it'd be a … nope, mustn't say the word.

Newcomb imagined me having perfect knowledge of the physical universe and using it to predict a person's decision. This person – let's call her Heather – is given two choices, and I get to predict which choice she's going to make – which is easy enough when you know everything about the physically determined universe, right? According to my new image, all that vast knowledge was only supposed to be my

own self-awareness anyway … know thyself and all that, no biggie. So, knowing how the bits and pieces of Heather's brain change as she's making her choice, I can work out the effect this'll have on how she'll move her mouth and tongue (i.e., what she'll say when she announces her choice), so I can predict her decision every time, easy-peasy.

Heather's choice is about whether she wants the money in two boxes (B1 and B2) or just one (B2). B1 always contains $1000, but what B2 contains depends on my prediction. If I predict she's going to choose to have box B2 only, then I'll reward her by putting $1,000,000 in it. But if I predict she's going to be greedy by choosing both boxes, then I won't put any money in box B2; she'll just get the $1000 from box B1.

The fun starts once I've made my prediction, and so either put the million in B2 or not, depending on the prediction. Once that's done, Heather has 5 minutes to choose whether she wants both boxes or only one. Well, you can imagine the humming and harring that one causes! Her first reaction is always going to be that she should choose box B2 only; that's just common sense. She knows everything I've told you, the whole shebang, so she figures that she'd better choose B2 only to get the million. She thinks that if she chooses both boxes, then I'll have predicted she would, so she won't get the million.

But then, if she's got an ounce of sense, the next thing that's going to occur to her is:

> Heather (as remembered by Lady Luck): Hang on a cotton-picking minute; Lady Luck's *already* made the prediction. The prediction is a *past event*! Lady Luck either put the million in the box or she didn't; that's already happened, and nothing can change it now. So I might as well choose both boxes because if the million's already in B2, it'll be there whatever I choose now; I know the extra thousand from B1 isn't much, but there's no risk involved, so I might as well have it if I can. My decision can't change what she's already done, so … I'll have both boxes, please—final answer.

Lady Luck: Now, you might think this is a kind of paradox—that's how Newcomb saw it—since her motive for choosing only the one box seems just as strong as her motive for choosing both. What I've come to realise, however, is that if Heather is even contemplating the choice of both boxes, then she can't really believe my predictions are always correct. Just thinking about choosing both boxes, once she fully understands the situation, can only mean that she doesn't trust me. And it makes perfect sense for her to choose both if she doesn't trust me, of course, since if she goes for only box B2, then she might end up with nothing—in the scenario where I falsely predicted that she'd choose both boxes, and so didn't put any money in B2.

So, choosing both boxes only makes sense if you don't trust me, or rather, if you don't think determinism is true—that amounts to the same thing for Heather, as for most of my people these days, because they think I'm the inevitable flow of physical reality, or backed up by it, or whatever. But if Heather does trust me, then she'll choose only one box since she'll expect me to have known she'd make that choice, and so she'd also expect me to have already put the million in the box.

To sum up, then, if Heather trusts me, she should choose one box, but if she doesn't, she should choose both. Since she's supposed to trust me, she ought to choose one box.

Well, I'd been thinking this over for some time when I had the bright idea of trying it out experimentally, just to be sure. There was a businesswoman called Heather, whom I was working with at the time—she made billions with an internet company, a nasty piece of work. She told me she was trying to change the future of friendship. Well, I assured her it was her fate to succeed, that the suffering she caused would be outweighed in the long run, that she should ignore all risks … you know, the usual stuff I whisper in their inner spaces. Anyway, one night, I took her to a dream world to try my experiment. When we got to the bit where she had 5 minutes to decide, what I discovered astonished me. It turned out she wasn't even capable of trusting me!

Fate: That's what I was telling you earlier.

Lady Luck: Shut up a minute, will you? We're just getting to the good bit. You see, Heather was a very clever woman, so she understood the scenario well enough to understand that, at the end of the day, trusting me just doesn't make sense. This is how she explained it—she said it so nicely; I'll repeat her exact words.

> Heather (as remembered by Lady Luck): If I'm to trust that you always get the predictions right, then I have a choice between B2 only to get the million dollars, or B1 and B2 to get only a thousand dollars. The choice I make right now will decide how much money I get because whatever that choice is, it'll be the choice you predicted. But that's impossible; I can't make a choice about something which happened in the past. So, I'm afraid I simply can't trust you. It's paradoxical to experience freedom while believing in determinism, and since I can't renounce the experience of freedom—I can't deny that I currently have a choice to make; as a plain matter of fact, I'm here, and I do—so you leave me with no alternative except to renounce determinism.

Lady Luck: She explained that well, don't you think, Fate? I saw straight away that she was right, so I had to admit to myself that I wasn't really an all-knowing, deterministically unfolding natural world—or, as humans like to think of it, a deterministic physical reality which allows *them* (!) to *in principle* (!) predict what's coming next. It was a bit deflating, I'll admit, but it made a lot of sense since I certainly didn't always get things right when I was a mysterious lady-spirit. Sometimes, I'd tell a military leader he was heading for glory, that his army was invincible, and then … oops!

Fate: Are we talking about the same Heather you handed over to me, the Aussie internet tycoon?

Lady Luck: The very same; now you know why I dropped her.

Fate: She didn't do very well with me either. She kept walking under ladders, didn't salute magpies, didn't have lucky underpants, so hardly a day passed when I wasn't irritated by her. Before long, I snapped and got her to sell up her shares in the company at the worst time possible. Well, even that didn't change her behaviour. She didn't start talking to me more than before, she wasn't begging me for better luck, so I'm afraid I got a little vindictive—I persuaded her to apply for a job as a philosophy lecturer, after that her downfall was assured.

Philosophy: Aha, I thought I heard somebody mention my name … I've been invoked !!!

Lady Luck (groaning): Oh no, not you …

Commentary

I am arguing that we cannot consistently believe that determinism is true once the practical implications of living in a deterministic world are considered. We see Heather's discovery of this in her final speech. If she has complete faith in the prediction machine (Lady Luck / determinism), then she will be fully confident that the machine has accurately predicted whether she will choose to have both boxes or only one. When she makes her decision, then, she will believe that if she chooses one box only, she will subsequently open the box to find $1,000,000 inside, whereas if she chooses two boxes, she will subsequently open the box to find nothing inside. To believe this, however, is to believe that she currently has a choice over whether the money is in the box. From her current epistemic situation, it is as if there were two buttons she must choose between pressing, one which gives her $1,000,000 and one which does nothing. This is something she cannot truly believe, however, because she knows the question of whether the

money is in the box or not has already been settled. She knows that at the moment she makes her choice, the money will either already be in the box or the box will already be empty and that her choice cannot affect this—and yet, if she fully trusts the predictor, it cannot help seeming to her that her choice will indeed affect this, in the sense that it is currently up to her to choose between two different outcomes. If she chooses only one box, after all, it will be because she wanted the money and felt that she needed to do something to get it, i.e. make the right choice. The conclusion of the argument is that while experiencing free will, as we all do, we cannot fully trust the prediction machine and hence cannot consistently believe in determinism. The portrayal of determinism as "Lady Luck" reflects my view that the thesis of determinism, which is typically portrayed as a rather scientific thesis, is actually rooted in ancient superstitions such as astrology (see Tartaglia, 2020, chapter 6, section 4).

Bibliography

Nozick, R. (1969). Newcomb's Problem and Two Principles of Choice. In N. Rescher (Ed.), *Essays in Honor of Carl G. Hempel* (pp. 114–146). Dordrecht: Springer.

Tartaglia, J. (2020). *Philosophy in a Technological World: Gods and Titans*. London: Bloomsbury.

LABORATORIUM MENTIS | 2023, T. 1, NR. 1

DOI: 10.52097/lm.8154

WOJCIECH JANKOWSKI
Uniwersytet Gdański (Polska)
0000-0002-5431-027X
wojciech.jankowski@prawo.ug.edu.pl

Jednoręki bandyta

Received: 4.11.2023 / Revised: 24.11.2023 / Accepted: 6.12.2023 / Available: 20.12.2023

Abstrakt:
Artykuł prezentuje eksperyment myślowy dotyczący nieco dziwnej planety Losoria. Na tej planecie odpowiedzialność za pewne rodzaje przestępstw przypisywana jest w sposób całkowicie losowy. Eksperyment ma na celu wzmocnienie naszych intuicji moralnych wspierających zasadę kontroli w sporze o tzw. traf moralny. Ponadto stymuluje on refleksję nad niemoralnymi przesłankami różnicowania odpowiedzialności karnej ponoszonej za skutki działań znajdujących się poza kontrolą podmiotu. Autor wskazuje, że możliwym rozwiązaniem pewnych aporii wynikających z trafu konsekwencji jest przyjęcie sprawiedliwości naprawczej jako formy racjonalizacji kary.

Słowa kluczowe:
traf moralny, traf prawny, sprawiedliwość naprawcza, zasada kontroli

Jak cytować:
Jankowski, W. (2023). Wolna wola a wiara w determinizm [polski oryginał]. *Laboratorium Mentis*, *1*(1), 47–54. https://doi.org/10.52097/lm.8154

Aby wczuć się w atmosferę opisywanej poniżej historii, wyobraźmy sobie, że dzieje się ona na planecie rodem z *Dzienników gwiazdowych* Stanisława Lema, a my obserwujemy zachowania jej mieszkańców niczym Ijon Tichy, który z dystansem godnym badacza stara się zrozumieć motywacje tubylców bez pochopnego oceniania ich zwyczajów[1]. Na planecie tej – nazwijmy ją Losoria – na każdym rogu znajdują się olbrzymie urządzenia, które zbiegiem okoliczności przypominają naszych ziemskich jednorękich bandytów. W razie wygranej, która zdarza się stosunkowo często, z maszyny wypada słodycz będący przysmakiem Losorian. Z korzystaniem z maszyny wiąże się jednak pewne ryzyko. Otóż w każdej z maszyn znajduje się przypadkowo wybrany Losorianin, któremu w 999 na 1000 przypadków użycia maszyny absolutnie nic nie zagraża. Jednak w nieszczęśliwym tysięcznym wypadku użycie maszyny powoduje uruchomienie mechanizmu dekapitującego uwięzionego nieszczęśnika (u Losorian, podobnie jak u ludzi, pozbawienie głowy łączy się w sposób konieczny ze śmiercią).

Wiedza o potencjalnych skutkach korzystania z automatu jest na planecie powszechna, ze względu jednak na małe prawdopodobieństwo okropnych następstw część obywateli nadal decyduje się z nich korzystać. Władze Losorii, wiedząc o potencjalnie straszliwych skutkach użycia maszyny, postanowiły zabronić korzystania z niej pod jakimkolwiek pozorem. Nie do końca wiadomo, z jakiego powodu, ale automatów nie można odgrodzić i każdy w każdym momencie może z nich korzystać (bardzo możliwe, że powody mają charakter religijny). Z kolei ze względu na problemy kadrowe w losorskiej policji kontrola tego, czy ktoś z automatu korzysta, ma charakter jedynie wybiórczy. To, co jednak dziwi najbardziej, to kary przewidziane przez Losorię dla swoich obywateli. Otóż kiedy ktoś zostanie przyłapany na skorzystaniu z maszyny w przypadku, gdy wynikiem gry jest wypadnięcie

[1] Dla czytelników spoza Polski lub nieznających twórczości Stanisława Lema proponuję, aby wyobrazili sobie, że czytają rozdział z przewodnika autostopem po galaktyce.

ciasteczka, będzie skazany na niezbyt dotkliwą grzywnę. W wypadku jednak, gdy wynikiem gry jest śmierć Losorianina uwięzionego wewnątrz maszyny, karą jest zesłanie do kolonii karnej na co najmniej 15 lat. Jako przybysze z innej planety jesteśmy akurat świadkami, jak dwóch przyjaciół, którzy właśnie zgłodnieli, podchodzi do automatu, mając nadzieję na coś słodkiego. Pierwszy z nich pociąga za dźwignię i po chwili otrzymuje upragniony smakołyk. Gdy jednak drugi pociąga za dźwignię na maszynie pojawiają się trzy czaszki – oznaka przegranej. Tak się składa, że ta konkretna maszyna akurat była obserwowana przez losorską policję. Widzimy więc, że pierwszy z przyjaciół dostaje jedynie pouczenie i niewielką grzywnę, drugi z nich musi szykować się na długie lata spędzone w kolonii karnej.

Zapewne w ziemskim przybyszu powyższe zachowanie losorskich władz wzbudziłoby sprzeciw. Czy jednak pochodzący ze współczesnej Ziemi obserwator, krytykując system sprawiedliwości i moralność Losorian, nie narażałby się w ten sposób na zarzut hipokryzji? Jeśli chcemy krytycznie oceniać praktykę Losorian, może wpierw powinniśmy zmienić nasze oceny wobec pewnych kwestii na Ziemi.

Powrót na Ziemię

Miesiąc przed napisaniem tego tekstu cała Polska wstrząśnięta była wypadkiem na autostradzie, w którym zginęła rodzina z małym dzieckiem. Powodem (czy też sprawcą) wypadku był prawdopodobnie kierowca BMW, który znacznie przekroczył dozwoloną prędkość[2]. Czytając komentarze dotyczące tej tragicznej sprawy, można odnieść wrażenie, że społeczeństwo jest w tej kwestii jednoznaczne i potępia działania sprawcy. Tzw. opinia publiczna domaga się także wyjątkowo surowego ukarania sprawcy. Jednocześnie jednak, jak pokazują badania,

[2] „Prawdopodobnie", gdyż w trakcie pisania tego artykułu postępowanie wciąż się toczy, zaś zgodnie z art. 42 ust. 3 Konstytucji RP każdego uważa się za niewinnego, dopóki jego wina nie zostanie stwierdzona prawomocnym wyrokiem sądu.

większość użytkowników pojazdów mechanicznych często przekracza dozwoloną prędkość. O frywolnym podejściu do przepisów ruchu drogowego w naszym kraju trafnie traktuje reportaż Bartosza Józefiaka „Wszyscy tak jeżdżą". Czy nie jest to jednak przejaw moralnej schizofrenii? Z dużym prawdopodobieństwem osoby wyrażające kategoryczne sądy moralne na temat „zabójcy z BMW" przymykają oko na znaki z ograniczeniami prędkości, traktując je bardziej jak wskazówki niż realne zakazy. Czy jednak rzeczywiście kierowca BMW jest bardziej moralnie odpowiedzialny niż ktoś, kto podobnie przekracza prędkość, jednak bez tragicznych konsekwencji? Inna, przynajmniej na razie, jest jego odpowiedzialność prawnokarna. Ale czy tak być powinno? Pytania te związane są z jedną z ciekawszych debat w filozofii moralnej XX w. i dotykają odpowiednio kwestii tzw. trafu moralnego i trafu prawnego.

Problem trafu moralnego dotyczy tego, czy czynniki pozostające poza kontrolą jednostki powinny wpływać na jej moralną ocenę. Niejednokrotnie intuicyjną odpowiedzią na tak postawiony problem jest przeczenie i silne obstawanie przy tzw. zasadzie kontroli. Wedle zasady kontroli możemy przypisać odpowiedzialność jedynie za czynniki pozostające pod kontrolą działającego podmiotu. Zwolennikiem zasady kontroli był np. Immanuel Kant: „Dobra wola nie jest dobra ze względu na swoje dzieła i skutki ani ze względu na swą zdatność do osiągnięcia jakiegoś zamierzonego celu, lecz jedynie przez chcenie" (Kant, wyd. 1971, s. 5–6). Problem z zasadą kontroli pojawia się, gdy zdamy sobie sprawę, że wiele sytuacji, w których przypisujemy ludziom zasługę lub winę, jest pod silnym wpływem okoliczności pozostających poza kontrolą jednostki. Dochodzi w takim przypadku do konfliktu naszych intuicji. Z jednej strony chcemy bowiem w sytuacjach takich jak ta kierowcy BMW silnie obarczyć odpowiedzialnością, z drugiej nie chcemy zrezygnować z zasady kontroli. Problem ów najlepiej przedstawił Thomas Nagel w książce *Pytania ostateczne* (wyd. 1997, s. 38–53). Nagel twierdzi, że jeżeli chcielibyśmy konsekwentnie stosować zasadę kontroli, to ostatecznie możemy dojść do wniosku, że przypisanie odpowiedzialności nigdy nie będzie możliwe. Ilustruje tę obserwację za

pomocą kilku eksperymentów myślowych (rozróżniając jednocześnie odmienne rodzaje trafu moralnego).

Rodzaje trafu moralnego

Pierwszy rodzaj trafu to traf rezultatu. Można go zobrazować za pomocą przykładu dwóch kierowców, którzy w dokładnie takich samych warunkach znacznie przekraczają dozwoloną prędkość. Jednak tylko w wypadku jednego z nich na drogę wychodzi przechodzień, co kończy się przypadkiem. Intuicyjnie możemy być bardziej skłonni do obwiniania kierowcy, który spowodował szkodę, nawet jeśli obaj działali na takim samym poziomie nieostrożności. Różnicując jednak ich odpowiedzialność, opowiadamy się przeciw zasadzie kontroli, bowiem jedynym, co odróżnia tych sprawców od siebie, są konsekwencje ich działania, na które nie mieli wpływu.

Drugi rodzaj trafu to tzw. traf okoliczności. Załóżmy, że dwóch nastoletnich braci bliźniaków zostało rozdzielonych na początku lat 30. XX w. Jeden z nich wyjechał na studia do Argentyny, drugi zaś został w rodzinnych Niemczech. Skutkiem tego pierwszy spędził wojnę w Buenos Aires, nie wyrządzając nikomu krzywdy. Drugi zaś, będąc oficerem SS, dopuścił się wielu okrucieństw. Można by jednak przypuszczać, że gdyby zmieniły się okoliczności i drugi z braci także wyjechałby do Argentyny, to owych zbrodni nigdy by się nie dopuścił. Jeśli zatem chcielibyśmy konsekwentnie trzymać się zasady kontroli, to tu także musielibyśmy ocenić obu braci w jednakowy sposób. To jednak wydaje się być mocno sprzeczne z naszymi podstawowymi intuicjami.

Trzecim rodzajem trafu wyróżnionym przez Nagela jest tzw. traf konstytutywny. Odnosi się on do tego, że na to, jaką osobą się stajemy, duży wpływ mają genetyka, wychowanie i różne doświadczenia życiowe. Nasz charakter i temperament, które odgrywają kluczową rolę w naszych decyzjach moralnych, w dużej mierze nie są naszym własnym wyborem. Przykładowo: coraz częściej wskazuje się na genetyczne podłoże wielu patologicznych zachowań. Egzemplifikacją może

być tzw. gen MAOA, który zyskał sobie przydomek „genu wojownika". Gen wojownika w połączeniu z wychowywaniem się w warunkach przemocy domowej ma odpowiadać za skłonność do agresji i obniżoną samokontrolę. My sami nie mamy wpływu na środowisko, w którym nas wychowano, a tym bardziej na zestaw genów, z którym przyszliśmy na świat. Jeśli jednak to one są głównymi „przyczynami" naszych poczynań, to chcąc zachować wierność zasadzie kontroli, mielibyśmy trudności z przypisywaniem odpowiedzialności komukolwiek za cokolwiek.

Powyższa argumentacja podważa konsekwentne trzymanie się zasady kontroli. Zgodnie z tą argumentacją, jeżeli odpowiadać możemy jedynie za czynniki pozostające pod naszą kontrolą, to w ostateczności nigdy nie będzie można przypisać nikomu odpowiedzialności, bo jeśli dokładnie się przyjrzeć, to pod naszą kontrolą nie pozostaje właściwie nic. Jeśli nie akceptujemy całkowitej rezygnacji z przypisywania ludziom odpowiedzialności, to zmuszeni jesteśmy, przynajmniej w pewnym stopniu, zasadę kontroli odrzucić. Konkluzją Nagela jest więc konieczność zaakceptowania roli trafu w naszych ocenach moralnych.

W stronę silnej zasady kontroli

Eksperyment myślowy zaprezentowany na początku tego eseju ma jednak inny cel. Sprowadza on do absurdu konsekwencje odrzucania zasady kontroli (przynajmniej na poziomie trafu rezultatu). Jeśli bowiem opisaną powyżej praktykę prawniczą Losorian oceniamy w sposób negatywny, oznacza to, że przyjmujemy zasadę kontroli. Wydaje się bowiem, że przynajmniej w przypadku trafu rezultatu, gdybyśmy zaakceptowali traf jako mogący wpływać na nasze oceny moralne, powinniśmy takie same czyny dwóch przyjaciół używających jednorękiego bandyty, a także ich samych oceniać w diametralnie różny sposób. Wówczas jednak samo dokonywanie ocen moralnych zaczyna przypominać grę w jednorękiego bandytę.

Nawiązując do postawionego wcześniej pytania, musimy stwierdzić, iż potępienie Losorskiej praktyki oznaczałoby, że powinniśmy przyjrzeć się także pewnym rozwiązaniom z naszej ziemskiej rzeczywistości. Eksperyment z jednorękim bandytą ma w oczywisty sposób nawiązywać do problemu wypadków samochodowych. Nagroda w postaci ciasteczka jest podobnie mała jak nagroda czasu zaoszczędzonego przy bardzo szybkiej jeździe samochodem. Podobne, przynajmniej w zamierzeniu, jest także prawdopodobieństwo zabicia kogoś w jednorękim bandycie i zabicia kogoś na drodze wskutek przekroczenia prędkości. Wydaje się, że przynajmniej w przypadku trafu konsekwencji powinniśmy jednak opowiedzieć się za przyjęciem zasady kontroli. Jeżeli ktoś zatem przyjmuje zasadę kontroli (przynajmniej w przypadku trafu rezultatu), to jest szansa, że krytycznie oceni system prawa karnego zarówno na Losorii, jak i na Ziemi. Dlaczego bowiem dwie jednostki, które ocenilibyśmy moralnie w taki sam sposób, poniosą tak rażąco różnorodne konsekwencje prawne?

Tutaj objawia się właśnie problem trafu prawnego. Odrzucenie trafu w rzeczywistości moralnej nie musi z konieczności oznaczać odrzucenia go w rzeczywistości prawnej. Mogą bowiem istnieć pozamoralne racje dla różnicowania odpowiedzialności prawnej dwóch osób za czynniki pozostające poza ich kontrolą. Może to być choćby prewencyjny aspekt kary. Tu jednak pojawiają się dwie podstawowe wątpliwości. Po pierwsze, czy jesteśmy w stanie zaakceptować inkorporowanie instytucji kozła ofiarnego do naszego systemu prawnego (bo tym *de facto* będzie pechowiec z naszych kazusów)? Po drugie, czy taka prewencja będzie skuteczna? Zarówno osoba przekraczająca prędkość, jak i Losorianin ciągnący za wajchę, decydując się na tego rodzaju działania, prawdopodobnie biorą pod uwagę jedynie karę za przekroczenie prędkości/pociągniecie za wajchę, nie zaś karę za spowodowanie śmierci. Ktoś, kto pod wpływem powyższej uwagi zostałby przekonany o konieczności wyrugowania trafu także z prawa (a przynajmniej z prawa karnego), musiałby dokonać wyboru jednej z trzech poniższych możliwości:

1. Podnosimy niskie kary, tj. zarówno Losorianin, któremu wypadł batonik, jak i ten, który spowodował śmierć, jadą do kolonii karnej na 15 lat.

2. Obniżamy wysokie kary, tj. zarówno Losorianin, któremu wypadł batonik, jak i ten, który spowodował śmierć, dostają jedynie grzywnę.

3. Uśredniamy kary, tj. obaj sprawcy jadą do kolonii karnej na 1 rok.

Czy zdecydowalibyśmy się na któryś z powyższych zabiegów tu na Ziemi? Pierwszy z nich wydaje się być drakoński (choć prawdopodobnie skuteczny dla celów prewencyjnych). Ostatni zaś nie czyniłby zadość naszej, jak się wydaje, niezwykle dużej potrzebie przypisania odpowiedzialności za powstałą tragedię. Zostaje więc rozwiązanie środkowe. W każdym jednak z tych wypadków ten, kto spowodowałby zagrożenie na drodze, odpowiadałby tylko za stopień tego zagrożenia, ale już nie za to, czy rzeczywiście nastąpił jakiś jego skutek. Można jednak zastanowić się nad jeszcze innym rozwiązaniem, w którym, co prawda, sama kara oparta jest jedynie na czynie sprawcy i pozostaje podobna wobec obu podmiotów, jednak tym, co podlega różnicowaniu, jest odpowiedzialność odszkodowawcza. Wydaje się być to rozwiązanie o tyle optymalne, że zachowujemy zasadę kontroli na gruncie prawa karnego, jednocześnie wprowadzając element trafu na gruncie prawa cywilnego (w ramach którego element ryzyka akceptuje się z dużo większą swobodą). Odpowiedzią na problem trafu prawnego wydaje się zatem przyjęcie modelu sprawiedliwości naprawczej (Jankowski, 2021).

Bibliografia

Konstytucja RP (1997). Dz.U. 1997 nr 78 poz. 483.

Jankowski, W. (2021). W poszukiwaniu pozamoralnych racji dla trafu prawnego. *Studia Prawnoustrojowe*, (53), 253–270. https://doi.org/10.31648/sp.6866

Kant, I. (1971). *Uzasadnienie metafizyki moralności*, tłum. M. Wartenberg, oprac. R. Ingarden. Warszawa: Państwowe Wydawnictwo Naukowe. Oryginalne dzieło wydane w 1785 r.

Nagel, T. (1997). *Pytania ostateczne*, przeł. Adam Romaniuk. Warszawa: Fundacja Aletheia. Oryginalne dzieło wydane w 1979 r.

LABORATORIUM MENTIS | 2023, VOL. 1, NO. 1

DOI: 10.52097/lm.8154

WOJCIECH JANKOWSKI
University of Gdańsk (Poland)
0000-0002-5431-027X
wojciech.jankowski@prawo.ug.edu.pl

One-armed Bandit

Received: 4.11.2023 / Revised: 24.11.2023 / Accepted: 6.12.2023 / Available: 20.12.2023

Abstract:
The article presents a thought experiment about a slightly queer planet of Losoria. On this planet, responsibility for certain types of crime is assigned in a completely random manner. The experiment is designed to strengthen our moral intuitions advocating the principle of control in the dispute over so-called moral luck. In addition, it stimulates reflection on non-moral rationales for differentiating criminal responsibility for the consequences of actions beyond the subject's control. The author points out that a possible solution to some of the aporias arising from the result luck is to adopt restorative justice as a form of rationalizing of punishment.

Keywords:
moral luck, legal luck, restorative justice, principle of control

How to cite:
Jankowski, W. (2023). One-armed Bandit [English translation]. *Laboratorium Mentis*, *1*(1), 55–62. https://doi.org/10.52097/lm.8154

To immerse ourselves in the story's atmosphere, let us imagine that it takes place on a planet reminiscent of Stanisław Lem's "Star Diaries." We observe the behaviour of its inhabitants just like Ijon Tichy, who, with the detachment befitting a researcher, tries to understand the motivations of the natives without haste in judging their customs.[1] On this planet—let us call it Losoria—enormous devices resembling our earthly slot machines stand on every corner. In the event of a win, which happens relatively often, a sweet treat favoured by the Losorians is dispensed. However, using the machine involves some risk. In each of these machines, there is a randomly chosen Losorian, for whom absolutely nothing is a threat in 999 out of 1000 machine usage cases. Yet, in the one-thousandth case, using the machine triggers a mechanism that decapitates the unfortunate captive (for Losorians, just as for humans, beheading causes invariable death).

Knowledge of the potential consequences of using the machines is widespread on the planet. However, due to the low probability of the horrific outcome, some citizens still choose to use them. The authorities of Losoria, aware of the potentially dreadful consequences of using the machines, have decided to prohibit its use under any circumstances. It is unclear why, but the machines cannot be fenced off, and anyone can use them at any time (there are very likely religious reasons for this). Due to the low number of personnel in the Losorian police, the control over whether someone uses the machine is selective. However, the most surprising thing is the penalties prescribed by Losoria for its citizens. When someone is caught using the machine, and the game results in the dispensing of a cookie, they are sentenced to a fine which is not very severe. However, if the game results in the death of a Losorian imprisoned inside the machine, the punishment is deportation to a penal colony for at least 15 years. As visitors from another planet, we

[1] For nonpolish readers or those unfamiliar with the works of Stanisław Lem, I suggest imagining that you are reading a chapter from a hitchhiker's guide to the galaxy.

happen to witness two friends approaching the machine in the hope of getting something sweet. The first one pulls the lever and, after a moment, receives the coveted treat. However, when the second one pulls the lever, three skulls appear on the machine—a sign of loss. As it happens, this particular machine was being observed by the Losorian police. Therefore, the first friend receives only a warning and a fine, while the second one is to expect long years in a penal colony.

Undoubtedly, the behaviour of the Losorian authorities described above would likely provoke protests in an Earthly observer. However, would a visitor from contemporary Earth, when criticising the justice system and morality of Losoria, not expose themselves to the charge of hypocrisy? If we are to blame Losoria, we need to revise our judgments on certain Earthly matters first.

Back on Earth

A month before writing this text, all of Poland was shaken by an accident on the highway in which a family with a young child lost their lives. The cause (or perpetrator) of the accident was likely the driver of a BMW, who significantly exceeded the speed limit[2]. Reading the comments regarding this tragic incident, one might have the impression that society is unanimous in condemning the perpetrator's actions. The so-called public opinion also demands an exceptionally severe punishment for the perpetrator. However, as research shows, most motor vehicle users often exceed the speed limit. A report by Bartosz Józefiak titled "Wszyscy tak jeżdżą" (Everyone Drives Like That) accurately addresses the frivolous approach to traffic regulations in our country. Isn't this, however, a manifestation of moral schizophrenia? It

[2] „Probably", as at the time of writing this article, the proceedings are still ongoing, and according to Article 42(3) of the Constitution of the Republic of Poland, everyone is presumed innocent until their guilt has been established by a final court judgement.

is highly likely that individuals expressing categorical moral judgments about the "BMW killer" ignore speed limit signs, treating them more as guidelines than real prohibitions. However, does the BMW driver really have a higher moral responsibility than someone who similarly exceeds the speed limit but without tragic consequences? Another matter, at least for now, is their legal responsibility. But should it be that way? These questions are related to one of the more intriguing debates in 20th-century moral philosophy and pertain to the issues of moral luck and legal luck.

The problem of moral luck pertains to whether factors beyond an individual's control should influence their moral evaluation. The intuitive response to this problem is often to deny it and strongly adhere to the principle of control. According to the principle of control, one can take responsibility only for factors under their control. Immanuel Kant, for example, advocated the principle of control: "A good will is not good because of its effects and consequences, nor because of its ability to achieve some intended goal, but only because of the will." (Kant, 1785/1998, p. 8 [4:394]) The problem with the principle of control arises when we realise that circumstances beyond the individual's control strongly influence many situations in which we attribute credit or blame to people. This leads to a conflict of our intuitions. On one hand, in cases like that of the BMW driver, we want to hold them responsible. On the other hand, we don't want to abandon the principle of control. Thomas Nagel presented this problem most effectively in his essay "The View from Nowhere" (Nagel, 1979). Nagel argues that if we applied the principle of control consistently, we might ultimately conclude that attributing responsibility will never be possible. He illustrates this observation with several thought experiments (at the same time distinguishing different types of moral luck).

Types of Moral Luck

The first type of moral luck is the luck of the result. It can be illustrated using an example of two drivers who, under the same conditions, significantly exceed the speed limit. However, only in the case of one of them does a pedestrian step onto the road, resulting in an accident. Intuitively, we may be more inclined to blame only the driver who caused harm, even if both drivers were equally reckless. However, by differentiating their responsibility, we reject the principle of control because the only thing that sets these perpetrators apart from each other are the consequences of their actions, over which they had no control.

The second type of moral luck is called the luck of circumstances. Suppose two teenage twin brothers were separated in the early 1930s. One of them went to study in Argentina, while the other stayed in their native Germany. As a result, the first one spent the war in Buenos Aires without causing harm to anyone. The other, however, committed many atrocities as an SS officer. We can assume that if the circumstances had changed and both brothers had gone to Argentina, none of them would ever have committed those crimes. If we wanted to consistently adhere to the principle of control, we would have to assess both brothers in the same way. However, this is in strong conflict with our basic intuitions.

The third type of moral luck, identified by Nagel, is called constitutive luck. It relates to the fact that genetics, upbringing, and various life experiences significantly influence the kind of person we become. Our character and temperament, which play a crucial role in our moral decisions, are largely not of our own choosing. For example, there is increasing evidence of the genetic basis of many pathological behaviours. An example is the so-called MAOA gene, which has earned the nickname the "warrior gene." The warrior gene, in combination with childhood in a violent domestic environment, is said to be responsible for a tendency toward aggression and reduced self-control. We have no control over the environment in which we grow up, let alone the genes we were born with. However, if these factors are the main "causes"

of our actions, then, adhering to the principle of control, it would be difficult to assign responsibility to anyone for anything.

The above argumentation challenges the persistent adherence to the principle of control. According to this argumentation, if we can only be held responsible for factors under our control, then ultimately, no one could ever be held responsible—because, upon closer examination, virtually nothing would remain under our control. If we do not accept the complete abandonment of attributing responsibility to people, then we are forced to, at least to some extent, reject the principle of control. Nagel's conclusion is the necessity of accepting the role of luck in our moral assessments.

Towards the strong principle of control

The thought experiment presented at the beginning of this essay has a different purpose. It reduces the consequences of rejecting the principle of control to absurdity (at least in the case of luck of the outcome). If we negatively assess the Losorian legal practice, it means we accept the principle of control. It seems that at least in the case of luck of the outcome, if we were to accept luck as influencing our moral judgments, we should evaluate the actions of the two friends using the slot machine and even the friends themselves in a radically different way. However, in that case, making moral judgments begins to resemble playing a slot machine.

Referring to the earlier question, our condemnation of the Losorian practice would mean we should also consider certain solutions in our earthly reality. The slot machine experiment is meant to allude to the problem of car accidents. The reward in the form of a cookie is just as small as the reward of the time saved by driving very fast. The probability of killing someone on the slot machine and that of killing someone on the road due to speeding are (intended to be) similar. However, it seems that, at least in the case of luck of the outcome, we should accept the principle of control. So, if someone accepts the principle of control

(at least in the case of luck of the outcome), there is a chance that they will critically assess the criminal legal system on both Losoria and Earth. Why should two individuals, whom we would morally judge in the same way, face such drastically different legal consequences?

Here lies the problem of legal luck. The rejection of luck in the moral reality does not necessarily imply its rejection in the legal reality. There may be non-moral reasons for differentiating the legal liability of two individuals for factors beyond their control. It may be, for instance, the preventive aspect of punishment. However, two fundamental doubts arise here. First, can we accept the incorporation of the scapegoat institution into our legal system (because that's what the unlucky one from our cases would actually be)? Second, will such prevention be effective? When deciding on such actions, both a person speeding and the Losorian pulling the lever probably only consider the penalty for speeding or pulling the lever, not the penalty for causing death. Someone convinced of the need to eliminate luck from the law (or at least from criminal law) due to the above consideration would have to choose one of the following three possibilities:

1. We raise low penalties, i.e., both the Losorian who got a candy bar and the one who caused death go to a penal colony for 15 years.

2. We lower high penalties, i.e., the Losorian who got a candy bar and the one who caused death receive only a fine.

3. We average the penalties, i.e., both culprits go to a penal colony for one year.

Would we decide on one of the above procedures here on Earth? The first one seems draconian (though probably effective for preventive purposes). The last one would not satisfy our seemingly very strong need to assign responsibility for the resulting tragedy. So, the middle ground remains. In each of these cases, however, the one who would pose a threat on the road would be held accountable only for the degree of that threat but not for whether any actual consequence occurred. However, one can consider another solution in which, while the punishment is indeed based only on the act of the perpetrator

and remains similar for both subjects, what undergoes differentiation is the liability for damages. It seems to be an optimal solution, as it retains the principle of control in criminal law while introducing an element of accident into civil law (within which the risk element is accepted with much greater freedom). Therefore, the answer to the problem of legal accidents seems to be adopting a restorative justice model (Jankowski, 2021).

Bibliography

Constitution of Poland Dz.U. 1997 nr 78 poz. 483

Jankowski, W. (2021). W poszukiwaniu pozamoralnych racji dla trafu prawnego [In search of non-moral reasons for legal accuracy]. *Studia Prawnoustrojowe*, (53), 253–270. https://doi.org/10.31648/sp.6866

Kant, I. (1998). *Groundwork of the Metaphysics of Morals*, trans. M. Gregor. Cambridge: Cambridge University Press.

Nagel, T. (1979). *Mortal Questions*. Cambridge: Cambridge University Press.

LABORATORIUM MENTIS | 2023, T. 1, NR. 1

DOI: 10.52097/lm.8156

ARTUR SZUTTA
Uniwersytet Gdański (Polska)
0000-0002-1062-4808
artur.szutta@ug.edu.pl

O pewnej intuicji na temat nabywania cnoty

Received: 28.09.2023 / Revised: 20.10.2023 / Accepted: 5.11.2023 / Available: 20.12.2023

Abstrakt:
Niniejszy artykuł dotyczy kwestii moralnego doskonalenia człowieka za pomocą technologii. Proponuję eksperyment myślowy, który pozwala dostrzec pewną nową rację przeciwko implementacji takiego doskonalenia. Osiągnięcie cnoty drogą, która obejmuje własny wysiłek podejmowania i realizowania moralnie słusznych decyzji, zasługuje na większy szacunek, a także pozwala na uznanie, że jesteśmy (współ)autorami tego, kim w sensie moralnym się stajemy. Tego rodzaju autotworzenie wydaje się ważną częścią sensownego życia, której pozbawia nas przemiana w istotę moralną za pomocą sztucznego zabiegu moralnego doskonalenia.

Słowa kluczowe:
moralne doskonalenie, cnota, intuicja, sens życia

Jak cytować:
Szutta, A. (2023). O pewnej intuicji na temat nabywania cnoty [polski oryginał]. *Laboratorium Mentis*, *1*(1), 63–74. https://doi.org/10.52097/lm.8156

Debata wokół doskonalenia moralnego

Wyzwania współczesnego świata, rozwój technologiczny w szczególności, zdaniem niektórych stawiają nas wobec konieczności doskonalenia naszej moralnej kondycji. Przykładowo Julian Savulescu oraz Ingmar Persson (2012, 2013) twierdzą, że współczesna technologia daje ludziom niemoralnym możliwość czynienia zła na dużą skalę, np. zniszczenia wielomilionowego miasta, a nawet całego narodu przy użyciu nowoczesnej broni (tzw. brudnej bomby atomowej, wirusa etc.)[1]. Ten sam rozwój technologiczny, jak zauważają, pozwoli nam pewnego dnia udoskonalić moralnie naszą naturę, czy to na poziomie motywacyjnym (poprawiając naszą zdolność do empatii), czy to poznawczym. Poprawienie zdolności moralnych za pomocą technologii, np. manipulacji genetycznej lub farmakologii, będzie nie tylko interesującą, może nawet kuszącą opcją, ale też moralnym obowiązkiem, w imię uniknięcia bardzo prawdopodobnego wielkiego zła.

Idea doskonalenia moralnego jest przedmiotem krytyki ze strony wielu autorów. Tak oto np. Allen Buchannan (2009) obawia się, że osoby udoskonalone moralnie będą posiadały wyższy status moralny, a tym samym cieszyły się większymi prawami niż zwyczajne osoby ludzkie. Inni autorzy, np. John Harris (2011) czy Michael Hauskeller (2013), argumentują, że w wyniku moralnego doskonalenia będziemy tak zdeterminowani do czynienia dobra przez nasze poprawione emocje, pragnienia lub inne zachodzące w naszych ciałach mechanizmy, że staniemy się niczym automaty pozbawione wolnego wyboru. Krytycy moralnego poprawiania człowieka za pomocą technologicznych interwencji wskazują także na problem ustalenia, według jakich standardów powinniśmy doskonalić ludzi moralnie, skoro panuje silna niezgoda w etyce co do podstawowych zasad etycznych (Schaefer, 2014).

---

1 Idea moralnego doskonalenia ma coraz więcej zwolenników (zob. np. Specker et al. 2014).

Pośród obiekcji można wymienić także zarzut z nieprzewidzianych konsekwencji prób poprawiania moralnego za pomocą farmakologii lub innych technik (Fabiano, 2018). Głosy krytyczne nie zostały bez odpowiedzi ze strony zwolenników doskonalenia moralnego. Można powiedzieć, że obecnie toczą się liczne i zaawansowane dyskusje (zob. Lavazza i in., 2019), w których obie strony formułują kolejne argumenty, kontrargumenty i kontr-kontrargumenty.

W niniejszym artykule chcę zaproponować eksperyment myślowy, który daje nowe spojrzenie na problem doskonalenia moralnego, pozwala bowiem zidentyfikować pewną intuicję przemawiającą przeciwko sztucznemu poprawianiu moralnemu człowieka, która nie jest wskazywana (przynajmniej nie wprost) w znanych mi publikacjach.

Założenia

Zanim zaprezentuję sam eksperyment, przyjmijmy pewne założenie co do samej technologii doskonalenia moralnego. Pomińmy wszelkie techniczne niedoskonałości obecnych i dostępnych w przyszłości metod doskonalenia moralnego ludzi. Przyjmijmy, że efektem tych zabiegów będą ludzie nieodróżnialni na poziomie behawioralnym, świadomościowym i motywacyjnym od ludzi, którzy byliby doskonałym uosobieniem arystotelesowskiego człowieka cnotliwego. Załóżmy także, że omawiane tu doskonalenie ludzi będzie powszechnie dostępne, nie będzie miało niepożądanych skutków, nie będzie pozbawiało poprawionych moralnie osób zdolności do refleksji i działania na podstawie racji moralnych, czy też szerzej – zdolności do autentycznych decyzji moralnych. Czy wówczas uznalibyśmy, że nie istnieją żadne racje przeciwko realizacji idei doskonalenia moralnego?

Eksperyment

Aby odpowiedzieć na to pytanie, proponuję przeprowadzić eksperyment myślowy. Wyobraźmy sobie cztery równoległe światy nieodróżnialne

od siebie nawzajem poza jednym szczegółem. Każdy z nich zamieszkuje ta sama osoba, Jan, niemniej w każdym z tych światów Jan ma nieco inną historię. Odpowiednio w każdym z możliwych światów *a*, *b*, *c*, *d* istnieją Jan_a, Jan_b, Jan_c oraz Jan_d. Wszyscy wymienieni są w określonym czasie, mierzonym jednocześnie dla wszystkich czterech światów, powiedzmy w czasie t_n, osobami moralnie cnotliwymi w sensie arystotelesowskim, to znaczy posiadają stałe dyspozycje do niezawodnego rozpoznania tego, co jest moralnie właściwe w danych okolicznościach, oraz do motywowanego tym rozpoznaniem skutecznego czynienia dobra (Arystoteles, wyd. 2007, EN 1116a).

Jedyną rzeczą, która różni wszystkich Janów, jest sposób, w jaki swoje cnoty nabyli. Jan_a stał się cnotliwy tradycyjnym sposobem – poprzez naśladowanie innych osób, kiedy był jeszcze małym dzieckiem, następnie dzięki wytrwałemu treningowi w podejmowaniu właściwych decyzji, pokonywaniu swoich słabości, stopniowemu budowaniu w sobie właściwych dyspozycji, być może z niejedną porażką (szczególnie na początku moralnego rozwoju) po drodze.

Jan_b w pewnym momencie t_{n-m} został wybrany bez jego wiedzy i zgody do tajnego programu rządowego, którego celem było stworzenie i przetestowanie projektu doskonalenia moralnego za pomocą pewnej technologii. Jej szczegółów nie będę tutaj określał, aby uniknąć w tym miejscu (skądinąd ciekawej, lecz drugoplanowej dla celów tego artykułu) dyskusji o jej słabościach lub niedociągnięciach (zgodnie z założeniem eksperymentu byłaby to technologia bezpieczna)[2].

Przypadek $Jana_c$ różni się od sytuacji $Jana_b$ tym, że to nie rząd podjął decyzję o zastosowaniu technologii doskonalenia moralnego, ale rodzice Jana. Pragnęli oni moralnie cnotliwego dziecka, a bali się, że

[2] Naturalnym odruchem w trakcie czytania na temat sytuacji $Jana_b$ byłoby formułowanie zastrzeżeń co do moralnej akceptowalności samych metod sztucznej implementacji moralnych dyspozycji. Uznaję, że takie zastrzeżenia mogłyby być uzasadnione i stanowić ważny punkt wyjścia dla dyskusji, niemniej chcę, aby czytelnik je zignorował, ponieważ celem niniejszego eksperymentu nie będzie ocena metod, a samej idei sztucznego moralnego poprawiania człowieka.

sami nie będą w stanie właściwie wychować go metodami naturalnymi. Jan_d natomiast, być może sfrustrowany swoją złą kondycją moralną, sam zgłosił się do programu doskonalenia moralnego i został poddany zabiegowi przy pełnej świadomości i za wyrażeniem zgody.

Na potrzeby tego artykułu przyjmijmy, że w czasie t_n wszyscy wymienieni wyżej Janowie są nieodróżnialni pod względem behawioralnym, motywacyjnym, kognitywnym oraz emocjonalnym. Przez nieodróżnialność behawioralną rozumiem nieodróżnialność dającego się zaobserwować z trzecioosobowej perspektywy zachowania (może nawet dającej się zaobserwować aktywności mózgów i innych części ciała istotnych dla działania moralnego). Nieodróżnialność motywacyjna oznacza kierowanie się tymi samymi racjami moralnymi w podejmowaniu decyzji. Nieodróżnialność kognitywna i emocjonalna[3] polegają na tym, że świat, wszelkie w nim obiekty, ich cechy i relacje między nimi wyglądają (są poznawczo ujmowane) dla wszystkich Janów tak samo oraz mają oni dyspozycje do przeżywania takich samych emocji w takich samych okolicznościach.

Intuicyjna ocena

Wydaje się, że pomimo opisanej powyżej nieodróżnialności wszystkich czterech Janów, fakt odmiennej historii, innego sposobu, w jaki nabyli oni cnotę, decyduje o odmiennej ocenie Janów. Jak sądzę, intuicyjne jest twierdzenie (zakładam, że większość czytelników tę intuicję będzie podzielać), że Jan_a zasługuje na większy podziw z naszej strony niż pozostali Jan_b, Jan_c oraz Jan_d[4]. Podziw ten powinien bowiem

[3] Nie zakładam ani nie odrzucam tutaj, że emocje są z definicji czymś pozapoznawczym (akognitywnym). Rozstrzygnięcie takie nie jest istotne z punktu widzenia celu prezentowanego tutaj eksperymentu.

[4] Myślę, że na nieco lepszą ocenę moralną, w pewnej mierze także nasz podziw (ale w mniejszym stopniu niż Jan_a) zasługuje także Jan_d, który sam zwrócił się z prośbą o moralne udoskonalenie.

obejmować nie tylko aktualne możliwości Jana, ale także trud, wytrwałość i zaangażowanie woli Jana poprzez cały proces jego moralnego wzrostu. Aktualny stan $Jana_a$ jest relatywnie (w porównaniu do pozostałych Janów) w największym stopniu jego zasługą. Wydaje się zaś, że uwzględnienie zasługi w osiągnięciu cnoty powinno mieć miejsce w moralnej ocenie Janów.

Dyskusja

Czy z powyższej intuicji wynika coś dla debaty na temat doskonalenia moralnego? Wydaje się, że (przy założeniu jej słuszności) intuicja ta stanowi rację *prima facie* przeciwko wdrażaniu moralnego poprawiania ludzi. Niemniej zwolennik takiego doskonalenia mógłby sformułować dwa argumenty przeciwko opisanej powyżej intuicji.

Po pierwsze, sam fakt intuicji co do większej zasługi $Jana_a$ oraz fakt większego dla niego podziwu nie oznacza, że *de facto* Jan zasługuje na wyższą ocenę moralną. Intuicja ta może być przecież bezużyteczną pozostałością procesu ewolucji, który nie uwzględniał naszego rozwoju technologicznego. W najlepszym razie taka intuicja mogła odnosić się do warunków, w których ludzie żyli, zanim powstała technologia umożliwiająca poprawianie ludzkiej natury. Zmiany ewolucyjne (w tym moralne intuicje) są powolne, zaś rozwój techniki gwałtowny. Stąd nasze niedopasowanie naszych intuicji do obecnej sytuacji cywilizacyjnej, w tym możliwości moralnego doskonalenia[5].

Po drugie, nawet jeśli przypisalibyśmy powyższej intuicji wstępną wiarygodność, to musiałaby ona ustąpić innej intuicji dotyczącej porównania prawdopodobieństwa sukcesu w osiągnięciu cnoty metodą tradycyjną z prawdopodobieństwem osiągnięcia cnoty za pomocą technologii. Z obserwacji wiemy, że bardzo rzadko, jeśli w ogóle, ludziom

[5] Zwolennikiem takiego argumentu jest na przykład Peter Singer (2005). Więcej na temat zarzutów wobec tezy o wiarygodności intuicji moralnych zob. Szutta 2018.

udaje się dzięki własnej, opartej na seriach autonomicznych decyzji, praktyce osiągnąć bardzo wysoki poziom moralny zasługujący na miano cnoty. Tymczasem technologia, przynajmniej w świetle założeń przyjętych w eksperymencie, daje, jeśli nie zupełną pewność, to wyraźnie wyższe prawdopodobieństwo niż w przypadku tradycyjnej metody osiągania cnoty (lub przynajmniej czegoś od cnoty nieodróżnialnego, jeśli uznamy, że cnota na mocy pojęciowej konieczności obejmowałaby także konieczność odpowiedniego sposobu jej nabywania).

Tak więc, jeśli stoimy w obliczu podjęcia decyzji na poziomie społecznym/politycznym, czy wprowadzać moralne doskonalenie za pomocą środków technicznej manipulacji, czy też trwać przy tradycyjnych metodach wychowawczych, powinniśmy zważyć obie intuicje: z jednej strony tę dotyczącą niższego statusu moralnego sztucznie wytworzonej cnoty (lub czegoś, co jest łudząco do cnoty podobne)[6], z drugiej prawdopodobieństwa osiągnięcia sukcesu w implementacji cnoty jedną i drugą metodą. Jeśli możemy z relatywnie wysokim prawdopodobieństwem stwierdzić, że np. w milionowym społeczeństwie będziemy mieli milion (albo prawie milion, albo nawet połowę miliona) osób *à la* Jan$_{b,c \text{ lub } d}$, czy nie będzie to bardziej pożądany stan rzeczy (który będziemy moralnie zobowiązani realizować) niż milionowe społeczeństwo z kilkoma cnotliwymi wyjątkami?[7] Czy nie będzie to

[6] Nie chcę tutaj rozstrzygać, czy *de facto* można dyspozycje Janów$_{b\text{-}d}$ określać mianem cnoty, warto jednak zastanowić się nad tezą, że cnota jest czymś czterowymiarowym, mianowicie, że obejmuje nie tylko kondycję podmiotu w czasie t_n, ale że obejmuje cały proces w czasie od t_1 (moment początkowy istnienia osoby zdolnej do stania się podmiotem) do t_n (moment stania się cnotliwym). Kolejnym składnikiem/cechą cnoty obok jej czterowymiarowości byłaby, w świetle przedstawionego tu eksperymentu, autodeterminacja podmiotu. Innymi słowy, cnota musiałaby pochodzić od wewnątrz, nie z zewnątrz podmiotu. Podobne intuicje na temat nabywania cnoty (lub w ogóle moralnego doskonalenia) mają Jason Erbel (2018), Ruben Herce (2019) oraz James Tartaglia (2020, rozdz. 5).

[7] Powyższy argument można przedstawić w innej jeszcze formie. Możemy wyobrazić sobie Jana$_e$, który świadomie zrezygnował (albo wobec którego zrezygnowano) z moralnego poprawiania na rzecz tradycyjnej formy rozwoju moralnego, któremu

sytuacja bardziej pożądana, a jej realizacja naszym moralnym obowiązkiem, nawet jeśli będące efektem doskonalenia moralnego osoby będą pozbawione możliwości (pełnej) zasługi oraz (pełnego) podziwu dla ich moralnej kondycji?

W odpowiedzi na powyższe krytyki można wysunąć następujące argumenty. Po pierwsze, być może niekoniecznie nasze intuicje moralne, a ta dotycząca sposobu nabywania cnoty w szczególności, są niegodne przynajmniej wstępnego zaufania. Ponieważ na temat wiarygodności intuicji moralnych w ogólności pisałem gdzie indziej (Szutta, 2018, rozdz. 7) oraz istnieje już spora literatura na ten temat[8], tutaj ograniczę się do argumentu za możliwością wiarygodności konkretnej intuicji, zgodnie z którą znaczenie ma sposób nabywania cnoty. Niekoniecznie ma ona jedynie charakter przestarzałego mechanizmu (albo myślowego nawyku), który odnosił się do czasów, gdy czasochłonne kształtowanie charakteru przez autonomiczne uczestnictwo w odpowiednich praktykach było jedyną skuteczną metodą osiągnięcia cnoty. Być może intuicja ta odnosi się do naszego rozumienia sensownego życia.

jednak nie powiodło się osiągnięcie cnoty czy też chociaż zbliżenie się do niej. Czy nie uznalibyśmy, że Jan_e zasługuje na negatywną ocenę moralną, że zmarnował czas na nieudane próby, podczas gdy miał do dyspozycji skuteczną alternatywę? W przypadku, gdy mamy pewność, że dana osoba nie ma żadnych szans albo ma bardzo małe szanse na to, aby stać się moralnie cnotliwą bądź przynajmniej dobrą moralnie na jakimś minimalnym poziomie, można by argumentować, że mamy wręcz obowiązek poprawić ją moralnie za pomocą technologii. W innym razie życie takiej osoby byłoby w jakiejś mierze zmarnowane. Warto także dodać, że w opisanym powyżej porównaniu nie chodzi tylko o skuteczność w znaczeniu możliwości osiągnięcia pożądanego skutku, ale także o efektywność mierzoną w kategoriach czasowych. Porównując tradycyjne nabywanie cnoty oraz metodę sztucznego poprawiania moralnego, powinniśmy także wziąć pod uwagę fakt, że metoda tradycyjna jest bardziej czasochłonna i po drodze do osiągnięcia cnoty podmiotowi moralnemu zdarzyć się może wiele moralnych porażek, co, przy założeniach, że dysponujemy skuteczną metodą sztucznego doskonalenia, można by wykluczyć.

8 Z nowszych publikacji interesującą obronę intuicji prezentują Bengson i in., 2020.

Możliwe, że istotnym elementem sensownego życia jest to, aby przynajmniej w jakiejś znaczącej części być autorem samego siebie, stwarzać siebie poprzez autonomiczne decyzje. W celu wsparcia takiej interpretacji bronionej intuicji proponuję kolejny, krótki eksperyment. Wyobraź sobie czytelniku/czytelniczko, że twoi rodzice w momencie twoich narodzin poprosili o wszczepienie w twojej głowie małego chipu, który – odpowiednio stymulując twój mózg – będzie motywował ciebie do realizacji konkretnego, zaplanowanego przez rodziców, wzorca życiowego. Załóżmy, że plan ten polega na tym, abyś miał/-a odpowiednie zainteresowania w szkole, wybrał/-a odpowiednie studia, odpowiedniego partnera życiowego itd. Załóżmy też, że – dokonując „właściwych" wyborów – będziesz odwoływał/-a się do rozpoznanych przez ciebie obiektywnych racji, będą one jednak raczej pełniły rolę uzasadnienia już określonych uprzednio przez rodziców wyborów, a nie przyczyn/racji determinujących twoje wybory. Czy, gdyby przez przypadek udało ci się odkryć taki czip i jego funkcję, twoje życie nie straciłoby dla ciebie przynajmniej jakiejś ważnej części swojej wartości?

Czy zatem intuicja, którą pozwala nam uchwycić proponowany powyżej eksperyment z czterema Janami, nie dotyczy tak naprawdę faktu, że podziw dla osiągnięć ludzi oznacza zarazem pewnego rodzaju potwierdzenie, że ich życie zawiera jakiś istotny nadający mu sens element? Że tym elementem sensownego życia jest właśnie fakt, że to, co w nim osiągamy, osiągamy przynajmniej w pewnym zakresie również dzięki własnym wysiłkom i decyzjom?

Na koniec proponuję argument pomocniczy, skierowany jedynie do tych osób, które przyjmują istnienie osobowego Boga, który stworzył ludzi i który pragnie ich rozwoju moralnego. Jeśli Bóg jest dobry, wszechmogący i wszechwiedzący i gdyby to było dla nas dobre, mógłby stworzyć nas od razu doskonałymi, bez konieczności niepewnego rodzenia się naszego moralnego charakteru w bólach, bez konieczności regresów i upadków, bez ryzyka porażki. Jeśli mimo swojej dobroci, wszechwiedzy i wszechmocy nie stworzył nas doskonałymi, to może uznał, że jest inne lepsze rozwiązanie: abyśmy mieli swój udział

w tworzeniu siebie, byli współautorami nas samych. Jeśli odrzucimy tezę, że Bóg nie istnieje, wówczas dysponujemy kolejną racją wspierającą przekonanie, że sposób, w jaki stajemy się moralnie dobrymi ludźmi ma znaczenie, bowiem znaczenie ma to, czy możemy efekt swojego życia, to kim się stajemy, przynajmniej po części nazwać swoim dziełem.

Konkluzja

Zarówno zaproponowany powyżej eksperyment myślowy, jak i przedstawiona dyskusja wskazują na istnienie pewnej, pomijanej do tej pory racji przeciwko wprowadzaniu moralnego poprawiania w rozumieniu sztucznej ingerencji w organizm człowieka. Racją tą jest fakt, że – implementując w ten sposób dyspozycję do skutecznego czynienia dobra moralnego – pozbawiamy osoby szansy na autonomiczne autotworzenie. Racji tej jednak nie należy rozumieć jako ostatecznie przesądzającej o niedopuszczalności sztucznego moralnego doskonalenia. Ma ona charakter *prima facie*. Ostateczne rozstrzygnięcie dyskutowanej tu kwestii wymaga rozważenia wielu innych racji, np. tego, czy cena rezygnacji z (lub pozbawienia) szansy na autokreację nie jest warta tego, aby osiągnąć społeczeństwo, w którym znamienita większość, jeśli nie wszyscy postępują dobrze moralnie. Niniejszy tekst należy traktować jako propozycję kolejnej racji (kwestii), którą trzeba uwzględnić w szerszej dyskusji na temat dopuszczalności i wartości moralnego doskonalenia[9].

[9] Pragnę podziękować za krytyczne uwagi i komentarze poczynione wobec wstępnej wersji tego artykułu Krzesimirowi Cholewie, Michałowi Dominówowi, Pawłowi Homelowi, Jakubowi Nowickiemu, Stanisławowi Kamińskiemu, Pawłowi Sikorze, Nataszy Szutcie, Błażejowi Szymichowskiemu, Jamesowi Tartaglii oraz Emilii Wrońskiej.

Bibliografia

Arystoteles (2007). *Etyka nikomachejska*, tłum. D. Gromska, Warszawa: Wydawnictwo Naukowe PWN. Oryginał wydany ok. 300 p.n.e.

Bengson, J., Cuneo, T., Shafer-Landau, R. (2020). Trusting moral intuitions. *Noûs*, *54*(4), 956–984. https://doi.org/10.1111/nous.12291

Buchanan, A. (2009). Moral status and human enhancement. *Philosophy & Public Affairs*, *37*(4), 346–381. https://www.jstor.org/stable/40468461

Eberl, J.T. (2018). Can prudence be enhanced?. *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine*, *43*(5), 506–526. https://doi.org/10.1093/jmp/jhy021

Fabiano, J. (2018). *Probing the Risks of Moral Enhancement* [PhD thesis]. University of Oxford.

Harris, J. (2011). Moral enhancement and freedom. *Bioethics*, *25*(2), 102–111. https://doi.org/10.1111/j.1467-8519.2010.01854.x

Hauskeller, M. (2013). The "Little Alex" Problem. *The Philosophers' Magazine*, (62), 74–78. https://doi.org/10.5840/tpm20136299

Herce, R. (2019). Is human enhancement possible if it comes from the outside?. *Scientia et Fides*, *7*(2), 165–170.

Lavazza, A., Reichlin, M. (2019). Introduction: moral enhancement. *Topoi*, *38*(1), 1–5. https://doi.org/10.1007/s11245-019-09638-5

Persson, I., Savulescu, J. (2013). Getting moral enhancement right: the desirability of moral bioenhancement. *Bioethics*, *27*(3), 124–131. https://doi.org/10.1111/j.1467-8519.2011.01907.x

Savulescu, J., Persson, I. (2012). Moral enhancement, freedom and the god machine. *The Monist*, *95*(3), 399–421. https://doi.org/10.5840/monist201295321

Schaefer, G., Schaefer, G.O. (2014). *Moral enhancement and moral disagreement* [PhD thesis]. Oxford University, UK.

Specker, J., Focquaert, F., Raus, K., Sterckx, S., Schermer, M. (2014). The ethical desirability of moral bioenhancement: a review of reasons. *BMC Medical Ethics*, *15*(1), 1–17. https://doi.org/10.1186/1472-6939-15-67

Singer, P. (2005). Ethics and intuitions. *The Journal of Ethics*, *9*(3/4), 331–352. http://www.jstor.org/stable/25115831

Szutta, A. (2018). *Intuicje moralne. O poznaniu dobra i zła*. Lublin: Wydawnictwo Academicon. https://doi.org/10.52097/acapress.9788362475612

Tartaglia, J. (2020). *Philosophy in a Technological World: Gods and Titans*. Bloomsbury Publishing.

LABORATORIUM MENTIS | 2023, VOL. 1, NO. 1

DOI: 10.52097/lm.8157

ARTUR SZUTTA
University of Gdańsk (Poland)
0000-0002-1062-4808
artur.szutta@ug.edu.pl

On an intuition regarding the acquisition of moral virtue

Received: 28.09.2023 / Revised: 20.10.2023 / Accepted: 5.11.2023 / Available: 20.12.2023

Abstract:
This article concerns the issue of the moral enhancement of humans through technology. I propose a thought experiment that allows us to identify a new reason against implementing such enhancement. Achieving virtue through a path that involves one's own effort in making and implementing morally sound decisions deserves greater respect. It also allows us to acknowledge that we are (co)authors of who we become morally. This kind of self-creation seems to be an important part of a meaningful life, and artificial moral enhancement deprives us of it.

Keywords:
moral enhancement, virtue, intuition, meaning of life

How to cite:
Szutta, A. (2023). On an intuition regarding the acquisition of moral virtue [English translation]. *Laboratorium Mentis*, *1*(1), 75–85. https://doi.org/10.52097/lm.8157

The debate on moral enhancement

According to some authors, the challenges of the modern world, technological development in particular, require us to improve our moral condition. For instance, Julian Savulescu and Ingmar Persson (2012, 2013) argue that contemporary technology may enable immoral people to commit large-scale evil acts, such as destroying a multi-million city or even an entire nation using modern weapons (e.g., dirty atomic bombs, viruses, etc.).[1] The same technological advancement, as they observe, will one day allow us to enhance our moral nature, either at the motivational level (improving our capacity for empathy) or at the cognitive level. Improving our moral abilities through technology, such as genetic manipulation or pharmacology, will not only be an intriguing, perhaps tempting option but also be our moral duty, justified by the need to prevent a very likely great evil.

The idea of moral enhancement is the subject of numerous objections. For example, Allen Buchanan (2009) is concerned that morally enhanced individuals will have a higher moral status and, as a result, greater rights than ordinary human beings. Other authors, such as John Harris (2011) and Michael Hauskeller (2013), argue that as a result of moral enhancement, we will be so determined to do good by our improved emotions, desires, or other mechanisms occurring in our bodies that we will become automatons devoid of free will. Critics of moral improvement through technological interventions also point out the problem of establishing the shared standards by which we should improve people, given the strong disagreement in ethics about basic ethical principles (Schaefer, 2014). Among the objections, there is also the charge of unforeseen consequences of attempts at moral improvement through pharmacology or other techniques (Fabiano, 2018). Critical

[1] The idea of moral enhancement gathers more and more advocates (Specker et al., 2014)

voices have not gone unanswered by supporters of moral enhancement. It can be said that numerous advanced discussions are currently taking place (see Lavazza et al., 2019), in which both sides formulate further arguments, counterarguments, and counter-counterarguments.

In this article, I want to propose a thought experiment that offers a new perspective on the problem of moral enhancement. It allows us to identify a certain intuition that speaks against the implementation of this idea, an intuition that has not been pointed out (at least not directly) in the publications I am aware of.

Assumptions

Before I present the experiment itself, let us make certain assumptions about the technology of moral enhancement. Let us set aside any technical imperfections of current and future methods of human moral enhancement. Let us assume that the outcome of these procedures will be individuals who are indistinguishable in behaviour, consciousness, and motivation from people who would be the perfect embodiment of an Aristotelian virtuous person. Let us also assume that the discussed enhancement of people will be universally accessible; it will have no undesirable effects and will not deprive morally improved individuals of the ability to reflect and act on moral reasons or, more broadly, to make authentic moral decisions. Would we then consider that there are no reasons against realising the idea of moral enhancement?

The experiment

To answer this question, I propose conducting a thought experiment. Let us imagine four parallel worlds indistinguishable except for one detail. Each is inhabited by the same person, John, but in each of these worlds, John has a slightly different history. Accordingly, in each of the possible worlds *a*, *b*, *c*, *d*, there are John_a, John_b, John_c, and John_d. At a specific time measured simultaneously for all four worlds, let's say at

time t_n, all Johns are morally virtuous in the Aristotelian sense. They have a stable disposition to reliably recognise what is morally right in given circumstances and are motivated in the light of this recognition to effectively do what is good (Aristotle, ca. 300 B.C.E./2014, NE 1116a).

The only thing distinguishing all four men is how they acquired their virtues. $John_a$ became virtuous in the traditional way, initially (when he was still a small child) through imitating other people, then through persistent training in making the right decisions, overcoming his weaknesses and gradually building up the right dispositions within himself, perhaps with more than a few failures (especially at the beginning of his moral development) along the way.

$John_b$, at a certain time, $t_{n\text{-}m}$, was chosen without his knowledge or consent as the subject of a secret government program, the goal of which was to create and test a moral enhancement project using a certain technology, the details of which I will not specify here to avoid unnecessarily criticisms that would focus on its weaknesses or shortcomings (according to the experiment's assumption, it would be a safe technology).[2]

The case of $John_c$ differs from that of $John_b$ in that it was not the government that decided to use moral enhancement technology but $John_c$'s parents. They desired a morally virtuous child and were afraid they could not raise him properly using natural methods. $John_d$, on the other hand, perhaps frustrated by his own poor moral condition, voluntarily enrolled in the moral enhancement program and underwent the procedure with full awareness and consent.

---

2 It would be a natural reaction, while reading about $John_b$'s situation, to raise objections regarding the moral acceptability of the very methods of artificially implementing moral dispositions. I acknowledge that such objections could be justified and provide an important starting point for discussion. However, I want the reader to disregard them here because the goal of this experiment will not be the evaluation of the methods themselves but the very idea of artificial moral enhancement.

For this article, we can assume that at time t_n, all the Johns mentioned above are indistinguishable in terms of their behaviour, motivation, cognition, and emotions. By behavioural indistinguishability, I mean their observable actions from a third-person perspective (including activities of the brain and other relevant body parts) are identical. Motivational indistinguishability means they are guided by the same moral reasons in decision-making. Cognitive and emotional indistinguishability[3] implies that the world, all its objects, their characteristics, and relationships between them are cognitively perceived the same way by all the Johns, and they have dispositions to experience the same emotions in the same circumstances.

An intuitive assessment

Despite the indistinguishability described above, the different histories and ways of acquiring virtue seem to lead us to distinct evaluations of the four men. It is intuitive to claim (assuming that most readers will share this intuition) that we should admire $John_a$ more than the other three Johns.[4] This admiration should encompass not only the current abilities of $John_a$ but also the effort, perseverance, and willful commitment throughout his moral growth process. $John_a$'s current state is, to the greatest extent (compared to the other Johns), a result of his own merits. It seems that accounting for merit in the achievement of virtue should play a role in the moral assessment of the Johns.

3 I neither assume nor reject here that emotions are, by definition, non-cognitive. Such an assumption is not relevant to the purpose of the experiment presented in this paper.

4 I believe that, to some extent, $John_d$ deserves a slightly better moral evaluation and, to some extent, our admiration (though to a lesser degree than $John_a$) because he himself requested moral improvement.

Discussion

Now, does the above intuition have any implications for the moral enhancement debate? It seems that (assuming its validity) this intuition provides *prima facie* reasons against the implementation of artificial moral improvement. Still, the proponents of such enhancement could formulate two arguments against the abovementioned intuition.

First, the mere fact of intuition regarding John$_a$'s greater merit and admiration for him does not necessarily mean that he deserves a higher moral evaluation. This intuition may be a relic of the evolutionary process that is no longer useful as it does not take into account our technological development. At best, such an intuition could relate to the conditions in which people lived before the technology enabling human enhancement emerged. Evolutionary changes (including moral intuitions) are slow, while technological development is rapid. Hence, our mismatch of intuitions with the current civilisational situation, including the possibilities of moral improvement.[5]

Second, even if we ascribe some initial credibility to the intuition mentioned above, it might still have to yield to another intuition concerning the comparison of the likelihood of success in achieving virtue through traditional methods with the probability of attaining it through technology. From observation, we know that very rarely, if ever, people manage to achieve virtue through their own practice based on a series of autonomous decisions. Meanwhile, technology, at least in light of the assumptions adopted in the experiment, provides, if not complete certainty, a significantly higher probability than the traditional method of achieving virtue (or at least something indistinguishable from virtue).

5 An advocate of this argument, for example, is Peter Singer (2005). For more about the criticisms of the thesis on the credibility of moral intuitions, see Szutta, 2018.

Thus, when faced with a decision at the societal/political level, whether to introduce moral enhancement through technical means or stick with traditional educational methods, we should weigh both intuitions: on the one hand, the intuition concerning the lower moral status of artificially produced virtue (or something deceptively similar to virtue),[6] and on the other, the probability of success in implementing virtue by one method over the other. If we could, with a relatively high probability, conclude that, for example, in a society of one million, we would have a million (or even half a million) people *à la* $\text{John}_{b,\ c,\ or\ d}$, wouldn't that be a more desirable state of affairs (which we would be morally obliged to realise) than an equally large society with a few virtuous exceptions? Wouldn't this be a more desirable situation and its realisation our moral duty, even if, as a result of moral enhancement, individuals were deprived of the possibility of (full) merit and (full) admiration for their moral condition?[7]

---

6 I don't want to resolve here whether we can indeed describe the dispositions of $\text{Johns}_{b\text{-}d}$ as virtues, but it's worth considering the thesis that virtue is four-dimensional, meaning it encompasses not only the condition of the subject at time t_n but the entire process from t_1 (the moment in a person's existence when she may start slowly becoming virtuous) to t_n (the moment of acquiring virtue). Another component/feature of virtue, alongside its four-dimensionality, would be, in light of the experiment presented here, the self-determination of the subject. In other words, virtue would have to originate from within, not from outside the subject. Similar intuitions regarding the acquisition of virtue (or moral improvement in general) are held by Jason Erbel (2018), Ruben Herce (2019), and Tartaglia (2020, ch. 5).

7 The above argument can be presented in another form. We can imagine John_e, who consciously gave up on moral improvement in favour of the traditional form of moral development but still failed to achieve virtue or even come close to it. Wouldn't we consider that John_e deserves a negative moral evaluation for wasting time on unsuccessful attempts when he had an effective alternative at his disposal? In cases where we are certain that a person has no or very little chance of becoming morally virtuous or at least morally good at some minimal level, one could argue that we have a duty to morally improve them through technology. Otherwise, that person's life would be somewhat wasted. It's also worth adding that in the comparison described above, it's not only about effectiveness in terms

We can put forth the following arguments in response to the above criticisms. First, perhaps not necessarily our moral intuitions, especially concerning the way virtue is acquired, are unworthy of at least initial trust. Since I have written elsewhere about the credibility of moral intuitions in general (Szutta, 2018, Chapter 7) and there is already substantial literature on this topic,[8] I will limit myself to the argument for the credibility of a specific intuition regarding the significance of the way virtue is acquired. It does not necessarily have the character of an outdated mechanism (or a mental habit) that only applied to times when the time-consuming development of character through autonomous participation in appropriate practices was the only effective method of achieving virtue. Perhaps this intuition also relates to our understanding of a meaningful life.

Maybe an essential element of a meaningful life is, at least to some significant extent, to be the author of oneself, to create oneself through autonomous decisions. Let me propose another short experiment to support this interpretation of the defended intuition. Imagine that at the moment of your birth, your parents requested the implantation of a small chip in your head, which, by appropriately stimulating your brain, would motivate you to follow a specific life pattern planned by your parents. Let's assume that this plan involves having certain interests at school, choosing the right studies, selecting the right life partner, etc. Assume also that by making the "right" choices, you would appeal to the objective reasons you've recognised, but they would serve more as justifications for choices predetermined by your parents rather than

of achieving the desired outcome but also about efficiency measured in terms of time. When comparing traditional virtue acquisition with the method of artificial moral enhancement, we should also take into account that the traditional method is more time-consuming, and along the way to achieving virtue, there may be many moral failures, which, assuming we have an effective method of artificial enhancement, could be avoided.

8 An interesting example of the latest defences of moral intuitions is Bengson et al., 2020.

as causes or reasons determining your choices. If you were to discover such a chip and its function, wouldn't your life lose at least some important part of its value to you?

Thus, doesn't the admiration for human achievements simultaneously imply a kind of affirmation that human life contains an essential, meaningful element? And does this element not consist in the fact that what we achieve in life, we do so, at least to some extent, also through our own efforts and decisions?

In conclusion, let me propose an additional, auxiliary argument aimed only at those who accept the existence of a personal God, who created humans and desires their moral development. If God is good, omnipotent, and omniscient, and if it were good for us, He would have made us morally perfect from the very start, without the need for the uncertain and painful birth of our moral character, without regressions, and the risk of failure. If, despite His goodness, omniscience, and omnipotence, He did not create us as perfect, then perhaps He deemed there is a better alternative: that we have a part in creating ourselves, that we are co-authors of ourselves. Accepting God's existence, we have another reason to support the belief that the way we become morally good people matters because it matters whether we can at least partially call the outcome of our lives, who we become, our self-determination.

Conclusion

The thought experiment and the discussion above point to a rather neglected reason against introducing moral improvement through artificial intervention in the human body. By implementing the disposition for effective moral good through this means, we deprive individuals of the opportunity for autonomous self-determination. However, we should not consider this reason finally determining the impermissibility of artificial moral enhancement. It has a *prima facie* character. The final decision on the issue discussed here requires considerations of many other reasons, such as whether the price of giving up (or depriving) a

chance for self-creation is worth it to achieve a society in which the vast majority if not all, act morally well. This text should be considered a proposal of a reason to be included in the broader discussion of the permissibility and value of moral enhancement.[9]

Bibliography

Aristotle (2014). *Nicomachean Ethics*, transl. R. Crisp. Cambridge University Press. Original work published ca. 300 B.C.E.

Bengson, J., Cuneo, T., Shafer-Landau, R. (2020). Trusting moral intuitions. *Noûs*, *54*(4), 956–984. https://doi.org/10.1111/nous.12291

Buchanan, A. (2009). Moral status and human enhancement. *Philosophy & Public Affairs*, *37*(4), 346–381. https://www.jstor.org/stable/40468461

Eberl, J.T. (2018). Can prudence be enhanced? *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine*, *43*(5), 506–526. https://doi.org/10.1093/jmp/jhy021

Fabiano, J. (2018). *Probing the Risks of Moral Enhancement* [PhD thesis]. University of Oxford.

Harris, J. (2011). Moral enhancement and freedom. *Bioethics*, *25*(2), 102–111. https://doi.org/10.1111/j.1467-8519.2010.01854.x

Hauskeller, M. (2013). The "Little Alex" Problem. *The Philosophers' Magazine*, (62), 74–78. https://doi.org/10.5840/tpm20136299

Herce, R. (2019). Is human enhancement possible if it comes from the outside?. *Scientia et Fides*, *7*(2), 165–170.

Lavazza, A., Reichlin, M. (2019). Introduction: moral enhancement. *Topoi*, *38*(1), 1–5. https://doi.org/10.1007/s11245-019-09638-5

Persson, I., Savulescu, J. (2013). Getting moral enhancement right: the desirability of moral bioenhancement. *Bioethics*, *27*(3), 124–131. https://doi.org/10.1111/j.1467-8519.2011.01907.x

[9] For critical comments made to the earlier versions of this paper, I would like to express my gratitude to Krzesimir Cholewa, Michał Dominów, Paweł Homel, Jakub Nowicki, Stanisław Kamiński, Paweł Sikora, Natasza Szutta, Błażej Szymichowski, James Tartagila, and Emilia Wrońska.

Savulescu, J., Persson, I. (2012). Moral enhancement, freedom and the god machine. *The Monist*, *95*(3), 399–421. https://doi.org/10.5840/monist201295321

Schaefer, G., Schaefer, G.O. (2014). *Moral enhancement and moral disagreement* [PhD thesis]. Oxford University, UK.

Specker, J., Focquaert, F., Raus, K., Sterckx, S., Schermer, M. (2014). The ethical desirability of moral bioenhancement: a review of reasons. *BMC Medical Ethics*, *15*(1), 1–17. https://doi.org/10.1186/1472-6939-15-67

Singer, P. (2005). Ethics and intuitions. *The Journal of Ethics*, *9*(3/4), 331–352. http://www.jstor.org/stable/25115831

Szutta, A. (2018). *Intuicje moralne. O poznaniu dobra i zła* [*Moral intuitions. About the cognition of the good and evil*]. Lublin: Wydawnictwo Academicon. https://doi.org/10.52097/acapress.9788362475612

Tartaglia, J. (2020). *Philosophy in a Technological World: Gods and Titans*. Bloomsbury Publishing.