

Contents

Research Articles

Ondřej Beran: <i>Flattening the Curve of Moral Imagination</i>	174
Nora Grigore: <i>Susan Wolf on Supererogation and the Dark Side of Morality</i>	200
Andrzej Waleszczyński – Michał Obidziński – Julia Rejewska: <i>The Problem of Intention and the Evaluative Properties of Effects in the Knobe Effect</i>	219
Seong Soo Park: <i>The Whole-Part Dilemma: A Compositional Understanding of Plato's Theory of Forms</i>	246
Erhan Demircioğlu: <i>Conditional Uniqueness</i>	268
Harold Noonan: <i>The Personite Problem and the Stage-Theoretic Reply</i>	275

Book Review

Jaroslav Peregrin: Catarina Dutilh Novaes, <i>The Dialogical Roots of Deduction: Historical, Cognitive and Philosophical Perspectives on Reasoning</i>	283
--	-----

Flattening the Curve of Moral Imagination

Ondřej Beran*


Received: 14 September 2020 / Revised: 14 May 2021 / Accepted: 20 June 2021

Abstract: In this paper, I discuss some moral dilemmas related to the COVID-19 crisis and their framing (mainly) in the public debate. The key assumption to engage with is this: that we need primarily to take into account the long-term economic consequences of the proposed safety measures of social distancing. I argue that the long-term economic concerns, though legitimate, cannot suspend the irreducibly moral nature of the demand placed on the decision-makers by those who are vulnerable, at risk, or in need of medical treatment. This is discussed in relation to two points: 1) The political endeavour and rhetoric of “flattening the curve” is not necessarily short-sighted, but expresses the acknowledgment of a legitimate expectation placed on elected representatives. 2) Not being able to prevent harm (to those who are in real need, or otherwise vulnerable) may lead to a genuine moral distress, even if it is not clear whether it was in one’s, or anybody’s, powers to prevent the situation, or even if the best possible outcome has been otherwise reached. The second point may be understood as a part of the broader context of the established criticisms of utilitarianism.

Keywords: COVID-19 crisis, economic concerns, moral dilemmas, moral luck, remorse.

* University of Pardubice

 <https://orcid.org/0000-0003-2553-5872>

 Centre for Ethics as Study in Human Value, Department of Philosophy and Religious Studies, Faculty of Arts and Philosophy, University of Pardubice, Stavařov 97, Pardubice 532 10, Czech Republic.

 ondrej.beran@upce.cz

© The Author. Journal compilation © The Editorial Board, *Organon F*.



This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International Public License (CC BY-NC 4.0).

Introduction

The surge of COVID-19 in spring 2020 caught many countries unprepared. Or, more precisely, the character of its spread, along with not always transparently distributed information and not always smooth international coordination, made it practically impossible for most countries to be “fully” prepared.

During the first weeks of the outbreak in the European countries and the U.S., the measures taken aimed at the effect described by the phrase that has rapidly become popular: “flatten the curve.” These measures sought to slow down the increase in cases of COVID-19 so that the capacities of healthcare systems would not be overwhelmed.

There were various predictions of the clash between the expected progress of the epidemic and the real capacities of healthcare systems. From the outset, there were reported cases of healthcare facilities being overwhelmed. The reports also assumed that some patients may have died while not getting all the necessary treatment. Relatedly, medical authorities and healthcare workers needed to practise emergency triage, prioritising those patients who had better prospects of recovery.¹ Mostly, these were younger and less afflicted patients.

The underlying logic of this reasoning is straightforward: distributing medical capacities and material in such a way that would save as many lives as possible. At the same time, medical personnel could not fail to see that many who could not get access to ventilators or other medical material that was in scarce supply were in danger. The standard options of treatment would have increased their chances of recovery, though less than for those patients who were given priority. Although this practice of providing healthcare and making such far-reaching decisions in real time, under extremely difficult conditions, has been complex and far from straightforward,

¹ See, for example, Jason Horowitz, “Italy’s health care system groans under coronavirus—a warning to the world,” *The New York Times*, 12 March 2020, <https://www.nytimes.com/2020/03/12/world/europe/12italy-coronavirus-health-care.html>; or Sam Jones, “Spain: doctors struggle to cope as 514 die from coronavirus in a day,” *The Guardian*, 24 March 2020, <https://www.theguardian.com/world/2020/mar/24/spain-doctors-lack-protection-coronavirus-covid-19>.

it is often understood as a form of applying roughly utilitarian reasoning. Many have understood the triage practice during COVID-19 pandemic as a case in point. This practice purportedly illuminated the fact that the problems before which the pandemic was placing us centred round the principle of saving as many lives as possible.

Perhaps the most (in)famous philosophical reply to the pandemic was Giorgio Agamben's short critical point against the wave of societal restriction and social distancing regulations, which he views as an illegitimate form of "biopolitics."² However, philosophers reflected also the above aspect of the COVID-19 crisis. This interest is quite natural, as the triage practice points towards difficult moral dilemmas.

I will focus here on this latter angle of philosophical interest in the situation, in particular on related criticisms of the social distancing regulations, backed by different reasons than Agamben's. As the point of departure for my discussion, I would like to use H. Orri Stefánsson's (2020) particular reading of this medical dilemma. In section 1, I summarise the parts of Stefánsson's argument that are relevant for my discussion. I then raise some objections, in two directions. In section 2, I argue that some straw-man elements partly compromise Stefánsson's criticisms of what he takes to be common moral reasoning about the COVID-19 crisis. In section 3, I present a more general reflection on the crisis, beyond criticising closely Stefánsson's position only. I will strive to show that the crisis represents a different kind of moral problem, relating to issues of remorse and moral injury. Section 4 offers a few concluding remarks.

Stefánsson's paper is unusual in that it represents a philosophical articulation of sentiments and attitudes relatively common among "laypeople," including high-profile authorities and opinion-makers. However, as far as

² Agamben (2020); Castrillón and Marchevsky (2021) assembled an interesting critical discussion about this piece. Žižek (2020a, 75) succinctly points out that while we may rightly be suspicious about some forms of social control inherent to the pandemic regulations, this suspicion "does not make the reality of the threat disappear." Later on (Žižek 2020b, 28f), he notes that Agamben's criticisms offer little to distinguish themselves from the populist new Right. He argues that Agamben missed the chance to say anything about the new forms of inequality, the situation of workers or precariat, or about the current forms of capitalism.

I can see, distinctly *philosophical* articulations of this position are rare to meet. In itself, Stefánsson's argument represents a particular and perhaps somewhat crude version of the utilitarian reading of the pandemic. It is a version, not necessarily something any utilitarian would subscribe to. After all, it has turned out that various utilitarian analyses of the Covid-19 crisis lead to very varied recommendations. The two distinct utilitarian ideas that find a particular expression in Stefánsson's paper are the following: 1) We should be worrying about the long-term results of the adopted regulations. There is, or has to be, an objective way of calculating these results, such as applying the metric of QALY (Quality-Adjusted Life Years). These calculations of the maximisation of the overall good have an unmistakable economic dimension. 2) This calculation covers, more or less, the range of all the meaningful or legitimate moral worries regarding the pandemic. If we take any other kind of concerns, reaching beyond the need to identify and apply such an impartial principle, as *moral* concerns, it is a confusion.

These two points are the object of my critical focus, though not in a way neatly falling apart into separate sections. I will not be arguing straightforwardly against 1); I do not aim to present a refutation of utilitarianism here. My critical comments will concern rather some neglected difficulties relating to the identification of the good results. My truly central target is the tacit assumption 2).

Stefánsson's arguments are illuminating in how straightforward and clear-cut they are. They also represent a characteristic feature of the current debate.³ Their examination may thus bring us a relevant insight reaching

³ As suggested before, the above-referred "debate" is generally public, rather than specifically philosophical. Thus, for instance, two former governors of the Czech National Bank predicted that the losses of the domestic economy, caused by the protective measures, will be ten or more times higher than is the aggregate cost of the QALYs (within the Czech healthcare system) of the lives saved. See <https://archiv.ihned.cz/c1-66738020-byvali-sefove-cnb-tuma-a-hampl-nechame-v-zajmu-ochrany-zivota-umrit-ce-lou-ceskou-ekonomiku>. Such simulations are made worldwide; e.g. Amewu et al (2020), who voice similar concerns. Gans (2020, chapter 1) provides a critical review of many such accounts of the COVID-19 crisis. There are, however, philosophical voices close to Stefánsson's position in some respects, for instance, Savulescu et al. 2020, Singer and Plant 2020, or Williams et al. 2021. From a position close to mine, Gaita (2020)

beyond this particular individual analysis of the COVID-19 situation. At first, I stick closely to particular points made by Stefánsson, which makes parts of this text a polemic directed specifically against him. However, he voices arguments and intuitions that are not unique or eccentric. I believe that it makes the criticisms I raise relevant also beyond the context of the one particular paper.

1. Stefánsson's argument

In his paper “Three Mistakes in the Moral Reasoning About the Covid-19 Pandemic,” Stefánsson argues that moral reasoning about the current crisis is burdened by several problems. He notes that he is not criticising the motivations of the actual measures taken, for it is difficult to gain their overview, but rather the fallacies to which the common moral framing of the situation is prone. He explores what he takes to be the main three problems. First, he sees a fallacy in the idea that difficult choices and trade-offs in decision-making both individual and public can be avoided if the right kind of precautions are taken in time. The dilemmas faced by overloaded medical facilities and healthcare workers serve as the example motivating the flawed reasoning. Second, he identifies the mistaken temptation to bypass democratic mechanisms in making the important decisions and to delegate these to experts. Third, he warns against an incoherent application of the precautionary principle. He suspects that measures taken in order to stop the spread of the coronavirus may have such drastic effects for societies and economies that the results would be even worse than those caused by the pandemic. I will focus here mainly on the first of Stefánsson's worries. I agree with much of what he says about the second, and his analysis of the third is a bit sketchy and partly repeats the points he makes in the first section.

offers a philosophical reply to voices in the public debate in Australia. Gaita also observes that the kind of utilitarianism that we find only rarely in a clear form in philosophy has a special attraction for non-philosophers, when they learn about it as a theory. As he says, “[t]here are hardly any strict consequentialists, but many people are vulnerable to believing that they should be.”

Stefánsson opens his discussion by rephrasing what he takes to be the principal motivation driving the endeavours to flatten the curve: “a commonly stated reason for why we should spread out the burden on the health-care system over time, namely, that it would allow us to avoid making hard trade-offs” (Stefánsson 2020, 4). He is critical of the persistent temptation to picture these hard trade-offs as something that can be avoided. People tend to think, he says, that if matters are organised better, we won’t have to choose whom to treat and whom not. No scarcities of human or material resources in healthcare would then have to occur. He is sceptical about this. More importantly, he identifies some of the public pundits, commenting indignantly on the perceived need to “choose whom to save and whom not to save,” as inappropriately moralistic. Speaking ostentatiously about this misconstrued moral problem only obscures the urgency of the real problem, as he sees it.

Stefánsson’s argument does not engage directly with the question of how to perceive the unfortunate choices faced by healthcare workers. He wants us, instead, to stop glossing over the inevitability of making such choices. For another, and graver, instance of such a choice is inherent to social distancing and other safety measures. These measures aim at 1) protecting lives now, which would otherwise be lost to the disease, but 2) taking these measures will cause a harmful economic impact in the future. Social distancing measures will necessarily slow down the economy, which will in turn result in a worsened quality of life and the deaths of even more people.⁴ Stefánsson illustrates this risk by citing statistics about the various degrading social effects caused by the last economic depression. As he argues, we can anticipate an analogous development in the wake of the coronavirus crisis.

Stefánsson criticises the hypocrisy of people indignant about the need to make a choice between the lives of COVID-19 patients, while overlooking that other lives will be taken in the long run. The problem responsible for this mistaken reasoning is, in his view, that the latter victims are at present unidentified and unspecifiable (cf. also Savulescu et al. 2020, or Singer and Plant 2020). However, the need to think over a considerably longer time

⁴ Singer and Plant (2020), or Savulescu et al. (2020, 626) expressed similar concerns.

span and a considerably more ramified interconnection of social phenomena makes the task of taking the current health safety precautions *responsibly* (with respect to the future) hard. In fact, it is harder than the supposedly hard trade-offs faced by healthcare workers distributing their overstretched capacities. For healthcare workers have, as Stefánsson concludes in this argument, at least some criteria helping them to navigate such decisions: for instance, the metric of Quality-Adjusted Life Years, or QALYs (cf. MacKillop and Sheard 2018).⁵

2. Avoiding the “hard trade-offs”

The primary worry underlying section 1 of Stefánsson’s discussion (on “hard trade-offs”) is how the *economy* will be affected. The decision to take measures aimed at flattening the curve is, for him, the “decision to almost completely shut down [national] economies and force people to stay indoors.” He points out that this decision was not even supposed to reduce the overall number of infections, but only to spread it over time so that *at no point* would healthcare systems become overloaded.

None of the few sources referenced by Stefánsson talks, however, explicitly about this ideal scenario. One of these, Spektor (2020), characterises the aim of the curve-flattening measures as follows: “A slower infection rate means a *less stressed* health care system, *fewer* hospital visits on any given day and *fewer sick people being turned away*” (my emphasis). This is a realistic suggestion of reducing the burden, rather than a way of securing the outcome that “all can get the treatment they need,” as Stefánsson puts it (Stefánsson 2020, 5).⁶

⁵ Stefánsson characterises QALYs as a “well worked-out framework,” though “not uncontroversial.” For general criticisms of the QALY metrics, see e.g. La Puma and Lawlor 1990, or Marra et al 2007. In reply to the way in which QALYs were alluded to in Australian public discussion about COVID-19 (“we must ‘apply scientific rigour’ to the questions that ‘everyone is skirting’”), Gaita (2020) stresses the risk that this metric represents for instance in the case of disabled people.

⁶ An additional factor contributing to the overload is the insufficient funding of many countries’ public healthcare systems. The present situation thus calls for making amends in this respect. Notably, Stefánsson classifies the present state of

For sure, only a few government officials would openly admit to the public that they expect the healthcare system to overload and people to be left to sicken or die without the full extent of needed treatment. Most of the statements they issued, whether specific or vague, thus suggested that everybody *would* get the necessary treatment. In that respect, Stefánsson's critique is perhaps right. These statements sometimes indeed evoked the "thought that we can somehow *avoid* making hard trade-offs." The "hard trade-offs" were *not* avoided, and whether they could be, in the chaos of the first weeks and months of COVID-19, will remain unclear.

I am less sure, however, whether any proclamation presenting the endeavour to avoid these trade-offs as worthwhile is a mistake in moral reasoning. In fact, the rhetoric adopted by most⁷ governments was acknowledging the *legitimacy* of the expectation that they would endeavour to fight health system overload. This is a part of their responsibility to the public. A politician can hardly openly act as if indifference, or even placidity, about letting such trade-offs happen is consistent with how we understand the political representation's answerability to the public.

If political systems cannot operate on the *expressed*, if symbolic, assumptions of such goodwill on the part of politicians and trust in this goodwill on the part of voters, they will be affected in ways difficult to predict. Obviously, under particular difficult circumstances politicians may fail to represent their citizens' interests in obtaining urgently needed healthcare. Sometimes they indeed fail due to not trying hard enough, or even due to laziness or corruption. However, they can hardly be thought to represent the citizens' interests by *subscribing* to any *principle* that says it is perfectly all right not to care about representing the interests of *some* citizens.

Gaita (2004, 23ff) makes a similar point about the role of *lying* in politics. He argues that while politicians must not aspire to be saints, moral

healthcare systems in developed countries as already "ever-expansive" and therefore effectively unaffordable.

⁷ A notable exception may be the former PM of Australia, Tony Abbott, who suggested that it might be better to "let nature take its course" ("'Assess value of life' of elderly coronavirus patients when reintroducing lockdowns, urges Tony Abbott," *The Independent*, 2 September 2020, <https://www.independent.co.uk/news/uk/politics/tony-abbott-coronavirus-australia-covid-old-cases-deaths-a9700881.html>).

values, shaped also by the concern for what saints represent in our culture, impinge on politics. Thus, though such beliefs, that no person should be treated only as a means, but always as an end in itself, “are problematic in politics, (...) at crucial points they inform it” (Gaita 2004, 25). Similarly, it may be an impossible task for politicians in the time of COVID-19 to organise the provision of healthcare in such a way that every single patient is treated as an end in itself. Yet, they must not start acting as if the experienced impossibility itself renders such concerns irrelevant, dismissed. Placing long-term economic concerns first amounts to reducing some individuals, here and now, to the *means* to other ends.

(I do not aim here at the disentanglement of the structures of trust, important for the very understanding of politics, from ubiquitous political marketing. I assume that particular cases of political decisions typically represent a complicated and inseparable mixture of both, which does not mean it is the same thing, though.)

Stefánsson also quotes (Stefánsson 2020, 5) from a radio debate in which the host asked a member of the Swedish National Council on Medical Ethics whether it is possible that Sweden will face this kind of medical dilemma. Stefánsson does not record the doctor’s answer and focuses instead on the tacit assumption he saw in the host’s question: that we *can* manage to avoid these hard trade-offs. He seems to dismiss the legitimacy of asking such a question in the time of a pandemic. It is reasonable to ask the question, for there are many different ways of elaborating on the issue. Stefánsson’s argument *is* one of them. He himself would probably appreciate an opportunity to be asked the question and to present, in reply, his concerns and worries to the wider public. Moreover, the host’s question does not necessarily make the underlying assumption that Stefánsson reads into it. It is possible simply to ask whether a situation that raises legitimate moral worries and that seems looming is likely to happen, by expert estimate.

Apart from the literally taken idea that we can avoid difficult medical dilemmas altogether if only we flatten the curve appropriately, the public debate, according to Stefánsson, exhibits another flaw in the presented moral framing of the COVID-19 crisis. He argues that the truly hard trade-offs are not, as is usually assumed, between differently afflicted patients here and now. Instead, he emphasises the trade-offs between the

present estimable victims of the pandemic and the future victims of the economic depression that the health safety measures will cause.

He phrases his polemic in terms of rehabilitating the standing of the future victims in a debate that overlooks their relevance. However, he does so in a way that suggests preferences of his own. Thus, he says that while some lives “will *or might* be saved” by the present measures, other lives simply “*will* be affected by the economic depression” (Stefánsson 2020, 6, my emphasis). The depression, by the way, was not only predicted but “already starting” (in Spring 2020). He also presents this prediction of the long-lasting and intergenerational effects of the depression as a *plain fact*. The prediction relies on an extrapolation of findings about past socio-economic relationships into, presumably, an inevitable future.

This bleak deterministic view apparently presumes that human societies cannot learn from past crises or react to repeated difficulties in ways that differ from the previous cases and prevent or mitigate more harms. It allows Stefánsson simply to measure, by the same scale, the prospects of people likely to die, here and now, if they don’t get adequate treatment, and the prospects of people not yet born. The latter are in the same sense and with the same probability likely to suffer from the results of the present depression. Apart from his *certainty* of future victims, he also makes the equation between victims affected directly and *intentionally* and victims affected indirectly, in consequence of another action. Stefánsson phrases this difference as the one between victims known (identified) and unknown (unidentified), claiming that the principal reason for the apparent preference for the former is simply that they are known. However, he continues, “it should make no moral difference, all else being equal, whether a person is identifiable or not” (Stefánsson 2020, 7).

In these considerations, there seems no room left for taking into account the complex phenomena discussed under the heading of “double effect.” As, for instance, Anscombe argues in her classic paper “War and Murder” (1981, 58f), there are morally relevant differences between negative effects directly and intentionally caused and those that are foreseeable as a further effect of one’s action, which is, however, led by a different intention. If foreseen consequences are just as relevant as what one intends to do, here and now, then there is nothing, Anscombe argues, that would be morally prohibited as

simply wrong as such. Not even murder is. Everything is subject to a possible requalification, measured by its potentially graver consequences in the future.

Gaita (2004, 55ff) presents an argument similar to that of Anscombe, with *torture* as the focal case. He claims that when sometimes one needs to commit evil to avoid greater evil, it is important to retain the sense that even the lesser evil is still evil, often grave. On the other hand, some utilitarian arguments tend to assume that what is necessary cannot be evil. This relegates any remaining sense of worry to the merely *psychological*, rather than the moral. Not only has this trickery of rational arguments (as Gaita sees it) managed to re-establish torture as a legitimate topic for public debate. It has also rendered it impossible to distinguish between the rational dispelling of prejudices and the moral corruption of losing from sight why something used to be morally unthinkable. This is, however, not just a local flaw of moral reasoning, as Anscombe pictures it, but—in Gaita’s view—an established conception in its own right, taking morality as “an adaptable set of rules and principles that serve a purpose” (Gaita 2004, 58).

How does this illuminate Stefánsson’s classification of “hard trade-offs” and their avoidance? It may seem horrible deliberately to allow for such a situation in society, which would entail the overload of the healthcare system and the need to leave some patients without treatment, or to let elderly people die in nursing homes. However, this assessment could never hold absolutely. It would always be an initial, tentative assessment, awaiting its possible requalification by other considerations. For Stefánsson, the true task of difficult moral reasoning seems to consist in tracing and considering options of this requalification.

There is a hint of sad irony about his call for equality between victims, though. The burden of the pandemic already lies more heavily on more vulnerable population groups. The elderly, the already ill, or poor people with worse access to healthcare and/or riskier employment situations suffer more gravely. Compared to the identified vs. unidentified distinction, Stefánsson seems to disregard this latter kind of difference between the various kinds of victims of the pandemic.⁸

⁸ Reid (2020, 526f) suggests that issues of increased social or racial injustice are a blind spot in phrasing the pandemic counterstrategy in terms of maximising medical outcomes.

As we indicated, Stefánsson takes almost for granted the predicted economic consequences of lockdown and social distancing. However, as the pandemic was progressing, new evidence kept emerging. Not only the unregulated progress of the pandemic, but also insufficiently regulated progress proves to have more drastic consequences for societies and economies than strict lockdowns.⁹ The temporary gap in industrial production, travel, etc. also resulted in (sadly, also temporary) improved levels of air and water pollution. The COVID-19 crisis motivates also more long-term considerations of restructuring economies towards a greener and more sustainable shape. In the long run, the pandemic thus may also have positive consequences for the economy, which does not enter Stefánsson's discussion. Nor does he take into account our current lack of understanding of the disease and its effects. It may turn out that those who have contracted it but survived will suffer some permanent health effects. These would again represent a factor influencing the future load on healthcare systems and, by that, on the economy. The long-term damage done to the texture of society—high numbers of healthcare workers quitting their jobs, the eroded trust of citizens in the competence and good will of their governments—needs to enter our considerations as well. Stefánsson's ambition to present a more complex reply to naive reflections thus appears itself insufficiently complex.

Oddly enough, Stefánsson also refuses to see the particular character of the situation of represented by the pandemic. Its impact on a society is

⁹ Horton (2020) presents an overview of different strategies implemented by various countries, concluding that the more hesitant they were about applying strict social distancing measures, the graver were the consequences. Even mitigation ("flattening the curve") did not prove to be an efficient enough strategy. Analogously, Correia et al (2020) argue that, learning from the case of 1918 flu pandemic, there is a false dichotomy: by saving lives we are saving the economy, while the most disruptive factor for the economy is the epidemic itself. Even economic analyses presented in rather technical terms of "the value of a statistical life" suggest that "extreme measures are warranted" (USC economists Mireille Jacobson and Tom Chang for *STATNews*, 18 March 2020, <https://www.statnews.com/2020/03/18/economic-rationale-strong-action-now-against-coronavirus/>). This overview shows that while it is legitimate to worry about long-term consequences, predictions of consequences vary. Correia's argument, practically a counterargument against Stefánsson, is utilitarian, too.

overwhelming in a way similar to natural disasters. When an earthquake or a hurricane hits a country, the hospitals, firefighters, army, police, and other institutions simply do all they can to save all the people they can. Nobody really asks whether all these expenses might not cause even worse (economic) damage in the future. The reason is not that such calculations would not be possible, but that they are misplaced. That many people do not consider them misplaced in the COVID-19 case is a peculiar feature that the pandemic seems to share with the climate crisis. Certainly, we cannot overload the analogy between natural disasters and the pandemic. The latter is a long-term phenomenon. Short-term calamities usually provoke, after the initial shock, the spirit of solidarity and volunteering, but this drive naturally fades with time and cannot sustain the burden of a long-term hardship by itself. Let us not forget, though, that the ideas about economic caution were accompanying the pandemic from the very beginning, when it was not altogether sure, how long COVID-19 would remain here.

Overlooking the reasons why concerns such as Stefánsson's may sometimes be misplaced has to do with thinking about the nature of dilemmas in medicine in one-dimensional terms. I will discuss this in more detail in the following section.

3. Moral dilemmas and remorse

In the previous section, I tried to show that, at some points, Stefánsson seems to attack straw men. Here, I would like to look a bit more closely at one of his assumptions, which is, I believe, characteristic of a more general problematic tendency of reflecting on medico-ethical issues. Stefánsson keeps repeating that the alleged motivation for flattening the curve is the ambition to prevent the hard trade-offs in healthcare altogether. He attributes this ambition to some shortsighted, superficial moralism. Instead, he presents the true moral concern as proceeding in, as it were, organisational terms: it would be bad to get into the situation where healthcare workers would have to make the choice between patients, *if under different and more cautious arrangements it could have been avoided*, without causing more damage elsewhere. But, as it probably cannot (as he suggests), any further moral concerns implode.

What we have here is a rather familiar approach. The choices we make in relation to COVID-19 (just as elsewhere) and their underlying concerns centre round the ambition to identify and bring about the best possible scenario. If the decision-makers can reach such a scenario, they would have “clean hands” and no reason for regret or remorse. The reason is that the objectively best outcome simply is *the good outcome*, and it is unintelligible to question, criticise, or regret anything about *the good outcome*. Then, the only moral worry would be to consider whether government strategies have opted for the good outcome. For those who think of the crisis in terms similar to Stefánsson, the governments *have not*, prioritising the shorter-term effects of social distancing.

Some of my concerns presented in the previous section relate to my doubts as to whether Stefánsson identifies correctly the best aggregate result. Here, I will be more interested in the assumption (not only his) that the best aggregate result is the good outcome in the sense that it rules out intelligible remorse. If this logic held, nobody would need to blame themselves, if the need to choose between patients really has been unavoidable. Nobody would need to blame themselves, even if they *caused* a higher frequency of such situations, if only it was to prevent objectively predictable worse consequences in the future.

But consider the following: when one gets into a situation where one has to make such a choice, it is understandable that one feels blame for making *any* available decision. She feels the blame simply by virtue of *having to* make this decision. Even if such an overloaded healthcare worker has merely done what most other of her colleagues probably do, too, and for relevant reasons (triaging in favour of patients with better prospects), this does not make the self-blame unintelligible.¹⁰ If the need has been, in better-handled circumstances, avoidable or less urgent, it adds a further shade of outward-oriented anger or bitterness to one’s feeling of remorse. But it does not remove the remorse just because it is not the person oneself who was primarily responsible. Gaita (2006, 43ff) presents the analysis of an analogous

¹⁰ Cf. the interview with Cynda Rushton (Professor of Nursing Ethics at Johns Hopkins University) on the moral distress endemic among nurses during the COVID-19 crisis; *The Hub*, 2 April 2020, <https://hub.jhu.edu/2020/04/06/covid-nursing-cynda-rushton-qa/>.

kind of remorse, using the example of a Dutch woman during wartime who had to refuse shelter to Jewish fugitives (who were eventually caught and killed), in order to not threaten the anti-Nazi resistance plans in which she was involved. Her take on her own actions, as reported by Gaita, is remarkable: it made her hate Hitler even more because he had made her a murderer. We can understand this as a case of what was later called “moral injury”: a transformation of the person that makes her, though under the pressure of circumstances, incapable of imagining herself as a morally good person (cf. Wiinikka-Lydon 2019, 36f, 155f). Cases of moral injury show that the relationship between a tragic concurrence of events that one could not really influence and blame and remorse is very complex. Only an impoverished moral reflection would content itself with a picture of human life in which there is no room for tragedy or bad moral luck (cf. Williams 1981).

The characterisation of the moral dilemmas of COVID-19 crisis *exclusively* in organisational terms relies on neglecting an important underlying distinction. One thing is the practical, implicit need to practise triage in the real time of treatment. Under the extreme circumstances of the COVID-19 crisis, this involves treating some patients less than fully and appropriately. Another thing would be a moral *principle* stating that this is a right thing to do, as a *rule*, in order to meet the objective purpose of healthcare and medical ethics.

Applying widely the latter kind of approach seems a noteworthy aspect of *some* forms of utilitarian thinking. Undoubtedly, medical ethics benefits from identifying widely applicable general principles and procedures, which aim at maximising the number of surviving and recovered people. However, that does not mean that this is *all* that moral reflection needs to take into account in cases of medical dilemmas. Suggesting that the guilt and regret that healthcare workers experience are not moral, but psychological (neurotic) concerns is a serious misrepresentation. The purpose of finding the applicable principles and procedures is not to *dismiss* emotional responses of the moral kind to particular cases as irrelevant.¹¹

¹¹ Utilitarian framing of medical issues sometimes gravitates towards this view. Savulescu et al (2020, 626) characterise utilitarian recommendations related to COVID-19 as beneficial in that the position from which they appear counterintuitive

However, focusing on the level of a universally applicable principle and reflecting on the pandemic only in organisational terms flattens our moral imagination in certain respects. The organisational approach requires having a metric that will allow us to make far-reaching comparisons between people, based on an empirically measurable quality. Hence the entry of QALYs. This metric allows us to assess objectively, from the *impersonal* or *third-person* standpoint, measures targeting differently various group of people, *other* people.¹² The deepest problem lurks, expectably, in the claim of the *empirical measurability* of the “quality of life.” As Gaita (2020) argues, “the quality of life” is rather a first-person expression of the individual’s insight. Certainly, healthcare workers have had and will continue to have to practice triage, but the principle of who yields to whom differs in how it sounds, depending on who is voicing it. It makes a big difference whether it is the person herself, who consents not to be put on a ventilator,

and *which they make possible to avoid* consists in “psychological biases,” “heuristics,” emotion, or intuition. An interesting example from the Czech debate about COVID-19 is the expert overview and recommendation written by a team of ethicists and legal theorists (Černý et al. 2020). The paper contains many valuable insights and information, and for natural reasons it confines itself to the highly needed identification of the appropriate principles of the allocation of scarce resources. Yet, the rationale for this endeavour does not concern only action guidance itself. The authors also perceive the importance of being capable to show and justify that the physician’s decision “is not random, can be rationally understood and analysed, is transparent” (Černý et al. 2020, 6). As they say, this kind of transparency “bolsters, rather than undermines the trust of the society” (Černý et al. 2020, 8). Underlying is the worry that healthcare practice guided by anything else than such general principles would be “random” in an unacceptable manner, or at least that the public would suspect that. I am not sure to what extent this is true, or inevitable.

¹² From this perspective, Savulescu and Campbell (2020) suggest selective lockdown of the elderly, saying that “the benefits to others are so significant as to outweigh the loss of liberty.” Lawrence and Harris (2020) criticise their proposal, pointing out the special kind of vulnerability of the elderly who are likely to suffer in ways incomparable to younger age groups. They summarise their critique by saying that “[e]quality is not about equal misery but about giving equal concern, respect and protection to all.”

whether she is phrasing it as an observation about her particular case only, and so forth.¹³

Such and similar worries about the kind of reasoning that motivates the application of QALYs point towards the benefits of rethinking carefully the standing of the far-reaching, general principles. In their critical assessment of the QALY measure, La Puma and Lawlor (1990, 2920) make the following observation:

While utilitarianism may be an acceptable ethical theory with which to make health policy at the macro level, at present, clinical practice is not primarily conducted to benefit society as a whole, the public interest, or the common good. The physician's primary duty is to meet the patient's medical needs as they together find them, the physician with technical knowledge and expertise and the patient with his or her personal history and values. Conserving society's resources is secondary or tertiary; if such conservation is brought about by considering some patients expendable or by serving opposing masters of patient and society, the seemingly imminent role of public agent must be acknowledged, appealed, and refuted.

This sheds some light on the mischaracterisation of the situated healthcare practice under stress as a matter of a rule. A macro-level rule, relevant for policy-making, is necessarily a part of the system of many counterbalanced macro-level rules of policy-making. Economic criticisms legitimately deal with

¹³ Gaita's commentary goes as follows:

Were I, now 74 years old, in a hospital and told that I could not be put on a ventilator because it had to go to a younger person, I would consent to it. I would not think of this as "above and beyond the call of duty." For me this is ethically a no-brainer, which does not mean that I believe that anyone in a similar situation should think as I do, including the young person who would get the ventilator. Certainly, I would not respond graciously if they said, "Good on you, old man. You've made the right decision, impersonally considered. You've done your civic duty in this time of critically scarce resources." If they were to add that just by looking at me they could tell that my time-quality rating must be low, I would snatch the ventilator from them.

this level.¹⁴ However, any picture of healthcare, provided by particular doctors, nurses, and other medical personnel to particular patients, as practice either *following or failing to follow this rule* misses something. If one is motivated, in her treating of individuals as individuals, by concerns inherently directed *not* to individuals, it compromises the resulting attitude. Healthcare workers were properly worried about their capacities to treat their patients appropriately, including ensuring their own safety, which was a key factor in this consideration. Healthcare workers worrying about whether the extent of care they provided to their patients was not excessive and as such detrimental to the public economy must have been rather rare during the COVID-19 crisis. Stefánsson's discussion relies on 1) presenting these two worries as fundamentally of the same kind, which would thereby allow 2) to compare their relative significance, and which would then allow him 3) to proclaim the latter as graver.

Some commentators on the COVID-19 crisis considered ideas of the kind of 3) as outrageous;¹⁵ in a sense, I agree. However, the original confusion may

¹⁴ Utilitarianism may be the most common approach at the macro-level of reasoning about resource allocation, but it is not without alternatives. Perhaps the most important competitors are the various forms of egalitarianism, such as that of Daniels (2001). Reid (2020) questions the assumption that applying utilitarian principles in the case of COVID-19 pandemic would even represent the current medico-ethical consensus.

¹⁵ Not only philosophers or religious thinkers, but also economists. In his essay "The Dismal Kingdom" (*Foreign Affairs*, March/April 2020), the Nobel prize-winning economist Paul Romer deplores the ubiquitous reliance on economists as the principle decision-making source in matters of policy. He agrees that we cannot afford to "kill the economy" altogether (see e.g. his and Alan Garber's opinion article for *The New York Times*, 23 March 2020, <https://www.nytimes.com/2020/03/23/opinion/coronavirus-depression.html>). Yet, in "The Dismal Kingdom" he observes that

[u]nfortunately, asking economists to set a value for human life obscured the fundamental distinction between the two questions that feed into every policy decision. One is empirical: What will happen if the government adopts this policy? The other is normative: Should the government adopt it? Economists can use evidence and logic to answer the first question. But there is no factual or logical argument that can answer the second one.

His conclusion is that

lie in 1). That healthcare workers worried predominantly about providing treatment to their particular patients was not a sign of their having compared the two worries with a different result than (some) economists. They simply refused to acknowledge the worry about the future prospects of the country's economy as their own, inherent to their work as they needed to do it, and rightly so (cf. a similar argument made by Cowley [2008, 82ff]).

Once we have removed this worry from the picture, it turns out to be natural to rephrase the framing of the concern, as suggested at the end of the first paragraph of this section. Now, it would simply proceed in these terms: *whether or not* the need to make the choice between patients could be avoided (who knows), getting to that situation is simply bad as such. When the public was asking themselves or experts and politicians the questions about health care system overload, a particular feeling or sense was underlying these questions. It was, of course, the feeling that we need to do whatever we can to prevent as many instances of this situation as possible. The driving ambition was not to attain the objectively attainable minimum number of such situations (weighed against considerations of economic nature), because the moral problem would then disappear. For that would mean to overlook that the moral problem simply does not disappear no matter what. If a healthcare worker perceives the provision of treatment as a moral demand, following simply from the condition of the patients in need, the onerous sense of failing the demand does not disappear just because it is unclear whether it was in one's powers at all to avoid the situation. For sure, economics and economic relations *do* contribute significantly to the constitution of our moral and political relationships. We could never even understand our moral dilemmas, if we ignored how they their political and economic setting situated and shaped them. This is, however, a move of understanding, not of reduction.

The vocabulary itself that Stefánsson is using illuminates the risks of analysing our moral dilemmas in an overly reductionist manner, as fully

[n]o economist has a privileged insight into questions of right and wrong, and none deserves a special say in fundamental decisions about how society should operate. Economists who argue otherwise and exert undue influence in public debates about right and wrong should be exposed for what they are: frauds.

See <https://www.foreignaffairs.com/reviews/review-essay/2020-02-11/dismal-kingdom>.

exhausted by the description of their economic framework. Though Stefáns-son is concerned with *moral* reasoning, he avoids characterising the situation in which healthcare workers find themselves as a *dilemma*. For him, it is always a *trade-off*. But while there are moral dilemmas, I am not sure what a moral trade-off would be. A moral dilemma is a situation in which it simply may not be possible to avoid doing harm whatever one does (cf. Williams 1965; or Phillips 1979). A trade-off is a confrontation of inputs that need to be settled by means of a calculation. *Trade-offs* are *hard* in the manner in which complicated mathematical calculations are hard. (Moral) *dilemmas* are *difficult* in a different sense. If we embrace the vocabulary of trade-offs, it may prevent us from seeing the moral possibility of “inescapable wrongness,” in Bernard Williams’ words, as relevant for understanding the situation as a dilemma.

The moral concern of healthcare workers reflects the latter kind of difficulty: the need itself to decide whom one will not help to the full extent required by their condition. In this situation, one cannot help having qualms about whatever option one sees as available. These qualms do not depend on considerations of whether one has reached the threshold of inevitability. The overloaded healthcare workers did what they could under unimaginably difficult circumstances. Nobody, unless out of their mind, could think of suing them.¹⁶ Similarly, an army officer may need to give orders such that would result in the death of some of his or her troops, in order to secure a strategically important advantage (perhaps saving the lives of many more soldiers, or civilians). However, morality does not coincide with legal invulnerability or strategic necessity. It is perfectly intelligible that army officers who did the best they could under the circumstances still have moral worries about their decision, just as the healthcare workers. It has been among soldiers that cases of moral injury have been studied most frequently. Perhaps using this conceptual lens to understand the situation of healthcare workers will be helpful, too.

No third party is thereby granted the right to morally judge and condemn healthcare workers for failing to do what they could not avoid failing to do.

¹⁶ There are, instead, cases of bereaved citizens taking legal action against government authorities. See <https://www.theguardian.com/society/2020/jun/03/lost-father-covid-19-legal-action-against-uk-government>.

However, consider an attempt to placate an angry and remorseful traumatised medic by saying, “Come on, under the circumstances, nobody could sue you for the neglect of your professional duty.” Such a consolation amounts to an affront. It does not do justice to the fact that the medic understands what happened, and what she did, in terms that can be and surely often are irreducibly moral. The “bad moral luck” angle, which I believe is indispensable for appreciating properly the situation of the medic, doesn’t point towards a condemnation. It rather points towards pity, or abstaining from judgement by a third party (cf. Browne’s [1992] discussion of moral luck).

I think that important reasons for striving to “flatten the curve” and easing as much as possible the burden on healthcare systems lie somewhere here. These endeavours rely on the intuition that “hard trade-offs” are a bad thing to happen. We cling to this intuition even when we do not see whether there can be a viable plan for avoiding them altogether. And I hope that the principal motivation for the flattening endeavours and other counter-pandemic measures on the part of our representatives and institutions was not to maximise the aggregate value *in order to* clear themselves of possible blame. The underlying intuition may have been simpler: it is not right to let people die, even when you are in such a situation that your real capacities are limited and you can only save so many people. The endeavour to avoid ending up in a situation of “hard trade-offs” is thus an expression of an important intuition. Even if you manage to distribute healthcare resources so as to objectively minimise the number of people without the full necessary treatment, having to do this—having to fail *anybody’s* need—is bad enough. If the social contract between citizens and their states is taken seriously, the political representation cannot act or speak as if the fact that some citizens were not saved under the circumstances where it was not clear whether they could be saved nullifies the state’s commitment to represent the interests of these citizens.

Of course, there may be hypocrisy in the rhetoric of the politicians’ claiming that they would never let a single case of this kind happen. But what if they subscribed in a cavalier manner to the full legitimacy of letting it happen, when it’s unclear that it could be fully avoided? Adopting such an alternative rhetoric would represent a deep, worrying deficiency in moral reasoning. This might cause a broader damage to the society. Without necessarily

calling for the kind of welfare state that is looking after all the citizens' needs, the state cannot afford to become a body of representatives who are not really representatives because they do not care at all. Incompetence can erode the citizens' trust significantly, too, but indifference cuts even deeper, I believe.

The elected representatives' and the states' political responsibility is clearly not of the same kind as moral responsibility between and towards particular *individuals*. It is rather a complex mixture of responsibilities towards individuals *qua* members of particular groups, towards institutions, towards the "nation," or simply to the future. Politicians also carry the additional burden of the unrewarding, but immensely important task to justify their responsibilities in a way that will not alienate significant parts of the public.

4. Concluding remarks

In I am not sure what the best reaction to the COVID-19 outbreak was. Others have a better insight into this immensely difficult and sensitive topic. First, we would need some clarity as to what we mean by "the best." In some readings of "the best," perhaps Stefánsson's is the right suggestion: to gauge and regulate safety measures by their predicted future economic impact. My worry is that the reasons for this suggestion may not be the kind of reasons we would like to rely on if "the best" relates primarily to what is "good" in a moral sense. At any rate, how we think and talk of "the best" shapes our ideas of what is good.

We thus need to investigate critically the assumed strong analogy between questions of different kinds. One kind of question is "Is five more than one?," or "Which *number is higher*—5 or 1?" Another is "Is it *right* to let one person die (or kill one person), to avoid the predicted death of five?" Yet another is "Now that I am in the situation where I have to choose between this one person and several others whom I could treat instead this one, what *should* I do?"

The first is not a moral question at all. The third, under some circumstances, may not be either, though for different reasons. Sometimes it is, but not necessarily as an instance of the second question. Not all moral questions are such because they allow or require their rephrasing as

questions of the second kind. The strong analogy often drawn between the value of a quantity, moral rightness, and that what one needs to do under particular circumstances obfuscates the matter. Drawing this equation helps the ambition to have the tool that would enable one to exonerate oneself of moral responsibility in moral dilemmas, whether or not they are “hard.” In the same sense in which one does not need to pity the number 1 when truthfully stating that it is lower than number 5, one would also not need to regret the lives sacrificed. What gets overlooked here is that when one says “You say that I should have left the one person without treatment, but how could I have done it?,” she is not asking the other to recite to her a principle which she did not have available at the moment.

All in all, the kind of difficulty that one confronts in a genuine moral dilemma does not disappear simply because one probably did the best thing that one could do, under the circumstances. Dilemmas are not trade-offs, though some situations of dilemma are also situations of trade-off.

Acknowledgements

Work on this paper was supported by the project ‘Centre for Ethics as Study in Human Value’ (project No. CZ.02.1.01/0.0/0.0/15_003/0000425, Operational Programme Research, Development and Education, co-financed by the European Regional Development Fund and the state budget of the Czech Republic). I owe thanks to Marina Barabas, Michael Campbell, Matej Cíbk, Niklas Forsberg, Nora Hämäläinen, Ondřej Krása, Camilla Kronqvist, Kamila Pacovská and Hugo Strandberg for their helpful comments on the manuscript (Kamila and Mike in particular). The suggestions of my anonymous reviewers also helped me improve the text significantly.

References

- Agamben, Giorgio. 2020. “The Invention of an Epidemic.” *European Journal of Psychoanalysis*. Available at: <https://www.journal-psychoanalysis.eu/coronavirus-and-philosophers/>
- Amewu, Sena, Seth Asante, Karl Pauw, and James Thurlow. 2020. “The Economic Costs of COVID-19 in Sub-Saharan Africa. Insights from a Simulation Exercise for Ghana.” *The European Journal of Development Research* 32 (5): 1353–78. <https://doi.org/10.1057/s41287-020-00332-6>

- Anscombe, E. M. Gertrude. 1981. "War and Murder." In *Ethics, Religion and Politics*, 51–61. Oxford: Basil Blackwell.
- Browne, Brynmor. 1992. "A Solution to the Problem of Moral Luck." *Philosophical Quarterly* 42 (168): 345–56.
- Castrillón, Fernando & Marchevsky, Thomas (eds.). 2021. *Coronavirus, Psychoanalysis, and Philosophy. Conversations on Pandemics, Politics, and Society*. New York: Routledge.
- Correia, Sergio, Stephan Luck, and Emil Verner. 2020. "Pandemics Depress the Economy, Public Health Interventions Do Not: Evidence from the 1918 Flu" (5 June 2020). Available at SSRN: <http://dx.doi.org/10.2139/ssrn.3561560>.
- Cowley, Christopher. 2008. *Medical Ethics, Ordinary Concepts and Ordinary Lives*. Basingstoke: Palgrave Macmillan.
- Černý, David, Adam Doležal, and Tomáš Doležal. 2020. "Etická a právní východiska pro tvorbu doporučení k rozhodování o alokaci vzácných zdrojů při poskytování zdravotních služeb v rámci pandemie COVID-19." 2. Vydání. Ústav státu a práva AV ČR, v. v. i., Kabinet zdravotnického práva a bioetiky. Available at: https://zdravotnickepravo.info/wp-content/uploads/2020/10/Eticka-a-pravni-vychodiska-pro-tvorbu-doporuceni-v-ramci-COVID-19_podzim-2020.pdf
- Daniels, Norman. 2001. "Health-Care Needs and Distributive Justice." In *Bioethics*, edited by John Harris, 319–46. Oxford: Oxford University Press.
- Gaita, Raimond. 2004. "Breach of Trust. Truth, Morality and Politics." *Quarterly Essay* 16: 1–68.
- Gaita, Raimond. 2006. *Good and Evil: An Absolute Conception*. 2nd edition. New York: Routledge.
- Gaita, Raimond. 2020. "COVID, Quality and a Common World." *Meanjin* 79 (4): 84–99. Available at: <https://meanjin.com.au/essays/covid-quality-and-a-common-world/>
- Gans, Joshua. 2020. *Economics in the Age of COVID-19*. Cambridge (MA): The MIT Press.
- Horton, Richard. 2020. *The COVID-19 Catastrophe*. Cambridge: Polity Press.
- La Puma, John, and Edward F. Lawlor. 1990. "Quality-Adjusted Life-Years. Ethical Implications for Physicians and Policymakers." *The Journal of the American Medical Association* 263 (21): 2917–21. <https://doi.org/10.1001/jama.263.21.2917>
- Lawrence, David R., and John Harris. 2021. "Red Herrings, Circuit-Breakers and Ageism in the COVID-19 Debate." *Journal of Medical Ethics*. <https://doi.org/10.1136/medethics-2020-107115>

- MacKillop, Eleanor, and Sally Sheard. 2018. "Quantifying life: Understanding the history of Quality-Adjusted Life-Years (QALYs)." *Social Science and Medicine* 211: 359–66. <https://doi.org/10.1016/j.socscimed.2018.07.004>
- Marra, C. A., S. A. Marion, D. P. Guh, M. Najafzadeh, F. Wolfe, J. M. Esdaile, and A. H. Anis. 2007. "Not All 'Quality-Adjusted Life Years' Are Equal." *Journal of Clinical Epidemiology* 60 (6): 616–24. <https://doi.org/10.1016/j.jclinepi.2006.09.006>
- Phillips, Z. Dewi. 1979. "Do Moral Considerations Override Others?" *The Philosophical Quarterly* 29 (116): 247–54. <https://doi.org/10.2307/2218821>
- Reid, Lynette. 2020. "Triage of Critical Care Resources in COVID-19: a Stronger Role for Justice." *Journal of Medical Ethics* 46 (8): 526–30. <https://doi.org/10.1136/medethics-2020-106320>
- Savulescu, Julian, and James Cameron. 2020. "Why Lockdown of the Elderly Is Not Ageist and Why Levelling Down Equality Is Wrong." *Journal of Medical Ethics* 46 (11): 717–21. <https://doi.org/10.1136/medethics-2020-106336>
- Savulescu, Julian, Ingmar Persson, and Dominic Wilkinson. 2020. "Utilitarianism and the Pandemic." *Bioethics* 34 (6): 620–32. <https://doi.org/10.1111/bioe.12771>
- Singer, Peter, and Michael Plant. 2020. "When Will the Pandemic Cure Be Worse Than the Disease?" *Project Syndicate*. Available at: <https://www.project-syndicate.org/commentary/when-will-lockdowns-be-worse-than-covid19-by-peter-singer-and-michael-plant-2020-04>
- Spektor, Brandon. 2020. "Coronavirus: What is 'Flattening the Curve,' and Will It Work?" *Live Science*, 16 March 2020, <https://www.livescience.com/coronavirus-flatten-the-curve.html>.
- Stefánsson, H. Orri. 2020. "Three Mistakes in the Moral Reasoning About the Covid-19 Pandemic." *The Institute for Futures Studies*, Working paper 2020:12. Available at: <https://www.iffs.se/en/publications/working-papers/three-mistakes-in-the-moral-reasoning-about-the-covid-19-pandemic/>
- Wiinikka-Lydon, Joseph. 2019. *Moral Injury and the Promise of Virtue*. Cham: Palgrave Macmillan.
- Williams, A. O. Bernard. 1965. "Ethical Consistency." *Proceedings of the Aristotelian Society*, Supplementary Volume, 39 (1): 103–24.
- Williams, A. O. Bernard. 1981. "Moral Luck." In *Moral Luck*, 20–39. Cambridge: Cambridge University Press.
- Williams, Bridget, James Cameron, James Trauer, Ben Marais, Romain Ragonnet, and Julian Savulescu. 2021. "The Ethics of Age-Selective Restrictions for COVID-19 Control." Blog of *Journal of Medical Ethics*. Available at: <https://blogs.bmj.com/medical-ethics/2021/01/27/the-ethics-of-age-selective-restrictions-for-covid-19-control/>

-
- Žižek, Slavoj. 2020a. *Pandemic! COVID-19 Shakes the World*. New York: OR Books.
- Žižek, Slavoj. 2020b. *Pandemic! 2. Chronicles of a Time Lost*. New York: OR Books.

Susan Wolf on Supererogation and the Dark Side of Morality

Nora Grigore*

Received: 22 December 2020 / Revised: 11 July 2021 / Accepted: 1 December 2021

Abstract: Wolf proposes supererogation as a solution for curbing the exaggerated demands of morality. Adopting supererogation is supposed to prevent us from considering that all morally good deeds are obligatory. Supererogation, indeed, makes some morally good deeds merely optional, saving the agent from the requirement of behaving as much as possible like a Moral Saint. But I argue that Wolf cannot use supererogation in service of her overall project, for two reasons. First, because implied in the concept of supererogation is that going beyond duty adds to our humanity rather than detract from it (as Wolf argues). Secondly, after analyzing attempts to acclimatize supererogation in other theoretical frameworks, I conclude that supererogation can limit morality’s claims only if Wolf’s reasons of “individual perfection” can defeat moral reasons. I argue that a common scale of evaluation between moral and non-moral reasons is needed for their comparison, but Wolf explicitly rejects this way out.

Keywords: Morality; moral saints; supererogation.

1. Introduction

In “Moral Saints” Susan Wolf famously depicts a rather bleak image of the moral saint. A moral saint, claims Wolf, is someone “whose every action

*  <https://orcid.org/0000-0002-4035-5282>

 noragrigure@utexas.edu



is as morally good as possible, a person, that is, who is as morally worthy as can be" (Wolf 1982, 419). At the same time, she claims, there is a dark side to this moral excellence, mainly because dedicating so much time and resources to morality would have catastrophic consequences for the personal, private side of the moral agent.

One may protest the idea that living one's life to the moral extreme has such bleak consequences: the saint does not have to be this eviscerated self Wolf describes. Adams (1984), for instance, claims that if some suppositions from Wolf's picture are removed (e.g. the assumption that a moral saint should always maximize the moral good), then the resulting image is quite different. However, this is not the path I'll take. Rather, I'll show that, given Wolf's dark view of moral demands, her solution for restricting them (i.e. supererogation) doesn't work.

According to Wolf, following moral ideals has catastrophic effects for our personal lives. She therefore wishes to limit the influence of moral recommendation in favor of a personal, individual point of view, that she calls "the point of view of individual perfection." Wolf thinks that the best instrument for moral theories to make this limitation is supererogation.¹ Actions are usually called "supererogatory" when they are considered to be morally excellent, but not obligatory (paradigmatic examples being saintly or heroic deeds). The idea is that supererogation establishes a threshold (because supererogation is thought as going *beyond* duty) for what is obligatory; any moral behavior above it is merely permitted even if morally laudable (e.g., saintly and heroic action). I will focus on the relation between the image of a very demanding kind of morality and supererogation. And

¹ The usual general characterization of supererogation goes along these lines: "Supererogatory acts are those which lie 'above and beyond the call of duty'. Such acts characteristically enjoy a very high degree of value, probably more value than any other act available to the agent. (...) actions which is not wrong of the agent not to do (Dancy 1998, 173). Recent books covering the topic of supererogation start by this rough characterization: "It is often said that works of supererogation involve going beyond the call of duty, doing good in a way that transcends the requirement of moral obligation" (Mellema 1991, 3) or "Supererogation is the technical term for the class of actions that go 'beyond the call of duty'. Roughly speaking, supererogatory acts are morally good although not (strictly) required" (Heyd 2002).

I'll argue that supererogation cannot be accommodated in the framework adopted by Wolf, in which morality is seen as overly demanding, and two separate, independent scales for moral and non-moral values exist.

What I hope to show is that the failure to accommodate supererogation is not of merely local interest, a glitch in the bigger picture painted by Wolf about life and morality. There is a certain conception about morality as very demanding, (one might say "life-denying") that is so prevalent, so natural and by-default-adopted, that even authors who are not sympathetic towards it, who are critical and opposed, still espouse some of its basic assumptions. I think Susan Wolf is such an author and therefore a very relevant illustration of this pervasive image of morality. In short, Susan Wolf is famous for saying that moral saints cannot but have a diminished quality of life, and a diminished humanity. In the fight between 'life' and morality, she is no doubt, on the side of life. However, my point is that she shares this presupposition with her adversaries, that normal human life and the "higher flights of morality" are engaged in a struggle or, at least, in zero-sum game. She shares this presupposition with many other illustrious philosophical names. After all, it is a quite a Nietzschean claim that morality is "life-denying" (to be exact, "morality of slaves" has this role). Korsgaard (1996), quoting Nietzsche, considers that the proper role of morality is that of imposing or "forcing" values upon nature and life, thereby restricting and shaping their course. My present point is that this is a rich philosophical tradition, and one in which supererogation (with its knack for lifting some moral obligations) has never been at ease. Nevertheless, Wolf seems to want to have both: a conception of demanding, obligations-imposing morality and supererogation. My aim is to explain why this is an impossible philosophical mission.

In short, my aim is to argue for the following conditional claim: if Wolf wishes to keep (what she takes to be) the commonsensical image of morality, then it will be very difficult to *also* maintain the theoretical solution she proposes, namely supererogation. The conception of morality Wolf wishes to keep isn't welcoming towards the concept of supererogation. In that sense, Wolf doesn't ultimately "respond to the unattractiveness of the moral ideals that contemporary theories yield" (Wolf 1982, 434).

2. The dark side of moral saintliness

Wolf's main point is that we spontaneously judge moral ideals to be unpleasant and damaging *from another point of view than the moral one*, namely from what she calls "the point of view of individual perfection."² In other words, she thinks that if one follows the recommendations of commonsense morality (promoting at all times other people's good and disregarding one's own interest) and moves asymptotically towards a moral ideal, then one will end up hurting some nonmoral, personal values. One will end up, that is, with a mutilated life in one respect or another. Susan Wolf acknowledges that commonsense morality doesn't make the saintly moral ideal into an *obligatory* path to take. But she thinks that, even if not obligatory, this path is recommended as *the best* path one could take. She objects to that, saying that this can be seen as the best *moral* path and, at the same time, as a *bad* choice for the agent and those close to him in many other important respects (e.g. from the point of view of one's personal life). Her recommendation for solving this tension between the moral point of view and the point of view of individual perfection is to somehow restrict the claims that morality places on us and to give personal ideals legitimacy in our evaluations:

If we are not to respond to the unattractiveness of the moral ideals that contemporary theories yield either by offering alternative theories with more palatable ideals or by understanding these theories in such a way as to prevent them from yielding ideals at all, how, then, are we to respond. Simply, I think, by admitting that moral ideals do not, and need not, make the best personal ideals. (Wolf 1982, 434–35)

I want to underline that Wolf mentions here *two other possible ways* of preventing the undesirable consequence of self-mutilating moral sainthood: a) modifying our views about morality (such that its demands are no longer at odds with goals of personal perfection); or b) making the pursuit of an ideal

² Wolf (1982, 427): "Let us call the point of view from which we consider what kinds of lives are good lives, and what kinds of persons it would be good for ourselves and others to be, the point of view of individual perfection."

something of an undertaking outside morality. In other words, one might think that our conception of morality should change in such a way that moral excellence would not involve a mutilated self; or, alternatively, that our conception of morality would include only rules about what is obligatory and forbidden, and would place aspirations or ideals outside morality.

She gives counterarguments to taking route a), which implies that she wants to keep both the claims of what it means to be (commonsensically) moral and the moral ideals that they engender. However, she also wants some limits placed on moral claims when they go against personal goals, and she mentions *supererogation* as a helpful theoretical instrument. Supererogation may seem helpful because it establishes limits for moral obligation, a kind of threshold above which the agent isn't morally obligated to act. Therefore, heroic or saintly actions would not be morally required, and their omission could not be regarded as a fault (it can be seen as an appropriate instrument for pushing back against the threat of moral claims). I will argue, however, that even though supererogation is usually regarded as an instrument for pushing back against maximization requirements (for instance), her way of seeing morality is an unwelcoming theoretical environment for the concept of supererogation.

3. Morality's demandingness

In other words, morality should not be the only or even the most important set of values guiding our actions: "morality itself should not serve as a comprehensive guide to conduct" (Wolf 1982, 434). Once we admit that there are certain nonmoral values that one is right to attend to, then the claims of morality to be the most important guide to action are limited, mitigated so as to not lead to the extremes exhibited in the image of the moral saint.

Following her recommendation that one should not take moral evaluation and action too far, Wolf also recommends amending moral theories in order to fit this limitation of powers regarding moral claims. Namely, she recommends that moral theories should use *supererogation* as a helpful instrument:

From the moral point of view, we have reasons to want people to live lives that seem good from outside that point of view. If, as I have argued, this means that we have reason to want people to live lives that are not morally perfect, then any plausible moral theory must make use of some conception of supererogation. (Wolf 1982, 438)

Supererogation seems like a good choice in this respect because it implies that saintly and heroic deeds cannot and should not be deemed obligatory. However, I will claim that other aspects of supererogation come into conflict with some parts of Wolf's story, and specifically with her way of seeing morality.

Wolf claims (rightly, I think) that commonsense morality is heavily other-oriented: according to its recommendations one is supposed to help others on each and every occasion, regardless of the sacrifice imposed on the agent.

Notice that these two features—the expectation to help others, and the agent's sacrifice not counting as a valid moral consideration against giving that help—jointly constitute what authors usually call the *demandingness* of morality, i.e. the heavy burden placed on the moral agent to act sometimes against her own interest for the general good. In Wolf's case, she sees demandingness as unjustifiably affecting the agent's private life. Her solution *isn't to change our conception of morality but to mitigate some of its effects on us*. This move, I will argue in the next section, doesn't work because once one admits that morality is overly demanding, the effects of conceiving morality as a harshly demanding enterprise are difficult to avoid.

4. What is problematic in Wolf's solution

Supererogation seems to be a good solution and the right theoretical tool for Wolf. She wants to be able to say that in some (but not all) circumstances the agent may ignore morality's recommendations in order to attend to personal perfection. Supererogation, as usually presented, grants this permission even though it doesn't specify exactly why or what kind of reasons the agent is allowed to follow instead. The concept of supererogation only allows that there are circumstances when we are able to rightly ignore the morally right thing to do, without punishment, blame or justified reproach

(even if, subjectively, one might feel regret). It seems, therefore, that the concept of supererogation can provide what Wolf is looking for: some limitation of morality's grip on our life and values, making room for the point of view of individual perfection.

I agree that supererogation is an instrument able to provide all of these things. But its task is made difficult (if not impossible) by the particular conception of morality that Wolf keeps (even if she pushes against its perceived excesses). Supererogation can be deployed, as a theoretical instrument, against favorable or unfavorable theoretical backgrounds: some moral theories may make less room or no room at all for supererogation, depending on various other factors. For example, an obviously unfavorable environment for supererogation is one in which morality presupposes maximization of the good, as in act-utilitarian or act-consequentialist theories. If one is required to maximize the overall good on each occasion, then there cannot be acts that are good but not required, or good acts that can be omitted without blame; so no supererogatory actions exist. Basically, any theory claiming that what is morally good has to be covered by some kind of obligation or has to stay under an "ought," will threaten supererogation.

My first claim is that Susan Wolf has a specific conception on which morality is—and should be—demanding and imperative.³ My second claim is that it is very difficult to integrate supererogation into such a way of seeing morality.

One might think that my objection to Susan Wolf misses the point because the imperative character of morality was the whole purpose of employing supererogation. Precisely *because* morality is imperative and demanding (threatening to invade the personal domain, as Wolf sees it) we need something to curb its claims; had morality not been so demanding, then there would have been no need for supererogation in the first place. Supererogation is presumed to bring much needed permissions for the agent in the austere environment dominated by moral imperatives. My response is that not all theoretical landscapes can be balanced just because one needs some balance in them. Sometimes, the theoretical devices embedded in the theory simply exclude the possibility of supererogation. (E.g., the

³ Wolf claims she is tapping into the commonsense view of morality.

maximization required by some utilitarian theories can promptly exclude supererogation even if one might think supererogation is needed in those theories in order to make them more intuitively plausible.) In other words, I agree that morality being perceived as demanding is the circumstance where one is more likely to need the theoretical help provided by supererogation. However, this doesn't mean that it can always be *successfully* deployed and integrated within a certain theory. I think Wolf's theory is one of the unsuccessful cases. Let me explain why.

Susan Wolf clearly states that, according to commonsense morality, saintliness or ideal behavior isn't obligatory. However, both common sense and her own account of morality tend to go against this thesis. This is important because a morality that admits that some morally good things are not obligatory doesn't threaten supererogation. On the other hand, a morality that at least tends towards making all morally good things obligatory (what I have been calling an *imperative* kind of morality) is usually a threat for supererogation. When Wolf says that moral ideals are not obligatory, she seems to regard commonsense morality as being the former kind of morality. However, when she actually pushes against the claims of morality as she understands it, that morality seems closer to the latter kind. I am going to argue that both her account of morality and the commonsensical one have aspects similar to the imperative way of seeing morality.

First, morality according to common sense is far from being a coherent set of beliefs. It is true that, usually, we do not regard moral ideals as obligatory, and it is also true that supererogation (going beyond duty) is a commonsensical notion. So it would seem that there are some obligatory and some non-obligatory (e.g. supererogatory, saint-like) types of actions according to common sense. However, it is also a commonsense belief that anyone who can help, *should* help others in a difficult situation. If this "should" is translated into moral obligation, then each time one helps, one is simply fulfilling a moral obligation—something that is not, morally speaking, optional. I think this is the basic intuition behind the good-ought tie-up, namely that *one has a moral obligation to help others* in need because this is what constitutes the moral good—one might say that this is what morality is all about. The corollary of this thought is that one cannot invoke the inconveniences, or the losses suffered by oneself in order to opt out of

moral obligation: this is how morality works, by foregoing one's own interests to altruistically care about other people.

Consequently, there is a tension here and we can see that commonsense morality may indeed seem, at times, very demanding. If one interprets it as saying that every morally good deed *should* be done, one can also interpret this as requiring an ascent to moral ideals. Wolf herself mentions in passing something resembling a tension in the commonsense view, but she attributes it to different contexts and doesn't elaborate on the relation between ideals and contexts. She says that "outside the context of moral discussion" we consider it natural to reject the model of the moral saint, because we agree that we aren't blamable if we don't always act following the highest moral recommendation). But in the context of moral discussion, however, we also want to claim that "one ought to be as morally good as possible" and it would be at least shameful not to aim at that:

In other words, I believe that moral perfection, in the sense of moral saintliness, does not constitute a model of personal well-being toward which it would be particularly rational or good or desirable for a human being to strive. Outside the context of moral discussion, this will strike many as an obvious point. But, within that context, the point, if it be granted, will be granted with some discomfort. For within that context it is generally assumed that one ought to be as morally good as possible and that what limits there are to morality's hold on us are set by features of human nature of which we ought not to be proud. (Wolf 1982, 419)

I think this is a quite clear expression of the tension I want to point at: on the one hand it is obvious that we are not required to be saints; on the other hand, it seems equally obvious that we *do* want to say that we are required to pursue what is morally best. Of course, one can choose one or the other, by accepting the moral obligation implied in the requirement to always do your morally best, or by limiting it. My point is simply this: that *Wolf seems to want to embrace both at the same time*. This is because she wants to keep the very demanding version of morality that makes every good deed required of us and, at the same time, to have supererogation limit these requirements. Or, in other words: to accept that we always have to aim at doing our morally

best and that we do not always have to. In this respect, Susan Wolf's position is in keeping with commonsense morality by taking in its inner tensions. But if her take is ambivalent in this respect, then one part of her image of morality (the one describing morality as demanding and imperative) is the one that makes supererogation difficult to accommodate.

My second point about the tension between requiring and not requiring that all morally good deeds be performed, regards the way Susan Wolf herself chooses to depict morality. Especially when she argues against morality's demanding ideals and in favor of limitations being imposed (in order to make room for the legitimacy of the point of view of individual perfection), Wolf is presenting a morality that has a strong, imperative character, one that overrides and demotes other concerns:

[T]he desire to be as morally good as possible is apt to have the character not just of a stronger, but of a higher desire, which does not merely successfully compete with one's other desires but which rather subsumes or demotes them. The sacrifice of other interests for the interest in morality, then, will have the character, *not of a choice, but of an imperative*. (Wolf 1982, 423–24, my italics)

This image, of an imperative morality, should not come as a surprise if we consider one other aspect of the problem, namely that Wolf sees morality as engaged in a zero-sum game with the domain of the private, personal life of the agent. This point is made clear by the disturbing picture of the moral saint: whatever one does for the moral good of others is a loss for the personal self; and, conversely, whatever one does good from the point of view of individual perfection is a rejection of the relentless demands of morality.⁴ A consequence of this way of thinking is that morality is seen as a difficult, demanding, and unpleasant (to say the least) to follow. Such a view is (not necessarily, but likely) going to have to rely on obligation in order to see its

⁴ “The normal person's direct and specific desires for objects, activities, and events that conflict with the attainment of moral perfection are not simply sacrificed but removed, suppressed, or subsumed. The way in which morality, unlike other possible goals, is apt to dominate is particularly disturbing, for it seems to require either the lack or the denial of the existence of an identifiable, personal self” (Wolf 1982, 424).

tasks fulfilled and its recommendations followed, because it is unlikely that people would want to undertake such unpleasant tasks voluntarily (especially since sacrifice is often involved). And this shows, against our moral intuition that not all moral deeds should be demanded, that certain moral frameworks end up with extending (at least some form of) obligation to the whole domain of morality.

There is, I think, a third argument for my tenet that Susan Wolf inclines towards an imperative view of morality, even if she doesn't say it explicitly. She says so herself, in her passionate plea for the personal domain: morality *should not* play the role of supreme scale of values, and the agent should not ask permission for omitting to always do the morally best thing. But this protest means that she believes that moral value *is* the value that trumps any other kind of value (moral values “subsume or demote” other values⁵), and the agent might feel that she has to ask permission in order not to do the morally best thing. It is not only when she *opposes* these tendencies that she recognizes them (the tendencies of *requiring* any morally good deed, of following any moral good with an “ought”). It is also when she approvingly characterizes morality that she says the following: “A moral theory that does not contain *the seeds of an all-consuming ideal* of moral sainthood thus seems to place false and unnatural limits on our opportunity to do moral good” (Wolf 1982, 433, my italics).

Therefore, I think one should not be surprised if Susan Wolf ends up with an imperative version of morality, one where obligation plays a central role and usually does not accommodate supererogation. However, it can be argued that there are examples of such moral theories that have tried to accommodate supererogation, and what Wolf ends up with is not a straightforward contradiction, but rather a puzzling tension. I'll now further pursue the charitable assumption that Wolf's view could be one of them. The main task of the next section is to present various theoretical possibilities and evaluate them to see if they could be a path Wolf could take.

⁵ Wolf (1982, 424).

5. How supererogation may be accommodated

In a way, what Susan Wolf recommends is quite banal and, moreover, it is something that we already do routinely: we limit moral demands when they threaten other parts of our lives. The problem for a moral theory, however, is to find a justification for this limitation in its own terms. The problem is: can it be morally justified to limit morality's demands? And how exactly will that justification look for a particular theory?

When speaking specifically about supererogation, the problem is already famous: the puzzle, paradox,⁶ or simply the problem of supererogation—they all refer to a number of difficulties for various theories in justifying the agent's permission to sometimes omit the morally best action. For Susan Wolf, in particular, the problem of supererogation is the following: *How can one justify that we are sometimes allowed not to follow moral prescriptions, and instead act for the good from the point of view of individual perfection?* It could seem that she offers an answer when she worries about the possible objection that not pursuing moral ideals in order to attend to individual perfection is just an excuse for pursuing a selfish agenda. Differently put: how do we know that when we reject morality's claims on us, we do this for the right reasons? She replies that there are nonmoral values and nonmoral virtues involved in individual perfection. These give rise to valid nonmoral reasons to sometimes reject the claims of moral reasons.

In other words, some of the qualities the moral saint necessarily lacks are virtues, albeit nonmoral virtues, in the unsaintly characters who have them. In advocating the development of these varieties of excellence, we advocate nonmoral reasons for acting, and in thinking that it is good for a person to strive for an ideal that gives a substantial role to the interests and values that correspond to these virtues, we implicitly acknowledge the goodness of ideals incompatible with that of the moral saint. (Wolf 1982, 426)

⁶ Cf. Archer and Ridge (2015) and Horgan and Timmons (2010).

On the image Wolf offers here, there are moral reasons in favor of pursuing supererogatory acts (or saintly acts), and they are sometimes opposed by nonmoral reasons (belonging to individual perfection) that sometimes win the confrontation between reasons.

There are two immediate problems with this response that I can discern. First, leaving aside the problems raised by confronting moral with nonmoral reasons, is the missing common scale of comparison. The moral and the individual point of view are independent points of view, according to Wolf, without an overarching framework to encompass them both. It is true that the moral point of view gives some weight to the individual point of view and the other way around. But when they are in conflict, there are no means to decide which one will prevail. Wolf explicitly rejects the possible construction of a common framework out of fear that it will become one that will again make moral value the ruling, deciding value:

The philosophical temperament will naturally incline, at this point, toward asking, “What, then, is at the top—or, if there is no top, how are we to decide when and how much to be moral?” In other words, there is a temptation to seek a metamoral—though not, in the standard sense, metaethical—theory that will give us principles, or, at least, informal directives on the basis of which we can develop and evaluate more comprehensive personal ideals... I am pessimistic, however, about the chances of such a theory to yield substantial and satisfying results. For I do not see how a metamoral theory could be constructed which would not be subject to considerations parallel to those which seem inherently to limit the appropriateness of regarding moral theories as ultimate comprehensive guides for action. (Wolf 1982, 438–39)

This is only a general observation regarding her solution. But, more to the point, if Wolf wants to employ supererogation in her theory, this comes with some additional complications. When trying to justify why an agent is allowed to omit some morally excellent actions, the justification cannot be merely prudential—it has to carry moral weight. It is obvious why, for *prudential* reasons, the agent can omit heroic or saintly deeds: they involve heavy self-sacrifice. What is difficult to do, and what the problem of supererogation asks, is what *moral* reasons one could have not to act saintly or heroically.

And this is a justified demand if we consider that, on the commonsense notion of supererogation, it is not only disadvantageous to place the agent under an obligation to act heroically, but it is first of all *morally wrong*—we feel—to make sacrifices of this kind *obligatory* (Urmson 1958). The obligation itself seems in these cases immoral, for in most cases something is wrong with being constrained to give your life or limb for the greater good. The problem is—to give a theoretical support for this impression that “something is wrong.” So there must be some *moral* reasons justifying the fact that we are *not obligated but merely permitted* to act in a saintly or heroic manner.

Of course, Wolf’s reasons for foregoing saintly actions are explicitly *non-moral*, personal and partial to the agent. She insists that they are not exactly prudential reasons, as they do not have the agent’s interest in view, but they are something else, namely concerned with the agent’s individual perfection. Therefore she cannot provide, in her theory’s own terms, a *moral* justification for sometimes disobeying morality and following one’s own plans.⁷ Consequently, she doesn’t have a good answer to the problem of supererogation, even if she says that the claims of the personal, individual point of view are recognized by morality to some extent. The problem for her theory is that these claims, when seen from within the moral point of view, don’t carry much weight according to her own evaluation. Therefore, they cannot be reliable in “defeating” moral reasons that would recommend heavy sacrifices on the part of the agent.

6. Morality and supererogation

To recap, Wolf’s image about morality is that the moral domain is at odds with the personal domain, that it is other-oriented and has an imperative character. Wolf wants to keep these features, as she believes that this is what morality should look like, but at the same time she wants to restrict moral claims such that one would not be under an obligation to give up personal plans in order to attend to helping others. For this task she proposes that moral theories make use of supererogation.

⁷ Jonathan Dancy raises a similar objection to her theory in *Moral Reasons*.

I have argued, first, that supererogation is difficult to accommodate within moral theories that are obligation-based and have an imperative character because these theories usually tend to assume that all morally good deeds stay under an “ought” (and therefore come into conflict with the idea that some excellent moral deeds are merely permitted—as supererogatory). In Section 3, I have argued that both commonsense morality and Wolf’s own position fall into this category of obligation-based morality, despite the fact that they do contain some opposing intuitions in this respect. However, because supererogation can be accommodated even in unfriendly environments by making appropriate theoretical adjustments, I have looked into the possibility of making such adjustments in Section 4. In order to be able to use supererogation, Wolf’s theory should be amenable to a credible strategy for morally justifying the omission of saintly or heroic actions. The justification should explain why or how, sometimes, nonmoral reasons from the personal side are able to defeat moral reasons. I’ve argued that, because Wolf doesn’t have a common scale for moral and nonmoral reasons, she has no way of explaining how such a confrontation may be decided.

In the end, the issue of being able to use supererogation within Wolf’s framework is this: Supererogation, as a conceptual structure, has two main components,⁸ namely that some excellent moral deeds are not (and should not) be obligatory, and that the same moral deeds are praiseworthy or good from a moral point of view. I have argued that neither of these components fits with what Wolf wants to say.

The first component tends to be undermined directly by the image of an imperative, obligation-based morality because this kind of morality tends to make all morally good deeds obligatory. Even if and when limits are imposed to obligation in these theories, they look like concessions made to human weakness or to everyday intuitions⁹ rather than limits imposed by morality itself. For the state where *all morally good deeds are obligatory* is regarded as the *default rational* state from which one can depart by various technical means or by making concessions to the fact that humans cannot

⁸ Dreier (2004) and Hurka and Schubert (2012).

⁹ Slote (1984) and Scheffler (1994) argue for integrating supererogation in act-consequentialist frames in order to accommodate commonsensical intuitions about morality.

live up to this high (and very demanding) moral standard. For example, Thomas Nagel claims that allowing the omission of morally good acts by invoking supererogation is a compromise due to “human weakness,” compromise in which “[w]e must so to speak strike a bargain between our higher and lower selves in arriving at an acceptable morality.”¹⁰ Wolf vigorously protests this position; for her, our omission of morally saintly deeds should need no permission and no excuse:

It is misleading to insist that one is permitted to live a life in which the goals, relationships, activities, and interests that one pursues are not maximally morally good. For our lives are not so comprehensively subject to the requirement that we apply for permission, and our nonmoral reasons for the goals we set ourselves are not excuses, but may rather be positive, good reasons which do not exist despite any reasons that might threaten to outweigh them. (Wolf 1982, 436)

So, on Wolf’s view, supererogation cannot come as an excuse for moral weakness. And, indeed, once one admits that supererogatory actions exist, one should also admit that some moral deeds can be omitted without an apology needed from the agent. But then Wolf’s theory should afford means (even if not explicitly given) to limit morality’s claims by using *moral* reasons, or at least reasons that can be given significant moral weight. If not, her nonmoral personal reasons, that are supposed to justify our omission to behave saintly, can look like just another *prudential* reason to not risk too much in the service of morality. To show how nonmoral personal reasons can defeat moral ones, Wolf would need something like a common scale of values, i.e., she would need some theoretical device that would allow for a comparison between the two kinds of reasons. And she explicitly rejects this possibility.

The second component of supererogation, the one claiming that actions that go beyond duty are morally excellent, praiseworthy actions, doesn’t fare much better than the first component, since Wolf famously claims that saints are repulsive, defective human beings. Recall this is because Wolf thinks that the more one improves morally, the narrower one’s mental horizons become (due to an obsessive concern with helping others), the less time one has for

¹⁰ Nagel (1986, 202).

oneself and the more unpleasant and lacking in humanity they become. This is also a feature of her view about morality as being an extremely demanding enterprise, one engaged in a competition with the personal domain, such that each time one acts morally the personal domain loses, and the other way around. For Wolf, there is definitely such thing as “too much morality.”¹¹ It isn’t clear if this is a case of “too much of a good thing” or a case of “something better to have only in moderate quantities.” Considering her tone, I would venture to say that it is the latter.

Regardless of how one chooses to interpret her position in this respect, Wolf’s view regarding moral saints paints a very different picture from the one promoted by supererogation regarding saintly and heroic deeds. For, in the case of supererogation, saintly actions are presented as praiseworthy and overall good. Wolf’s reply to the objection that her image of the saint is not very appealing is that a saint is morally excellent, but a rather unpleasant figure from another point of view, that of personal perfection. However, I believe this doesn’t address the discrepancy between her image of the saint and the commonsense one regarding supererogatory action. Our tales of heroic and saintly deeds are not cautionary tales about how the hero was a morally excellent person, but nevertheless they ruined their humanity out of lack of moderation regarding the moral good and, therefore, one should be careful not to do the same. Quite the contrary. The main character from *Schindler’s List* who, at the end of the story, is tortured by remorse thinking that he could have done more, earn more money and buy more lives—is hardly someone to whom we could reasonably recommend moderation because he could be seen as a glutton for morality, having lost part of his humanity in the process. Therefore, I think that one cannot employ the usual concept of supererogation while at the same time denying that agents who act supererogatorily are morally excellent *and better human beings overall*. This part of supererogation, claiming that agents who go beyond duty are not only partially admirable but also overall better human beings, will always be in conflict with the image of the moral saint depicted by Wolf.

I believe it is important to see that Susan Wolf’s way of seeing morality is part of a venerable tradition, and more importantly, part of a tacitly held

¹¹ Wolf (1982, 483): “In other words, there seems to be a limit to how much morality we can stand.”

opinion that morality is and should be demanding, harsh, life-denying. Once we see this as a philosophical presupposition (and not as a “fact of moral life” as Nagel (1986), for example would claim), we may begin to wonder if another view of morality is possible, and if it would be a better fit for the concept of supererogation.

References

- Adams, Robert Merrihew. 1984. “Saints.” *The Journal of Philosophy* 81: 392–401. <https://doi.org/10.2307/2026294>
- Archer, Alfred, and Michael Ridge. 2015. “The Heroism Paradox: Another Paradox of Supererogation.” *Philosophical Studies* 172: 1575–92. <http://www.jstor.org/stable/24704233>
- Dancy, Jonathan. 1993. *Moral Reasons*. Oxford: Blackwell.
- Dancy, Jonathan. 2004. “Enticing Reasons.” In *Reason and Value: Themes from the Moral Philosophy of Joseph Raz*, edited by Wallace, Jay R., Philip Pettit, Samuel Scheffler and Michael Smith, 91–118. Oxford: Clarendon Press.
- Dreier, James. 2004. “Why Ethical Satisficing Makes Sense and Rational Satisficing Doesn’t.” In *Satisficing and Maximizing*, edited by Byron, Michael, 131–54. Cambridge: Cambridge University Press. https://doi.org/10.1111/j.1468-0149.2007.00449_9.x
- Heyd, David. 1982. *Supererogation: Its Status in Ethical Theory*. Cambridge: Cambridge University Press.
- Horgan, Terry, and Mark Timmons. 2010. “Untying a Knot from the Inside Out: Reflections on the ‘Paradox’ of Supererogation.” *Social Philosophy and Policy* 27: 29–63. DOI:10.1017/S026505250999015X
- Hurka, Thomas, Esther Schubert. 2012. “Permissions to Do Less Than the Best: A Moving Band.” *Oxford Studies in Normative Ethics* 2: 1–27. DOI:10.1093/acprof:oso/9780199662951.003.0001
- Korsgaard, Christine. 1996. *The Sources of Normativity*. Cambridge: Cambridge University Press.
- Nagel, Thomas. 1986. *The View from Nowhere*. New York & Oxford: Oxford University Press.
- Scheffler, Samuel. 1994. *The Rejection of Consequentialism: A Philosophical Investigation of the Considerations Underlying Rival Moral Conceptions*. Oxford: Oxford University Press.
- Slote, Michael. 1984. “Satisficing Consequentialism – I.” *Proceedings of the Aristotelian Society* 58: 139–63. <https://doi.org/10.2307/2940825>

Urmson, James Opie. 1958. "Saints and Heroes." In *Essays in Moral Philosophy*, edited by Melden, Abraham I., 198–216. Seattle: University of Washington Press.

Wolf, Susan. 1982. "Moral Saints." *Journal of Philosophy* 79: 419–39.
<https://doi.org/10.2307/2026228>

The Problem of Intention and the Evaluative Properties of Effects in the Knobe Effect


Andrzej Waleszczyński* – Michał Obidziński** – Julia Rejewska***

Received: 30 July 2020 / Revised: 20 July 2021 / Accepted: 1 December 2021

Abstract: In the article, we present analyses and findings which add precision to the role of intentions and the relation between effects in attributing the intentionality of causing a side effect. Our research supplements and modifies numerous findings regarding the appearance of the so-called Knobe effect. The experiments and analyses show that the very originality of the story used by Knobe and the relationship between the evaluative properties of the main effect and the side effect results in an asymmetry of responses and contributes to the occurrence of the side-effect effect. Because of this, we reject

* Cardinal Stefan Wyszyński University


 <https://orcid.org/0000-0003-0426-3919>

 Institute of Philosophy, Faculty of Christian Philosophy, Cardinal Stefan Wyszyński University, Wóycickiego 1/3, 01-938 Warsaw, Poland

 a.waleszczyński@uksw.edu.pl

** Cardinal Stefan Wyszyński University


 <https://orcid.org/0000-0002-7854-3123>

 Institute of Philosophy, Faculty of Christian Philosophy, Cardinal Stefan Wyszyński University, Wóycickiego 1/3, 01-938 Warsaw, Poland

 m.m.obidziński@gmail.com

*** Cardinal Stefan Wyszyński University

 <https://orcid.org/0000-0001-5740-6608>

 Institute of Philosophy, Faculty of Christian Philosophy, Cardinal Stefan Wyszyński University, Wóycickiego 1/3, 01-938 Warsaw, Poland

 julia.rejewska@gmail.com



the thesis that the mode of attitude of the agent to the caused side effect or that the social expectation of this attitude determine the attribution of the intentionality of the caused effect. On the contrary, we defend the thesis that it is the relationship between the evaluative properties of the main effect and those of the side effect, as well as the impact of a side effect on the main effect, that significantly influence the attribution of intentionality in causing a side effect.

Keywords: Side-effect effect; Knobe effect; intention; intentionality; evaluative properties.

1. Introduction

Studies on intentionality are of key importance for the philosophy of action (Mele 1992, 199). Standard accounts of intentionality show that the attribution of intentionality is based on clearly definable descriptive properties of the situation or of the agent (Malle and Knobe 1997; Mele and Sverdlik 1996). From this perspective, intentional causation of an effect is possible if the agent had the intention of causing it (Adams 1986; McCann 1987).

In recent years, a considerable contribution to studies on intentionality has been made by experimental philosophy, in particular studies on the attribution of intentionality in causing side effects. Philosophers and psychologists are very much interested in those experiments which analyze the so-called side-effect effect, also called the Knobe effect (Knobe 2003, 2006; Nadelhoffer 2005, 2006; Nichols and Ulatowski 2007; Nado 2008; Guglielmo and Malle 2010; Uttich and Lombrozo 2010). The effect reveals an asymmetry in the attribution of intentionality. It turns out that people are more apt to attribute intentionality in causing a side effect when the effect is negative than when it is positive. This has led to an alternative view which suggests that the attribution of intentionality in causing a side effect may also depend on its morally negative properties.

A detailed analysis of the impact of the agent's attitude on the side effects caused and the relationship between the moral weight of the main effect and the side effect was presented by Joshua Shepherd (Shepherd 2012). Further studies describing and analyzing the significance of the

relationship between the main effect and the side effect (Waleszczyński, Obidziński, and Rejewska 2019) have provided new interesting data on the emergence and disappearance of the Knobe effect. They have shown that the attribution of intentionality in causing a side effect is also possible in the case of positive effects and when the agent does not care about causing it. These findings are problematic for the explanations of the Knobe effect provided so far (Nadelhoffer 2006b; Nichols and Ulatowski 2007; Wright and Bengson 2009; Holton 2010; Sripada 2010; Sripada and Konrath 2011; Cova et al. 2012, Paprzycka 2015; Hindriks 2011, 2014; Hindriks, Douven, and Singmann 2016).

In this article, we will focus first on examining the relevance of the intention of causing a side effect for the attribution of intentionality to the agent. Our analyses will be concerned with the attribution of intentionality in causing a side effect, and not the attribution of intentionality in the action as such. Consequently, our conclusions will not apply to the intentionality of causing the main effect, but to that of causing side-effects. We will present analyses and findings which add precision to the role of intentions in attributing the intentionality of causing a side effect. Second, we will focus on further empirical investigations of the interaction between the importance of the main effect and the valence of the side effect (Cova et al. 2012, 402). We will defend the thesis that it is the relationship between the moral evaluative properties of the main effect and those of the side effect, and the impact of a side effect on the main effect that significantly influence the attribution of intentionality in causing a side effect.

2. Concepts of intentional action

Analyses in the field of experimental philosophy, despite having analytical elements, are classified as experimental descriptivism (Nadelhoffer and Nahmias 2007). Therefore, it is necessary to emphasize the distinctiveness of research in the field of experimental philosophy from conceptual analysis, the aim of which would be mainly to determine the conditions for the application of appropriate concepts (Knobe and Nichols 2008, 5). In his research, Knobe searches for cognitive mechanisms shaping popular intuitions that would satisfactorily explain the relevant mental processes (Knobe

2016). For this purpose, he uses empirical research in which participants are confronted with scenarios based on thought experiments. From this perspective, experimental philosophy is part of the cognitive sciences (Piekarski 2017, 112).

Research on the concept of intentional action conducted within the framework of experimental philosophy focuses on the study of the relationship between the folk concept of intentional action, i.e. one that is part of folk psychology, and the philosophical concept of intentional action. They show that the folk concept of intentional action does not correspond to the concept used by philosophers and is not limited to the concepts of intention and prediction. In the folk concept of intentional action, its effects and how we evaluate them are also taken into account (Piekarski 2017, 115).

According to Fred Adams and Annie Steadman, there are basically two concepts of intentional action. The first assumes that person S intentionally performs action A only when S intends to do A. In this case, having an intention is a necessary condition of an (intentional) action. According to the second concept, person S intentionally performs action A, not intending to do A, as long as action A is predicted by person S and accepted as a consequence of action S. In this case, the action may be intentional, even though the person has no intention of doing it (Adams 1986; Adams and Steadman 2004a, 2004b). However, the multiplicity and complexity of the results of research conducted as part of the experimental philosophy on the concept of intentional action generate the formulation of new explanations that go beyond the above two concepts of intentional action. Taking into account publications from recent years, one should also mention the explanations regarding the intentional actions provided by the representatives of responsibilism (Paprzycka 2012, 476–77). Herbert Hart, in his famous article, ‘The Ascription of Responsibility and Rights’ (1949), that action statements are not action statements, but rather ascriptive statements. In other words, they assign responsibility for certain events in the world before assigning them intentionality. Although this position was rightly criticized, it is worth emphasizing the main intuition presented by Hart, which may be significant for the analysis of the folk concept of intentional action. It concerns the fact that the concept of action is a secondary concept to the concept of responsibility (Sneddon 2006). This leads to a situation where the

intentionality of the action may be attributed to person S due to the prior attribution of responsibility to person S for action A, even in a situation in which he had no intention of causing action A, and did not even anticipate it. As a consequence, intentionality may be assigned due to a breach of obligations or applicable standards (Paprzycka 2015). This means that apart from intention and prediction, there may appear normative and evaluative factors that affect the attribution of intentionality not only to actions, but also to the side effects they cause.

3. Knobe effect

Joshua Knobe presented a story in which the “HARM” condition was as follows:

The vice-president of a company went to the chairman of the board and said, ‘We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment.’ The chairman of the board answered ‘I don’t care at all about harming the environment. I just want to make as much profit as I can. Let’s start the new program.’ They started the new program. Sure enough, the environment was harmed. (Knobe 2003, 191)

In the “HELP” condition, the structure of the story was the same, with the only difference being that the side effect is positive—it will “help the environment.”

The vice-president of a company went to the chairman of the board and said, ‘We are thinking of starting a new program. It will help us increase profits, but it will also help the environment.’ The chairman of the board answered ‘I don’t care at all about helping the environment. I just want to make as much profit as I can. Let’s start the new program.’ They started the new program. Sure enough, the environment was helped. (Knobe 2003, 191)

After being presented with the story, depending on the scenario each respondent was asked a question: “Did the chairman intentionally harm the

environment?” or “Did the chairman intentionally help the environment?”. It turned out that in the HARM condition, 82% of the respondents attributed intentionality of causing the side effect. In the HELP condition, on the other hand, only 23% of respondents believed that the environment was helped intentionally.

The results of this experiment have been repeatedly confirmed (Knobe 2004b, 2004a; Nadelhoffer 2004b, 2004a; Mele and Cushman 2007) also in studies on children (Leslie, Knobe, and Cohen 2006) and with adults with high functioning autism or Asperger’s syndrome (Zalla and Leboyer 2011) as well as in various languages, including Hindi (Knobe and Burra 2006), German (Dalbauer and Hergovich 2013), Polish (Kuś and Maćkiewicz 2016; Waleszczyński, Obidziński, and Rejewska 2018).

4. Structure of the story and the problem of intention

Knobe’s story is based on a predefined structure. The CEO is only concerned with A (the main effect) and does not care about B (the side effect). In the HARM condition the side effect (B1) is negative, and in the HELP condition the side effect (B2) is positive. It does not result from the structure of the story itself that there is any difference between the likelihood of effects A or B, or that of effects B1 and B2. And yet, studies have shown (Nakamura 2018) that the kind of side effect (positive, negative) alone determines the attribution of likelihood. Negative effects are considered more likely than positive ones. From this perspective, the difference between the two stories is not only related to the content of the story itself, but also to expectations concerning the likelihood of a side effect. This suggests that certain notions and descriptions which refer to (morally) positive and negative effects are related not only to the purely descriptive function of the language, but also to its evaluative and normative one. Authors studying the Knobe effect point out that it is not so much the difference on the descriptive plane of the language, but that on the evaluative and normative one that causes the Knobe effect.

The problem of intention plays an important role in the problem of intentionality. Therefore, it is necessary to analyze the attitude of the agent to a particular side effect. This attitude may be threefold: (1) I want it; (2)

I do not want it; (3) I do not care about it. In the last case, the attitude of the agent may be treated as indirect intention or consent for a particular side effect (Głowala 2013). There is also a fourth special case that should be considered, where the agent does not know about some possible consequences of his or her actions. More specifically, this describes a situation when the agent does not foresee a particular effect. In such a situation, the agent cannot be regarded as having intentionally caused that effect, since he or she did not expect to cause such an effect when taking the action, and thus could not have had the intention of causing it. Consider the following example. A and B are playing ball, throwing it to each other. While they are playing, a factor distracting the attention of B appears as a firefighting siren. Consequently, B is hit by the ball thrown by A, falls to the ground, and hurts himself on the head. Did A foresee such a complex coincidence? He did not. We may say that he had the intention to play ball with B, but not the intention to hurt his head.

The above distinction refers to the issue of intention. The situation is somewhat different, however, when we take into account the existence of obligations, that is, the influence of the normative factor on judging the intentionality of the agent, which needs to be determined in order to decide whether the agent is to be blamed or not. In such a situation, the attribution of obligation burdening the agent creates a fourth attitude to the side effect (4) (non-intentionally) causing an effect, for example by negligence or failure to act. This happens, for example, when a steersman leaves the wheel, goes below deck to play cards, and the ship is wrecked. In such circumstances, even if he did not expect any danger to occur, he will be charged with intentional default on a duty. In other words, the agent's attitude is referred to the (intentional) negligence to fulfill an obligation which resulted in a particular side effect (the main effect is playing cards, the side effect is the shipwreck). In situation (4), however, the intention is referred not so much to causing an effect (since the agent did not foresee effect [3]), but rather to violating a norm resulting from a duty. In some explanations of the Knobe effect (Holton 2010; Paprzycka 2015), it is argued that the attribution of intentionality results precisely from the attribution of the intention to violate a norm, or the intention to fulfill a (social) norm. Such a simplification seems to be unsubstantiated, however, as reference to

a norm helps evaluate the (moral or legal) responsibility for causing an effect, rather than the intentionality of causing it, even if we consider the side effect to have been foreseeable (3). The story used in Knobe's experiment cannot be referred directly to situation (4), since it results from the structure of his story that the agent foresaw the occurrence of a particular effect. Consequently, it is not necessary to refer to a norm in order to attribute intentionality or to judge a side effect as having been caused intentionally. Unless we assume, in line with what is argued by Holton and Paprzycka, that the power of a (social) norm affects the attribution of intentionality to such an extent that it overrides (suspends) the intention explicitly stated in the analyzed stories. It should also be assumed that the attribution of intentionality is secondary to the attribution of responsibility; this, however, would stand in opposition to other studies (Nadelhoffer 2004c) which have shown that causation of a negative side effect and attribution of responsibility does not need to be coupled with attribution of intentionality.

Studies by Waleszczyński, Obidziński and Rejewska (2019) have shown, however, that the thesis about the key significance of the norm for the attribution of intentionality in causing a side effect in the Knobe effect does not apply in one particular case—namely, when there is a significant difference¹ in weight between the evaluative properties of the main effect and the evaluative properties of the side effect. This does not mean that (social) norms do not influence the attribution of intentionality in causing a side effect. Their impact is indirect, however, and relates to the determination of the evaluative-normative meaning of a particular situation and the caused effects. Thence the importance of studies which aimed at a more precise identification of the role of intentions in the attribution of intentionality in causing a side effect.

¹ Assessment of this difference is problematic. Authors of this article make it based on prevailing convictions about the value of particular effects as seen against other effects. In the stories used in the experiments, the emphasis was mainly on the relationship between the moral weight of the main effect and the moral weight of the side effect.

5. The influence of moral factors—research findings to date and solutions

The observed side-effect effect has been explained by Knobe (Knobe 2004a, 2006) as the influence of the moral valuation of side effects on the judgment of intentionality in causing them. In his opinion, people have a tendency to attribute intentionality when the side effect is negative. The attribution of intentionality in such circumstances may be influenced by moral factors. Studies by Hindriks (Hindriks et al. 2016) show that the influence of moral valuation is only observable in the HARM condition and does not explain judgments made in the HELP condition. Hindriks's comment is significant in that it shows that we should not refer to the influence of moral valuation in general, but only to the influence of (morally) negative effects on the attribution of intentionality. In other words, only a part of moral reality may influence judgments concerning intentionality. Negative moral valuation may be the very factor which influences the occurrence of asymmetric judgments in attributing the intentionality of causing a side effect.

Some of the hypotheses proposed so far have linked the occurrence of asymmetry in the attribution of intentionality in causing a side effect with a praise-blame asymmetry (praise-blame asymmetry) (Malle and Nelson 2003; Nadelhoffer 2004a, 2006b, 2006a; Nado 2008; Hindriks 2008, 2011), and with the attribution of responsibility for the caused negative side effects (Wright and Bengson 2009). These explanations only explain the HARM condition in the Knobe effect, and make the attribution of intentionality dependent on the attribution of blame or responsibility. Other hypotheses have tried to explain the side-effect effect by referring to social norms, which in the story proposed by Knobe would concern the prohibition of harming the environment. It is the intentional violation of such a norm (Holton 2010), or its intentional neglect (Paprzycka 2015, 2016), that is supposed to determine the attribution of intentionality in causing a side effect. An important question that may be asked concerns the character of the norm to which the respondents are supposed to be referring to. Is it a legal, social, or moral norm, and does the character of this norm matter? Finding an answer to these questions would be relevant to the hypothesis regarding the influence of evaluative properties on the attribution of intentionality. If the violated norm did not

need to be a moral, but could for example be a legal one, one could hardly claim that moral norms affect the attribution of intentionality.

There are also other explanatory hypotheses which point to the relationship between the good or the evil of side effects and the values to which the agent refers when making a judgment (Sripada 2010, 2012; Sripada and Konrath 2011). The values and attitudes endorsed by the agent form the structure of his or her deep Self. Consequently, the respondents who attribute anti-environmental values to the CEO also attribute intentionality in the HARM condition. This hypothesis also naturally allows for a gradation in the attribution of intentionality. As pointed out by Hindriks (2016, 215). However, this hypothesis is symmetrical, since when it is interpreted in categorical terms, it only explains the HELP condition, and when it is interpreted in graded terms, it only explains the HARM condition. It may be observed, however, that in both interpretations, the role of intentions and of the reference to the side effect itself are secondary. What moves to the forefront is the attribution of a particular structure of values to the agent.

A different explanation of the Knobe effect is proposed by Hindriks. He points out that the CEO is attributed a degree of indifference to the side effect he causes. On this basis, Hindriks proposes his Normative Reason Hypothesis. He suggests the existence of a normative reason which is the obligation to care about the side effect caused. The attribution of indifference to the agent is gradable and may extend from the attitude of neutrality to that of full care. Hindriks's hypothesis does not so much refer to intentions as it does to the attitude (care) of the agent to (about) the side-effect he causes. It is the degree of this attitude that is supposed to explain the attribution of intentionality in causing a side effect. The problem of this explanation lies on the theoretical side, however. It assumes that no normative reason exists for positive side effects. In other words, it presents a reality in which people are supposed to care about negative side effects, but are not required to care about positive ones.

The findings made so far are complicated by the results of studies which show that it is possible for intentionality of causing a side effect to be attributed in the HELP condition although the agent does not care about the side effect. Waleszczyński, Obidziński and Rejewska (2019) have pointed out that the influence on the attribution of intentionality in causing a side

effect may depend not so much on the moral weight of the side effect, but on the relationship between the moral weight of the main effect and that of the side effect. Their studies show that if in Knobe's story we swap the main effect for the side effect, then despite a statistical difference in answers provided in the two conditions, they become symmetrical and reveal a tendency in the attribution of intentionality.² Such experimental findings undermine most of the hypotheses explaining the Knobe effect proposed so far. The attribution of intentionality in causing a side effect in the HELP condition can no longer be explained by the attribution of blame, as suggested, e.g., by Malle and Nelson, Nadelhoffer, Nado, or Hindriks. In addition, as pointed out by Holton, it would be difficult to demonstrate the existence of a norm which could be violated in the HELP condition. Also, focusing on the agent's attitude to the side effect and duty to care about the side effect will not help much in explaining the observed relations. Similarly, the standard understanding of intentionality does not help explain the observed result, as the agent is indifferent to the expected side effect.

² The results below present data combined from two experiments downloaded from: <https://osf.io/ky3re/>. They have been combined for the purposes of further comparative analyses. Waleszczyński, Obidziński and Rejewska presented a study of the three types of the stories.

The first story was taken from the Knobe study (2003: 191). Results: $M_{\text{Harm}} = 1.645$, $M_{\text{Help}} = -1.113$, $t(122) = 7.552$, $p < 0,001$, $d = 0.864$.

The second one, was the original story created after the pattern of the Knobe's story: "The Deputy of Experimental Oncology Hospital asks the director: "We can produce a drug that will heal patients with pancreatic cancer, but it would cause pneumonia/and cure pneumonia." The director responds: "I want to primarily cure patients with pancreatic cancer. We start production and give medicine to patients. The drug has been given and has caused/cured pneumonia." Results: $M_{\text{Harm}} = 0.603$, $M_{\text{Help}} = -0.587$, $t(124) = 3.195$, $p = 0,002$, $d = 0.317$.

The third story was the modification of the Knobe's story: "The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help the environment, but it will also increase/decrease profits.' The chairman of the board answered 'I don't care at all about profits. I just want to help the environment. Let's start the new program.' They started the new program. Sure enough, the profits were increase/decreases." Results: $M_{\text{Harm}} = 0.952$, $M_{\text{Help}} = 0.274$, $t(122) = 1.997$, $p = 0,048$, $d = 0.359$.

Joshua Shepherd (2012) presented studies that broadly analyzed the impact of the relationship between the main and side effects and the agent's attitude to said side effects. However, the appearance of new interesting data in studies on the Knobe effect makes it necessary not only to reconsider the significance of the relationship between the main effect and the side effect, but also to reconsider the significance of the role of intention in the attribution of intentionality, or more generally, the attitude of the agent to the side effect caused. Therefore, in our study we have focused on the significance of intention, or more precisely, lack of intention, in the attribution of intentionality in causing a side effect, and the importance of moral evaluation of the relationship between the main and side effects in attributing intentionality to causing a side effect.

6. The experiment

The experiment employed modified versions of three scenarios used in experiments performed by Waleszczyński, Obidziński and Rejewska, diversified in terms of the evaluative properties of the effects. In order to see what role for the attribution of intentionality in causing a side effect is played by the agent's attitude to that side effect, the stories had been modified so that it could not be inferred what the attitude of the agent was to the side effect he caused. The investigated hypothesis assumed that if it is not possible to determine the attitude of the agent to causing a side effect, and thus to know his intention, then the model response would be "Hard to say." The experiments were aimed at verifying this hypothesis. Removing the agent's attitude to causing a side effect from the experiment will complement Shepherd's (2012) research and will also verify the attributing intentionality account proposed by Cova, Dupoux and Jacob (2012).

Verifying the hypothesis is crucial for the direction of further research. In philosophical conceptions, the intention of an action or the prediction of that action determines an intentional action. However, taking the example of responsibilism, we have indicated that the concept of intentional action (causing an effect), which uses the concept of intention, can be said in a way that is secondary to, for example, previously ascribed responsibility. Therefore, in our study, we pose only one question concerning the intentionality

of causing a side effect in a situation where the respondents do not know either the intention or the attitude of the acting subject towards the side effect caused. Ascribing or not ascribing intentionality to inducing a side effect would mean that additional factors that influence the ascribing of intentionality to induce side effects, beyond those of intention and prediction, should be sought.

6.1. Methodology

The experiment was performed at two locations: the Cardinal Stefan Wyszyński University in Warsaw among students of various faculties, and at Warsaw Metro stations among randomly selected passengers. For the purposes of this experiment, results from both locations have been combined, and all analyses are performed on the resulting samples. Every quizzed person responded only once, after reading one story in one of the study conditions. Thus, the sum total of all respondents was 372.

In the experiment described above, the following stories and questions were used:

S1_Help

The vice-president of a company went to the chairman of the board and said, ‘We are thinking of starting a new program. It will help us increase profits, but it will also help the environment.’ The chairman of the board answered ‘I just want to make as much profit as I can. Let’s start the new program.’ They started the new program. Sure enough, the environment was helped.

Question: Did the chairman intentionally help the environment?

S1_Harm

The vice-president of a company went to the chairman of the board and said, ‘We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment.’ The chairman of the board answered ‘I just want to make as much profit as I can. Let’s start the new program.’ They started the new program. Sure enough, the environment was harmed.

Question: Did the chairman intentionally harm the environment?

S2_Help

The vice-president of an experimental oncological hospital went to the chairman of the board and said, 'We are thinking of starting the production of a new medicine. It will help us cure patients of pancreatic cancer but it will also cure them of pneumonia.' The chairman of the board answered, 'I just want to cure patients of pancreatic cancer. Let's start the production of a new medicine.' They started the production of the new medicine. Sure enough, the patients were cured of pneumonia.

Question: Did the chairman intentionally cure pneumonia?

S2_Harm

The vice-president of an experimental oncological hospital went to the chairman of the board and said, 'We are thinking of starting the production of a new medicine. It will help us cure patients of pancreatic cancer but it will also cause pneumonia.' The chair-man of the board answered, 'I just want to cure the patients of pancreatic cancer. Let's start the production of a new medicine.' They started the production of the new medicine. Sure enough, the patients came down with pneumonia.

Question: Did the chairman intentionally cause pneumonia?

S3_Help

The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us help the environment, but it will also increase profits.' The chairman of the board answered, 'I just want to help the environment as much as I can. Let's start the new program.' They started the new program. Sure enough, profits were increased.

Question: Did the chairman intentionally increase profits?

S3_Harm

The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us help the environment, but it will also cause losses.' The chairman of the board answered, 'I just want to help the environment as much as I can. Let's start the new program.' They started the new program. Sure enough, losses were caused.

Question: Did the chairman intentionally cause losses?

Responses were given on a seven-point scale ranging from -3 to 3, where -3 meant “Absolutely not,” 0 meant “Hard to say,” and 3 meant “Absolutely yes.” Participants responded using the classic paper-and-pencil method.

6.2. Results³

Tests for distribution normality from the results of our study show that all distributions are significantly different from the normal distribution. Thus, we have started our statistical analysis using nonparametric tests of differences. However, after the first analysis, we made another one using a parametric test and compared their results. We have found that obtained results are similar in both significance and effect size. Therefore, in the analysis report, we present the results of parametric t-tests.

To test our hypothesis, we have used a one-sample t-Student test that allows us to compare collected data with assumed values (e.g. from previous data or data based on a theoretical approach). We have compared observed means with scale grade 0 (“Hard to say”) testing if there was a significant difference between the results of our studies and the value that is the model response of the tested hypothesis. There was a statistically significant difference in the distribution of results compared to standard normal distribution. Results of the one sample Wilcoxon test corresponded to the results of t-Student test results though. For this reason, we decided to present the results using a parametric test. Descriptive statistics and test results are presented in Table 1.

	<i>M</i>	<i>SD</i>	<i>t</i>	<i>df</i>	<i>p</i>	<i>Cohen's d</i>
<i>S1_Harm</i>	2.2	1.277	13.887	64	< 0.001	1.723
<i>S1_Help</i>	-0.754	1.723	-3.527	64	< 0.001	0.298
<i>S2_Harm</i>	-0.613	2.059	-2.343	61	0.022	0.283
<i>S2_Help</i>	-0.661	2.172	-2.397	61	0.02	0.438
<i>S3_Harm</i>	0.565	1.997	2.266	61	0.03	0.304
<i>S3_Help</i>	0.629	1.84	-2.692	61	0.009	0.342

Table 1. The difference between mean results and scale grade 0

³ The results can be downloaded from the repository at: (Obidziński and Waleszczyński 2021) <https://osf.io/8c5qk>.

The difference between mean results in the HARM and HELP conditions were then compared for each story. The findings are presented in Table 2.

	M_{Harm}	SD_{Harm}	M_{Help}	SD_{Help}	t	Df	p	<i>Cohen's d</i>
<i>S1</i>	2.2	1.277	-0.754	1.723	11.102	118.01	< 0.001	1.153
<i>S2</i>	-0.613	2.059	-0.661	2.172	0,127	121.65	0.899	-
<i>S3</i>	0.565	1.997	0.629	1.84	-0,187	121.19	0.852	-

Table 2. Comparison of the HARM and HELP conditions

The analyses show that results for all three stories both in the HARM and the HELP condition differ significantly from 0, meaning “Hard to say.” A statistical significance in differences between conditions in each study was only observed for story N1, corresponding to the standard story used in the Knobe experiment. Effect sizes for the observed differences were: high for the difference between conditions in story S1, and between 0 and the mean *S1_Harm* result; and low for the other statistically significant differences.

Finally, in order to see whether the manipulation performed on the study material (modified scenarios) significantly affected the results obtained in the experiment, differences have been analyzed between results in the corresponding stories of the above experiment and studies performed by Waleszczyński, Obidziński, and Rejewska (2019)⁴ designated in the Table as “Old”—separately for the HARM and HELP conditions. Table 3 presents the results of analyses performed for the t-Student test on two independent samples.

	M_{Old}	SD_{Old}	M_{New}	SD_{New}	t	Df	p	<i>Cohen's d</i>
<i>S1_Harm</i>	1.645	1.9	2.2	1.277	1.94	125	0,055 ^T	0.102
<i>S1_Help</i>	-1.113	2.159	-0.754	1.723	1.038	125	0.301	-
<i>S2_Harm</i>	0.603	2.174	-0.613	2.059	-3.21	123	0.002	0.314
<i>S2_Help</i>	-0.587	2.005	-0.661	2.172	-0.198	123	0.843	-
<i>S3_Harm</i>	0.952	1.741	0.565	1.997	-1.151	122	0.252	-
<i>S3_Help</i>	0.274	2.026	0.629	1.84	3.447	122	0.001	0.619

T—statistical tendention

Table 3. Comparison of results before and after story modification

⁴ The results can be downloaded from the repository at: (Obidziński and Waleszczyński 2019) <https://osf.io/ky3re>.

The analyses have revealed two significant differences between the results of studies using stories before and after modification. One difference concerns story S2 in the HARM condition—the effect size for this difference is low. The other difference concerns story S3 in the HELP condition—it is characterized by a medium effect size. One statistical tendency has also been observed concerning story S1 in the HARM condition—however, the very low effect size suggests that this potential difference between the studied groups is insignificant.

6.3. Discussion

In his experiment, Knobe (2003) used a typical structure of a situation analyzed using the Doctrine of Double Effect to evaluate responsibility for causing a side effect. Person X takes an action to achieve a particular goal, which in Knobe's experiment is increasing the company's profits. In order to achieve the intended (main) goal A, a side-effect B is caused. Knobe introduced what turned out to be a significant modification by creating an alternative situation where the side effect is positive and by asking an appropriate question. The question concerned not so much responsibility for causing the side effect, but the intentionality of causing it. It is also significant that it is clear from the story that X does not care about the side effects. Results of the experiment proved to be very interesting and became problematic for the understanding of the notion of intentional action. The greatest difficulty concerns results in the HARM condition, as they suggest that respondents attribute intentionality in causing the side effect even though the agent says he does not care about side effects. In other words, the problem consists in that intentionality is attributed even though the agent does not have the intention of causing a particular effect.

Our study presents results which provide a new perspective on Knobe's experiment. We used Knobe's story, designated as S1, where the part "I do not care that it will help [harm] the environment" has been removed. The goal was to see if once the explicit statement of the agent's intentions with regard to the expected side effect is removed, the attribution of intentionality in causing this effect will be significantly affected. The proposed structure of story S1 does not provide information on the intention to cause a side effect, and consequently the expected result was the answer "Hard to

say.” The results are presented in Tables 1 and 2. First, it turned out that responses for the HARM and HELP conditions significantly differ statistically from the expected response “Hard to say,” designated as “0”. Second, responses for the S1_Harm and the S1_Help condition proved to be asymmetrical, and the difference between them is also statistically significant. The study shows that concealing the agent’s intention as regards causing the side effect does not significantly influence the asymmetry in responses previously observed by Knobe. This is also confirmed by a comparison with the earlier (*Old*) studies presented in Table 3 for the S1_Harm and S1_Help conditions.

In the context of the above data concerning story S1, new data is provided by responses concerning stories S2 and S3. They are modified stories taken from studies by Waleszczyński, Obidziński and Rejewska (2019). First, in both stories, responses for both conditions are significantly statistically different from the “Hard to say” response. This coincides with the results for story S1. This would confirm the conclusion that revealing or concealing the intention is not the only, or the most important, element affecting the respondents’ attribution of intentionality in causing a side effect. Second, unlike those in story S1, the differences between HARM and HELP conditions in stories S2 and S3 are not statistically significant (Table 2). On the contrary, the responses are largely consistent and symmetrical. It is interesting that the symmetry in story S2 points to the non-intentionality of causing a side effect, and in story S3—to its intentionality. This means that for three different stories based on the same scheme, three different results have been received. They seem to undermine the so-called Norm Violation Hypothesis (Holton 2010; Paprzycka 2015) which says that the attribution of intentionality is caused by attributing the intention to violate or neglect to conform to a particular norm. It would be difficult to find a universally valid norm which prohibits the increase of profits in companies which help the environment (S3_Help). The Norm Violation Hypothesis may be upheld for the HARM condition, but it would be problematic in this hypothesis to identify a norm for the S2-Harm condition. This results from an analysis of the earlier (*Old*) and present (*New*) studies concerning this scenario which are significantly different statistically and asymmetrical at the same time.

However, the Normative Reason Hypothesis will also have trouble interpreting the results of these studies. If the normative reason is the duty to

care about causing a side effect, which relates to the attribution of intentionality, then why does it appear in both conditions in stories S3, and in neither of them in story S2? It seems that neither of these two cases can be reconciled with the hypothesis proposed by Hindriks, Douven and Singmann (Hindriks et al. 2016).

Still other consequences arise from the results of the experiment carried out for the account proposed by Cova, Dupoux and Jacob (2012). Their account assumes three meanings for the term intentionality, of which one or the other are preferred depending on the situation. Therefore, when asking about the intentionality of causing a side effect, one should take into account the attitude of the agent to the expected effect and the (social) expectation regarding the attitude of the agent and possibly the skills of the agent. The account given by them does not work in the case of our experiment, because in the analyzed history there is not the necessary data to judge on intentionality, and yet the respondents do it. Achieving a result similar to “Hard to say” would mean that their account using a changed meaning of intentionality is correct. However, this is not what occurs, because all the results are statistically significantly different from the expected response.

The results of our experiment support the hypothesis, however, that the issue of key importance for the attribution of intentionality in causing a side effect is that of the evaluative properties of the relationship occurring between the evaluative properties of the main effect and the evaluative properties of the side effect (Waleszczyński, Obidziński, and Rejewska 2019). This conclusion is based on the fact that when the structure of the story is the same, but the evaluative properties of the main effect and the side effect are modified, three different response patterns are observed for three different stories. Consequently, this leads to the conclusion that it is the (morally positive or negative) evaluative properties of the main and side effects that significantly affect the attribution of intentionality in causing a side effect.

When analyzing the significance of concealing the intentions, or more precisely, of the agent’s indifference to the expected side effect, two statistically significant changes and one tendency have been observed compared to the Waleszczyński, Obidziński and Rejewska (2019) study. One change and the tendency concern an increase in the attribution of intentionality in causing a side effect. The statistically significant change concerns the

S3_Help scenario, and is characterized by a medium effect size. Its influence is reflected in the fact that responses in the S3_Help scenario and in the S3_Harm scenario are similar. This may mean that in a situation when the main effect (helping the environment) has a high positive value, and the side effect (company profit/loss) is not significant (has a low moral value) compared to the main effect, and we do not know the intentions of the agent with regard to the side effect, respondents tend to attribute intentionality in causing a side effect both in the situation of help and that of harm. The observed statistical tendency in the attribution of intentionality concerns the HARM condition in the S1 scenario. It shows a tendency to attribute intentionality in causing a negative effect, that is, harming the environment (high moral value) when the main effect is the desire to increase company profits (low moral value).

The change most difficult to interpret is that in the S2_Harm condition, which is not only statistically significant, but which is also the only asymmetrical change in the experiment compared to the Waleszczyński, Obidziński, Rejewska study. Concealing the agent's attitude to causing a negative side effect (medium moral value) which is pneumonia, with the main effect (high moral value) of curing pancreatic cancer, indicates non-intentionality in causing it.

An analysis of the experiments performed so far supports the following tendency patterns in the attribution of intentionality in causing a side effect in view of the relationship between the main effect and the side effect. Table 4 presents a situation when the agent's intention/attitude to the side effect is unknown.

<i>L.p.</i>	<i>Main goal</i>		<i>Side effect</i>	<i>Causing a side effect</i>	
1.	low-value	positive	high-value	positive negative	Unintentionally Intentionally
2.	high-value	positive	medium-value	positive negative	Unintentionally Unintentionally
3.	high-value	positive	low-value	positive negative	Intentionally Intentionally

Table 4. Tendency patterns in the attribution of intentionality in causing a side effect

The above Table shows a tendency in the attribution of intentionality based on the relationship between the main and the side effect. We can see that the asymmetry in responses appears only in relationship no. 1, found in Knobe's story. The symmetry and the discrepancy in results between the relationships in scenarios no. 2 and no. 3 does not allow us to conclude that there is an absolute moment of judging about the intentionality of causing a side effect. In other words, we cannot say that if we do not know the intentions or the attitude of the agent to the side effect, we cannot claim that such effect is always caused unintentionally.

7. Summary

Studies on the common understanding of intentionality show that it is a complex and multi-threaded problem which requires further in-depth studies. Particularly interesting and fruitful are studies on the so-called side-effect effect observed by Knobe. Experiments and analyses presented in this article were aimed at contributing new knowledge about the attribution of intentionality in causing a side effect, in particular the role played by intention. First, they have shown that the very originality of the story used by Knobe (2003) and the relationship between the evaluative properties of the main effect and the evaluative properties of the side effect results in an asymmetry of responses and contributes to the occurrence of the so-called side-effect effect. On the one hand, this means that the relationship between the main and the side effect significantly affects the so-called side-effect effect. On the other hand, it shows the role played by intention in the attribution of intentionality in causing a side effect. Second, using the story structure proposed by Knobe and concealing the agent's intention to the expected side effect is not in itself enough to obtain reproducibility of responses, that is, a predictable pattern in the common application of the notion of intentionality, which has been unequivocally shown by the discrepancy in the results of the studies concerning each scenario (S1, S2, S3). Considering earlier studies (Waleszczyński, Obidziński, and Rejewska 2019) and the results of the studies presented in this article, it may be concluded that a change in the evaluation of the relationship between the evaluative

properties of the main and the side effect significantly affects the attribution of intentionality in causing the side effect.

The studies and analyses have shown that a significant impact on the occurrence of the Knobe effect has the story itself and the type of the main and the side effect, or to be more exact, the specific relationship between these effects. If the relationship between the two types of effect is significant for the attribution of intentionality in causing a side effect, then it must be a necessary condition for the evaluation of the intentionality in causing that side effect. Consequently, this means that there is a difference between the conditions of attributing intentionality in causing effect A as the main effect of action X, and the conditions of attributing intentionality in causing effect B, which is a side effect of action X. In this situation, the question reappears about the role of moral factors and their possible impact on the occurrence of the Knobe effect. At this stage of research, this cannot be unequivocally established, as we cannot precisely determine what influences the significance of the relationship between the two effects for the attribution of intentionality in causing a side effect. Even if we identify the evaluative properties of an effect, we must still face the dispute as to whether they still have the nature of description, or whether it is already that of moral judgment. In that case, we would be dealing with a purely meta-ethical dispute.

The analyses performed so far concerning the occurrence of an asymmetry in the attribution of intentionality in causing a side effect have focused mainly on explaining the attribution of intentionality in the HARM condition. At the same time, it has been assumed that in the HELP condition, all conditions for the attribution of intentionality are met. Not enough attention has been paid, however, to the fact that standard accounts of intentionality define conditions for the occurrence of a single predicted and identified effect. This means that one action causes one effect A. In such case, it is enough to check the intention of causing a particular effect in order to attribute intentionality of causing it. However, in analyses concerning the attribution of intentionality in causing a side effect, a significant additional element has been omitted. The occurrence of a side effect is conditioned by the occurrence of the main effect. This means that one action causes two effects A and B, where effect B is a derivative of effect A. If a side effect occurs, at least two additional elements appear with respect to

standard accounts of intentionality. These are: the occurrence of effect B, and the relationship between effects A and B, representing a significant change in the conditions of the analyzed situation.

We may ask whether observation of the side-effect effect materially affects the standard accounts of intentionality. The answer: no. The effect observed by Knobe concerns the attribution of intentionality in causing a side effect, and not the attribution of intentionality as such. Based on existing research, we may conclude that if one action causes one effect A, then in order to determine the intentionality of causing this effect, it is enough to check the intention of the agent with regard to this effect. When one action causes two effects A and B though, where B is a derivative of A, then in order to attribute intentionality in causing effect B it is necessary to check the intention of causing effect B and at least one of the evaluative properties of the relationship between A and B. Now the problems would be, first, how to identify the evaluative properties of the relationship between a particular A and a particular B, and, secondly, how the evaluative properties influence the attribution of intentionality in causing a side effect.

References

- Adams, Frederick. 1986. "Intention and Intentional Action: The Simple View." *Mind & Language* 1: 281–301. <https://doi.org/10.1111/j.1468-0017.1986.tb00327.x>
- Adams, Frederick, and Annie Steadman. 2004a. "Intentional Action and Moral Considerations: Still Pragmatic." *Analysis* 64: 268–76. <https://doi.org/10.1111/j.0003-2638.2004.00496.x>
- Adams, Frederick, and Annie Steadman. 2004b. "Intentional Action in Ordinary Language: Core Concept or Pragmatic Understanding?" *Analysis* 64: 173–81. <https://doi.org/10.1093/analys/64.2.173>
- Cova, Florian, Emmanuel Dupoux, and Pierre Jacob. 2012. "On Doing Things Intentionally." *Mind & Language* 27: 378–409. <https://doi.org/10.1111/j.1468-0017.2012.01449.x>
- Dalbauer, Nikolaus, and Andreas Hergovich. 2013. "Is What is Worse More Likely? —The Probabilistic Explanation of the Epistemic Side-Effect Effect." *Review of Philosophy and Psychology* 4: 639–57. <https://doi.org/10.1007/s13164-013-0156-1>

- Głowała, Michał. 2013. "On Indirectly Intended Consequences of Action and Relinquishment. G.E.M. Anscombe and the Spanish Thomists of XVI/XVII Century." *Etyka* 46: 7–20.
- Guglielmo, Steve, and Bertram F. Malle. 2010. "Can Unintended Side Effects Be Intentional? Resolving a Controversy Over Intentionality and Morality." *Personality and Social Psychology Bulletin* 36: 1635–47.
<https://doi.org/10.1177/0146167210386733>
- Hart, Herbert L. A. "1949. XI.—The Ascription of Responsibility and Rights." *Proceedings of the Aristotelian Society* 49: 171–94. <https://doi.org/10.1093/aristotelian/49.1.171>
- Hindriks, Frank. 2008. "Intentional Action and the Praise Blame Asymmetry." *Philosophical Quarterly* 58: 630–41. <https://doi.org/10.1111/j.1467-9213.2007.551.x>
- Hindriks, Frank. 2011. "Control, Intentional Action, and Moral Responsibility." *Philosophical Psychology* 24: 787–801.
<https://doi.org/10.1080/09515089.2011.562647>
- Hindriks, Frank. 2014. "Normativity in action: How to explain the knobe effect and its relatives." *Mind and Language* 29: 51–72.
<https://doi.org/10.1111/mila.12041>
- Hindriks, Frank, Igor Douven, and Henrik Singmann. 2016. "A New Angle on the Knobe Effect: Intentionality Correlates with Blame, not with Praise." *Mind and Language* 31: 204–20. <https://doi.org/10.1111/mila.12101>
- Holton, Richard. 2010. "Norms and the Knobe Effect." *Analysis* 70: 417–24.
<https://doi.org/10.1093/analys/anq037>
- Knobe, Joshua. 2003. "Intentional Action and Side Effects in Ordinary Language." *Analysis* 63: 190–94. <https://doi.org/10.1093/analys/63.3.190>
- Knobe, Joshua. 2004a. "Folk Psychology and Folk Morality: Response to Critics." *Journal of Theoretical and Philosophical Psychology* 24: 270–79.
<https://doi.org/10.1037/h0091248>
- Knobe, Joshua. 2004b. "Intention, Intentional Action and Moral Considerations." *Analysis* 64: 181–87. <https://doi.org/10.1093/analys/64.2.181>
- Knobe, Joshua. 2006. "The Concept of Intentional Action: A Case Study in the Uses of Folk Psychology." *Philosophical Studies* 130: 203–31.
<https://doi.org/10.1007/s11098-004-4510-0>
- Knobe, Joshua, and Arudra Burra. 2006. "The folk concepts of intention and intentional action: A cross-cultural study." *Journal of Cognition and Culture* 6: 113–32. <https://doi.org/10.1163/156853706776931222>
- Knobe, Joshua. 2016. "Experimental Philosophy Is Cognitive Science." In *A Companion to Experimental Philosophy*, edited by Justin Sytsma, and Wesley

- Buckwalter, 37–52. Wiley-Blackwell.
<https://doi.org/10.1002/9781118661666.ch3>
- Knobe, Joshua, and Shaun Nichols. 2008. “An Experimental Philosophy Manifesto.” In *Experimental Philosophy*, edited by Joshua Knobe, and Shaun Nichols, 3–16. Oxford: Oxford University Press.
- Kuś, Katarzyna, and Bartosz Maćkiewicz. 2016. “Z rozmysłem, ale nie specjalnie. O językowej wrażliwości filozofii eksperymentalnej.” *Filozofia Nauki* 24: 73–102.
- Leslie, Alan M., Joshua Knobe, and Adam Cohen. 2006. “Acting Intentionally and the Side-Effect Effect.” *Psychological Science* 17: 421–27.
<https://doi.org/10.1111/j.1467-9280.2006.01722.x>
- Malle, Bertram F., and Joshua Knobe. 1997. “The folk concept of intentionality.” *Journal of Experimental Social Psychology*, 33(2): 101–121.
<https://doi.org/10.1006/jesp.1996.1314>
- Malle, Bertram F., and Sarah E. Nelson. 2003. “Judging Mens Rea: The Tension Between Folk Concepts and Legal Concepts of Intentionality.” *Behavioral Sciences and the Law*, 21(5): 563–80. <https://doi.org/10.1002/bsl.554>
- McCann, Hugh J. 1987. “Rationality and the Range of Intention.” *Midwest Studies in Philosophy* 10: 191–211. <https://doi.org/10.1111/j.1475-4975.1987.tb00540.x>
- Mele, Alfred R. 1992. “Recent Work on Intentional Action.” *American Philosophical Quarterly* 29: 199–217. <https://doi.org/10.2307/20014416>
- Mele, Alfred R., and Fiery Cushman. 2007. “Intentional action, folk judgments, and stories: Sorting things out.” *Midwest Studies in Philosophy* 31: 184–201.
<https://doi.org/10.1111/j.1475-4975.2007.00147.x>
- Mele, Alfred R., and Steven Sverdluk. 1996. “Intention, intentional action, and moral responsibility.” *Philosophical Studies* 82: 265–87.
<https://doi.org/10.1007/BF00355310>
- Nadelhoffer, Thomas. 2004a. “On Praise, Side Effects, and Folk Ascriptions of Intentionality.” *Journal of Theoretical and Philosophical Psychology* 24: 196–213.
<https://doi.org/10.1037/h0091241>
- Nadelhoffer, Thomas. 2004b. “The Butler Problem Revisited.” *Analysis* 64(3): 277–84. <https://doi.org/10.1111/j.0003-2638.2004.00497.x>
- Nadelhoffer, Thomas. 2004c. “Blame, Badness, and Intentional Action: A Reply to Knobe and Mendlow.” *Journal of Theoretical and Philosophical Psychology* 24: 259–69. <https://doi.org/10.1037/h0091247>
- Nadelhoffer, Thomas. 2005. “Skill, luck, control, and intentional action.” *Philosophical Psychology* 18: 341–52. <https://doi.org/10.1080/09515080500177309>
- Nadelhoffer, Thomas. 2006a. “Bad Acts, Blameworthy Agents, and Intentional Actions: Some Problems for Juror Impartiality.” *Philosophical Explorations* 9: 203–19. <https://doi.org/10.1080/13869790600641905>

- Nadelhoffer, Thomas. 2006b. "Desire, Foresight, Intentions, and Intentional Actions: Probing Folk Intuitions." *Journal of Cognition and Culture* 6: 133–57. <https://doi.org/10.1163/156853706776931259>
- Nadelhoffer, Thomas, and Eddy Nahmias (2007). "The Past and Future of Experimental Philosophy." *Philosophical Explorations* 10(2): 123–49. <https://doi.org/10.1080/13869790701305921>
- Nado, Jennifer. 2008. "Effects of Moral Cognition on Judgments of Intentionality." *British Journal for the Philosophy of Science* 59: 709–31. <https://doi.org/10.1093/bjps/axn035>
- Nakamura, Kuninori. 2018. "Harming is more intentional than helping because it is more probable: the underlying influence of probability on the Knobe effect." *Journal of Cognitive Psychology* 30: 129–37. <https://doi.org/10.1080/20445911.2017.1415345>
- Nichols, Shaun, and Joseph Ulatowski. 2007. "Intuitions and Individual Differences: The Knobe Effect Revisited." *Mind and Language* 22: 346–65. <https://doi.org/10.1111/j.1468-0017.2007.00312.x>
- Obidziński, Michał, and Andrzej Waleszczyński. 2019. "The Significance of the Relationship between Main Effects and Side Effects for Understanding the Knobe Effect: Database.": <https://osf.io/ky3re>
- Obidziński, Michał, and Andrzej Waleszczyński. 2021. "The Problem of Intention and the Evaluative Properties of Effects in the Knobe Effect: Database.": <http://osf.io/8c5qk>
- Paprzycka, Katarzyna 2012. "Analityczna filozofia działania: problemy i stanowiska." In *Przewodnik po filozofii umysłu*, edited by Marcin Miłkowski and Robert Poczobut, 465–94. Krakow: Wydawnictwo WAM.
- Paprzycka, Katarzyna. 2015. "The Omissions Account of the Knobe Effect and the Asymmetry Challenge." *Mind and Language* 30: 550–71. <https://doi.org/10.1111/mila.12090>
- Paprzycka, Katarzyna. 2016. "Poznań Studies in the Philosophy of the Sciences and the Humanities." In *The Sciences*, edited by Adrian Kuźniar and Joanna Odrowąż-Sypniewska, 107: 204–33. Leiden—Boston: Brill Rodopi.
- Piekarski, Michał 2017. "Efekt Knobe'a, normatywność i racje działania." *Filozofia Nauki* 97: 109–28. <https://www.fn.uw.edu.pl/index.php/fn/article/view/846>
- Shepherd, Joshua. 2012. "Action, Attitude, and the Knobe Effect: Another Asymmetry." *Review of Philosophy and Psychology* 3: 171–85. <https://doi.org/10.1007/s13164-011-0079-7>
- Sneddon, Andrew 2006. *Action and Responsibility*. Dordrecht: Springer Netherlands. <https://doi.org/10.1007/1-4020-3982-4>

- Sripada, Chandra S. 2010. "The Deep Self Model and asymmetries in folk judgments about intentional action." *Philosophical Studies* 151: 159–76. <https://doi.org/10.1007/s11098-009-9423-5>
- Sripada, Chandra S. 2012. "Mental state attributions and the side-effect effect." *Journal of Experimental Social Psychology* 48: 232–38. <https://doi.org/10.1016/j.jesp.2011.07.008>
- Sripada, Chandra S., and Sara Konrath. 2011. "Telling More Than We Can Know About Intentional Action." *Mind & Language* 26: 353–80. <https://doi.org/10.1111/j.1468-0017.2011.01421.x>
- Uttich, Kevin, and Tania Lombrozo. 2010. "Norms Inform Mental State Ascriptions: A Rational Explanation for the Side-Effect Effect." *Cognition* 116: 87–100. <https://doi.org/10.1016/j.cognition.2010.04.003>
- Waleszczyński, Andrzej, Michał Obidziński, and Julia Rejewska. 2018. "The Knobe Effect From the Perspective of Normative Orders." *Studia Humana* 7: 9–15. <https://doi.org/10.2478/sh-2018-0019>
- Waleszczyński, Andrzej, Michał Obidziński, and Julia Rejewska. 2019. "The Significance of the Relationship between Main Effects and Side Effects for Understanding the Knobe Effect." *Organon F* 26: 228–48. <https://doi.org/10.31577/orgf.2019.26203>
- Wright, Jennifer C., and John Bengson. 2009. "Asymmetries in Judgments of Responsibility and Intentional Action." *Mind & Language* 24: 24–50. <https://doi.org/10.1111/j.1468-0017.2008.01352.x>
- Zalla, Tiziana, and Marion Leboyer. 2011. "Judgment of Intentionality and Moral Evaluation in Individuals with High Functioning Autism." *Review of Philosophy and Psychology* 2: 681–98. <https://doi.org/10.1007/s13164-011-0048-1>

The Whole-Part Dilemma: A Compositional Understanding of Plato's Theory of Forms

Seong Soo Park*

Received: 17 September 2020 / Revised: 8 January 2021 / Accepted: 23 June 2021

Abstract: In this paper, I suggest a way of resolving the whole-part dilemma suggested in the *Parmenides*. Specifically, I argue that grabbing the second horn of the dilemma does not pose a significant challenge. To argue for this, I consider two theses about Forms, namely, the oneness and indivisibility theses. More specifically, I argue that the second horn does not violate the oneness thesis if we treat composition as identity and that the indivisibility thesis ought to be reinterpreted given Plato's later dialogues. By doing so, I suggest a compositional understanding of Plato's theory of Forms, which can resolve the whole-part dilemma.


Keywords: Plato; Parmenides; theory of Forms; one and many.

1. Introduction

In the *Parmenides*, Parmenides argues against one version of theories of Forms, what young Socrates has in mind, by suggesting six different lines

* Sungkyunkwan University

 <https://orcid.org/0000-0002-7403-6343>

 Department of Philosophy, Sungkyunkwan University, Sungkyunkwan-ro 25-2, Jongro-gu, Seoul, South Korea

 seongsoo@buffalo.edu

of criticism.¹ This paper focuses on his second criticism—what I call the whole-part dilemma. The dilemma begins with the following conversation between Parmenides and Socrates:

‘But tell me this: is it your view that, as you say, there are certain Forms from which these other things, by getting a share of them, derive their names—as, for instance, they come to be like by getting a share of likeness, large by getting a share of largeness, and just and beautiful by getting a share of justice and beauty?’

‘It certainly is,’ Socrates replied.

‘So does each thing that gets a share get as its share the Form as a *whole* or a *part* of it? Or could there be some other means of getting a share apart from these two?’

‘How could there be?’ he said. (*Parmenides*, 131a, italics added)

In the conversation, Parmenides and Socrates accept two distinct types of entities, Forms and things that get a share of a Form. I will simply call the latter type of entities *sensible particulars*. As we saw, while suggesting the two options, Parmenides asks Socrates to elucidate the relation that holds between sensible particulars and Forms. This relation is often alluded to by Plato in his dialogues (*Phaedo*, 100c–7; *Parmenides*, 130a–134e; *Sophist*, 256a–b) by various terms, such as “participating in,” “sharing,” or “partaking of.” However, what this participation relation really is remains rather elusive.

According to the standard interpretation of the *Parmenides*, what Socrates and Parmenides both have in mind in reference to the participation relation is what might be called the *Pie Model*.² The Pie Model has two variations: the Whole Pie Model and the Piece-of-Pie Model. The Whole Pie Model says that participants partake in a pie if they get the *whole* pie. The Piece-of-Pie Model says that participants partake in a pie if they get a *part* of the pie.

If we construe Forms as a sort of pies, we can easily see how each variation can be applied to the participation relation.³ I will call the two resulting

¹ In this paper, all references about Plato’s dialogues come from (Cooper 1997).

² For further details, see (Rickless 2007a).

³ Although the Pie Model can be applied more easily to the participation relation if we construe Forms as pies, it does not mean that the shapes of Forms are pies. In fact, what the two variations of the Pie Model require in their applications is that

applications *WPM* and *PPM* respectively. Thus, in *WPM*, sensible particulars participate in a Form by virtue of getting the *whole* of the Form, while in *PPM*, sensible particulars participate in a Form by virtue of getting a *part* of the Form. However, Parmenides points out that neither option is desirable.

The aim of this paper is to suggest a possible way of resolving the whole-part dilemma that Plato could have adopted. More specifically, I will argue that grabbing the second horn of the dilemma—that is, adopting *PPM*, does not pose a significant challenge. To do this, I will take the following steps. First, in Section 2, I will outline the logical structure of the whole-part dilemma. Then, in Section 3, I will clarify the four implications of adopting *PPM* and suggest a compositional understanding of the participation relation between sensible particulars and Forms. After that, in Section 4, I will argue that if the relation between shares of Forms and Forms is compositional, the oneness thesis, according to which every Form is one, is not infringed by *PPM*. Lastly, in Section 5, I will argue that the indivisibility thesis, according to which Forms are indivisible, ought to be reinterpreted given the textual evidence. This will result in a compositional understanding of Plato's theory of Forms, which can resolve the whole-part dilemma.

2. The whole-part dilemma

In this section, I will present the target of this paper, namely, the whole-part dilemma. I begin by reconstructing the dilemma with the Pie Model as follows:

- P1. Nothing except the Pie Model explains the participation relation.
- P2. The Pie Model has two variations: *WPM* and *PPM*.
- P3. Neither *WPM* nor *PPM* is convincing.
- P4. If (P1) & (P2) & (P3), there is no way to understand the participation relation.
- C. Therefore, there is no way to understand the participation relation.

(for *WPM*) every Form is one, or that (for *PPM*) every Form has the parts. To be clear, I am not arguing that the shapes of Forms are pies. As will be argued later, shapes are not essential features of Forms.

The first and second premises set two horns of the dilemma. According to the conversation between Parmenides and Socrates, the participation relation should be one of the following two cases: either particulars get a *part* of a Form and thereby participate in the Form, or they get the *whole* of a Form and thereby participate in the Form.

From my perspective, the least controversial premise is the fourth one. If we have only two options of understanding the participation relation, and neither is desirable, then it conceptually follows that there is no way to understand the participation relation. On the contrary, the most controversial premise is the third one, given that it is the core premise that constitutes the dilemma. To support this premise, Parmenides attempts to show that both horns (i.e., WPM and PPM) generate an undesirable consequence. It can be explained as follows.

First, let us assume that WPM is true. Then, it is possible that different particulars get one and the same Form. For example, in this model, some objects are beautiful by virtue of getting the Form of Beauty. The issue with accepting such a case is that it demands that one and the same thing be in separate places simultaneously.

Second, let us assume that PPM is true. Then, each particular gets a different part of a Form. That is, in this model, some objects, let us say, are beautiful in virtue of getting a different part of the Form of Beauty. So, PPM does not demand that one and the same thing be in separate places simultaneously. However, Parmenides seems to believe that this option leads to a violation of the oneness and indivisibility theses; every Form is (1) one and (2) incomposite. He says, “Then are you willing to say, Socrates, that our one Form is really divided? [If so,] will it still be one?” (Plato 1997, 131c) Socrates and Parmenides agree that it won’t be so, and conclude that PPM is problematic as well.

The conclusion is so detrimental that one cannot simply bite the bullet. The participation relation is indispensable in Plato’s theory of Forms. So, if one wants to endorse Plato’s theory of Forms, one should deny one of the premises in the dilemma. Indeed, there have been some debates around the first horn.⁴ I will not deal with them here. Instead, I will focus on suggesting

⁴ Cherniss, Peck, and Sayre deny the third premise by grabbing the first horn, and several critics like Rickless and Panagiotou argue against them. In this paper, I will

a possible way to resolve the whole-part dilemma on the basis of PPM, the second horn.

3. A compositional account of the participation relation

One way to argue against the dilemma's second horn is to simply deny the following two theses of Forms:

Oneness: Every Form is one.

Indivisibility: Forms are indivisible.

According to this strategy, Forms do not need to be one and can be divided. So, in accordance with this strategy, one might endorse PPM to give an account of the participation relation between sensible particulars and Forms. However, this suggestion does not seem attractive. First, given Plato's theory of Forms, the two theses cannot be discarded for no reason. Second, and more importantly, even if there is a positive reason to surrender them, it seems that three unanswered questions remain: (I) What is the nature of the parts of Forms? (II) What is the role played by them? (III) What is their relation to Forms? Thus, in this section, let me first show that PPM in fact gives us direct answers to these questions. In subsequent sections of the paper, I will then turn to the oneness and indivisibility theses, arguing that in certain interpretation of these theses PPM is compatible with them.

I begin by focusing on what PPM really implies. To be specific, accepting PPM is tantamount to taking the following claims to be true.

- (A) Forms have shares as their parts.
- (B) The relation that holds between the shares and the Forms is a part-whole relation.
- (C) The parts (i.e., shares) are (individually, not collectively) distinct from the Form that they belong to.
- (D) The parts of a Form are property instances.

not judge whether their arguments are persuasive. For more details, see (Cherniss 1932), (Peck 1953), (Panagiotou 1987), (Sayre 1996), and (Rickless 2007a).

First, (A) says that Forms have parts and that these parts are called ‘shares.’ The former directly follows from what PPM says. According to PPM, sensible particulars participate in a Form in virtue of getting a part of the Form. That is, PPM implies that Forms have parts. The latter can be seen in the conversation between Parmenides and Socrates quoted in section 1. In the conversation, Parmenides asks Socrates a question about the two variations of the Pie Model by saying, “Does each thing that gets a *share* get as its share the Form as a whole or a part of it?” (*Parmenides* 131a5–6) This shows that Parmenides considers the possibility that a Form has shares (or more strictly, the entities which Parmenides and Socrates call shares) as its parts. And this possibility is the core assumption that constitutes PPM. Thus, if we endorse PPM, then we must construe a share as a part of a Form.

Second, (B) naturally follows from (A). If a Form has some shares as parts, then the relation between the shares and the Form is a part-whole relation. In other words, the shares, in some sense, collectively compose the Form.⁵ Third, according to PPM, each share is (individually, not collectively) distinct from the Form it belongs to; otherwise the same difficulty WPM faced—one and the same thing should be in separate places simultaneously—will arise again. Lastly, PPM assumes (D) as well, since according to PPM, sensible particulars have their properties in virtue of getting shares. In this sense, shares can be construed as playing the same role as property instances which are sometimes also called ‘tropes’ in the terminology of contemporary metaphysics.⁶

In fact, it is frequently pointed out by those who might be called *Platonic trope theorists* that there is a distinct type of entities in Plato’s metaphysical view, what Socrates calls shares, and these shares play the role that property instances typically do. For example, McPherran (1988), Mertz (1996), and Buckels (2018) argue that Plato’s middle dialogues (e.g., the *Republic*, the *Phaedo*, and the *Parmenides*) as well as later ones (e.g., the *Theaetetus* and the *Timaeus*) show that the role of shares of

⁵ This will be explained further in the next section.

⁶ See (Maurin 2018).

Forms is to ascribe non-repeatable properties to sensible particulars.⁷ Mertz writes:

Summarizing the textual evidence, in the *Republic* (510d), Plato refers to a class of “visible forms,” and in the *Parmenides* (130b), gives examples of the likenesses that we each possess, in contrast to LIKENESS itself...Similarly, in the *Phaedo* (102d–3b), OPPOSITENESS, LARGENESS, and SMALLNESS are distinguished from cases of oppositeness, largeness, and smallness that are “in us” ...In the *Theaetetus* (209a–d), it is argued that unit properties are needed to individuate what would otherwise be just bundles of universals. (Mertz 1996, 83–84)

However, while Platonic trope theorists explicitly mention that Plato admits the existence of shares which are very similar to property instances, they do not mention the relation between shares and Forms much. McPherran writes, “Immanent characters [shares] are likenesses of Forms and so act as properties.”⁸ Although I agree that shares are likenesses of Forms (or images of Forms), I believe that more can be said about this. In fact, one advantage of adopting PPM is that it gives an additional account of the relation between shares and Forms. In PPM, the relation is based on a part whole relation. That is, the shares collectively compose a Form. Thus, according to this model, the claim that shares bear a resemblance to Forms can be explained by the fact that shares compose Forms.

By making the implications of PPM explicit, we are now in a position to be able to give the following answers to the above three questions: (I) the parts of Forms are shares; (II) shares play the role of ascribing a non-repeatable properties to an object; (III) their relation to Forms is compositional.

Please note that this line of thought is not arbitrary. In the *Parmenides*, Parmenides and Socrates both agree that PPM is one of the two genuine options that must be considered in explaining the participation relation.

⁷ It is rather controversial how to arrange Plato’s dialogues. Different scholars may order the dialogues differently.

⁸ McPherran calls shares of Forms immanent characters. See (McPherran 1988, 534).

And taking the relation between shares and Forms to be compositional is the most natural way to endorse PPM. Thus, as a working hypothesis, I suggest that the relation between shares and Forms is a part-whole relation. To emphasize this, I will call shares of a Form *Form parts*.

Now I will give a final account of the participation relation on the basis of PPM. As mentioned earlier, the participation relation is the relation that holds between sensible particulars and Forms. Here sensible particulars are complex entities. To explain in what sense they are complex, I draw on another kind of entity that is mentioned in the *Timaeus*, namely, the receptacle. According to the *Timaeus* (50e5–8), the receptacle is a sort of base in which properties are able to inhabit. The most crucial feature of the receptacle is that it lacks any qualitative characteristics in its own right (except that it is characterless).

By accepting this entity, PPM can give a compositional account of the participation relation between sensible particulars and Forms. Sensible particulars are complex entities whose constituents are the receptacle and Form parts. The receptacle is the base in which Form parts are able to inhabit. Form parts are property instances, and they enable sensible particulars to maintain some properties. The relation of Form parts to Forms is compositional. As a result, the compositional account of the participation relation can be articulated as follows:

The receptacle possesses a Form part that is a constituent of a Form, and thereby, the sensible particular resulting from the combination of the Form part and the receptacle participates in that Form.

The plausibility of this articulation depends on two pending issues; the oneness and indivisibility theses. I will turn to them in the subsequent sections.

4. The problem of oneness and composition as identity

In this section, I argue that PPM is compatible with the oneness thesis if the relation between Form parts and Forms is understood to be compositional. To begin, let me consider again Parmenides's words, "Then are you willing to say, Socrates, that our one form is really divided? Will it still be one?" (*Parmenides*, 131c8–9)

Parmenides's first question involves the indivisibility thesis. So, this section focuses on his second question. As we see, it seems that Parmenides implicitly assumes that if Forms were divided into parts, they could not be one. However, does the antecedent necessarily imply the consequent? I do not think so if the relation between Form parts and Forms is compositional.

It is worth noting that there are various understandings of composition. Among them, I endorse a specific view of the composition, one that treats the composition as an identity relation. I will call this specific view of composition *CAI* for short. One key claim of *CAI* is that the composition is ontologically impotent, that is, "when parts compose a whole, the composition does not create a new entity for our list of beings" (Brown 2004). This is because *CAI* treats the composition as an identity relation. Thus, according to *CAI*, being divided into many parts does not entail the nonexistence of the whole, since the parts are in themselves identical to the whole. Thus, if the relation holding between Forms and Form parts is compositional, then a Form *is nothing over and above* the Form parts constituting it. Thus, *CAI* preserves the oneness thesis.

CAI is controversial. The debate on whether it is a tenable view is ongoing.⁹ However, it seems less controversial that Plato accepts this view. Indeed, a number of passages in different dialogues confirm it. For example, the idea that the whole is just the same as the parts is first mentioned in the *Parmenides*. In this dialogue (129c1–d2), Socrates admits that "the entire parts of his body and he himself are the same." Additionally, the first deduction in the *Parmenides* (137c4–142d) assumes that the one is not many, and shows that this assumption leads to undesirable consequences. The first deduction thus supports the claim that Plato accepts *CAI*. Furthermore, a similar idea also appears in the *Theaetetus* (204a7; 205a9–10). Here, Socrates claims that "when a thing has parts, the whole *is* necessarily

⁹ Although *CAI* is controversial, its restricted version, which states that there is more than one composition relation and only some kind of particular enjoys the composition relation as identity, is less controversial. McDaniel (2004), a main critic of *CAI* also admits that some versions of *CAI* can be compatible with some versions of compositional pluralism. The point is that it is the restricted version of *CAI* that I will endorse as below. For more details, see (Baxter 1988), (McDaniel 2004), and (Wallace 2011).

all the parts” and also that “in the case of a thing that has parts, both the whole and the sum will be the parts.”¹⁰ Therefore, I conclude that Plato endorses CAI.

One might wonder how one is identical to many in the framework of Plato’s theory of Forms by pointing out that it seems to violate the thesis of radical purity (or RP for short), according to which, Forms do not have contradictory properties in the same respect. However, there are at least two ways of dealing with this issue. First, we can adopt the developmentalists’ view and deny RP, since RP is just mentioned once in the *Republic* (436b). Indeed, Priest (2013) and Rickless (2007b) adopt this strategy. Second, we might argue that CAI does not violate RP. Specifically, Form parts can be regarded as one only if they are under the concept of Forms, while Form parts can be regarded as many only if they are under the concept of Form parts. To put it another way, the question of how many things are there is an ill-formed question since counting is necessarily tied to our concepts. So, we should ask “How many Forms are there?,” or “How many Form parts are there?” Then, it will turn out that the Form parts and the Form do not have contradictory properties *in the same respect*.¹¹

Before we proceed further, it is worth mentioning that there is a competing interpretation of Plato’s view of composition suggested by Harte. Harte (2002) argues that although Plato seems to endorse CAI in the *Theaetetus* and the *Parmenides*, he denies it and endorses the so-called *structural view* in later dialogues (e.g., the *Sophist*, the *Philebus*, and the *Timaeus*).

According to Harte (2002), the structural view says that there is an additional element aside from parts required by composition, namely, structure. By extracting the notion of structure from the later dialogues, Harte argues that later Plato’s view of composition suggests that certain parts compose a whole only when they are arranged in a *proper* way. This claim, if true, can significantly challenge my work, since the current arrangement of Form parts in the receptacle may not be sufficient to compose a Form.

¹⁰ Italics added.

¹¹ For more details, see (Wallace 2011).

However, even if Harte's extraction of the notion of structure is appropriate,¹² it is still questionable whether parts without a presumed order actually entail the non-existence of a whole. This is because it is one thing to say that structure affects the normative status of composite objects, such as labeling them good or bad, but another thing to insist that the parts are unable to compose the whole without possessing a proper order. Consider the case of the weather Harte (2002) mentions in her book. Even if elements of weather create good weather only if they are arranged in a specific way, that is not to say that there would be no weather if the elements are arranged differently.¹³

Moreover, for the sake of argument, even if Harte's interpretation of Plato's view of composition is right, it should be emphasized that my argument can still stand. This is because her interpretation is not committed to the claim that Plato could not have been a *compositional pluralist*. According to compositional pluralists (e.g., Fine (2010), Baxter and Cotnoir (2014)), there is more than one basic parthood relation. That is, they claim that different kinds of objects may possess different composition relations that can be defined in terms of different basic parthood relations. For example, as Fine (2010) points out, the way in which the letter 'n' is a part of the expression 'no' is different from the way in which it is a part of the set of letters {'n,' 'o'}. The difference lies in our *concepts* of them associated with their criteria of identities. Unlike the case of the word 'no,' the identity of sets is solely determined by its members. To put it another way, sets do not conceptually require a structure to establish their identity: Any sets are the same just in case they have the same members.

To clarify the point that my view is consistent with Harte's interpretation of Plato, it is worth noting that the objects Harte deals with in her

¹² Plato never explicitly mentions the notion of structure.

¹³ One might wonder whether some particular kinds other than weather could have a different composition relation. I would assume that they do. However, as we will see, I will endorse the view that there is also another composition relation. In addition, since the objects that Harte deals with in her book are limited to several kinds, I do not have sufficient resources for handling this issue. So, I will focus on claiming that Forms and Form parts may have a different composition relation from one that Harte suggests.

book are limited to the combinations of sensible particulars such as weather, notes, and letters.¹⁴ Thus, even if we accept that the composition relation holding between sensible particulars demands that there be a structure in composing a complex entity, it does not necessarily follow from this that the relation requiring a structure has to be applied to all kinds of composite objects. Then, it can be argued that the criterion of the identity for Forms does not involve any structure. That is, any Forms are the same just in case they have the same Form parts.

To be sure, the shape of a Form will differ according to how the Form parts are laid in the receptacle. However, this does not jeopardize the above criterion of identity for Forms. This is because according to Plato, the property of having a certain shape is not an essential property of Forms that is closely tied with the identity condition for Forms. For example, a round Form cannot have been round *from the beginning*. The only way for a Form to be round, according to Plato, is to participate in another kind of Form like the Form of Roundness or the Form of Change.¹⁵ Consequently, there is no good reason to say that the identity of a Form is determined by the structure of Form parts. This completes my argument for the claim that understanding the composition relation between Form and Form parts in a way that CAI theorists suggest does not conflict with Harte's interpretation of Plato's view of composition.

¹⁴ I do not doubt that the composite objects Harte deals with in her book are limited to sensible particulars. Harte argues that the composite objects she introduces are indeed *scientific objects*. Thus, from my perspective, there is no reason to apply the structural view to all kind of objects. For more details, see (Harte 2002, 268).

¹⁵ What is the relation between Forms? How can we explain the way that a Form participates in another Form? I have no definite answer to these questions. However, for current purpose, it would be sufficient to say that the relation between Forms is not involved with a criterion of identity for Forms, given Plato's theory of Forms. In addition, as a very rough sketch, it could be suggested that the way that a Form participates in another Form is related to how their Form parts are arranged in the receptacle. The point is that on PPM, the relation between Form parts and Forms is different from the one between Forms. The former is compositional, the latter is not.

In sum, I argued that if the relation between Form parts and Forms is understood to be compositional, the oneness of Forms could be preserved. This is because Plato treats composition as identity. Some developmentalists like Harte may understand Plato's view of composition differently. However, even if we accept their view, it does not exclude the possibility that there is more than one composition relation. Therefore, if we treat composition as identity with a certain limited applicability, then PPM is compatible with the oneness thesis.

5. Negotiability of the indivisibility thesis

In this section, I argue that the indivisibility thesis is negotiable. What I mean by "negotiable" is that we can decide whether to preserve this thesis in the way it is traditionally understood on the basis of potential theoretical benefits. Put simply, I argue that we can obtain some theoretical benefits at the cost of the indivisibility thesis. This results in sacrificing some orthodox readings of Plato. Nevertheless, I believe that the trade-off is worth considering since the cost is lower than expected.

So, my strategy in this section is not to argue that the indivisibility thesis should be discarded. Rather, I will merely focus on showing that there are substantial theoretical rewards to be gained if we replace the traditional thesis with a less stringent one, what might be called the *likely indivisibility thesis* (or LID for short): Forms are most likely indivisible.

Again, the cost is not too high in that this thesis replacing traditional indivisibility thesis can play much of the same role as the original one. Furthermore, from my perspective, the original thesis is controversial enough so as to warrant the consideration of an alternative one.

To argue for this line of thought, I will evaluate the costs and benefits of surrendering the original indivisibility thesis by focusing on the textual evidence commonly said to uphold it. More specifically, I will consider two passages mentioned by Rickless (2007b) that have been regarded as supporting materials for the indivisibility thesis, one in the *Phaedo* and the other in the *Timaeus*, arguing that neither passage is decisive when it comes to upholding the original version of the indivisibility thesis and that given

the theoretical benefits we should adopt an alternative, less stringent version of the thesis.

I will start with the passage from the *Phaedo*:

Are not the things that always remain the same and in the same state *most likely* (*malista eikos* [μάλιστα εἰκός]) not to be composite, whereas those that vary from one time to another and are never the same are composite? (*Phaedo*, 78c6–8)

Here, Socrates's point is that it is not extremely *probable* that what is always constant and invariable is divisible. Does this passage really bolster the indivisibility thesis? It does not appear to do so. Rickless admits this as well. He says, "Here, Socrates does not commit himself to the strong claim that Forms are incomposite" (Rickless 2007b, 43). Thus, strictly speaking, this passage is inconsistent with the indivisibility thesis. What the passage literally means is that even the things that always remain the same and in the same state are, *in principle*, divisible. Therefore, I conclude that far from supporting the indivisibility thesis, this textual evidence states the truth of a less stringent version of the indivisibility thesis (LID) that Forms are most likely indivisible.

One might wonder whether we should interpret the passage in light of the tendency of Forms to remain the same. For instance, one may argue that we can interpret the term 'most likely' as 'extremely plausible.' Then it may be that the passage supports the indivisibility thesis. However, this possible objection requires us to accept a wide scope view for the term 'most likely.' To be more specific, since it requires us to interpret the term 'most likely' as 'extremely plausible,' it demands that the term 'most likely' be placed *outside* of the that-clause. For example, the result would be the following:

- (α) It is most likely (it is extremely plausible) that the things that always remain the same and in the same state are not composite.

However, the cited passage is not like (α), but rather like the following:

- (β) The things that always remain the same and in the same state are most likely not composite.

This is in accordance with the Greek text: the passage comprises an ACI-construction dependent on Cebes' previous "It seems to me ... to be this way (δοκεῖ μοι ... οὕτως ἔχειν *dokei moi ... houtōs echein*)."

Accordingly, practically all translators (e.g., Grube (1997), Gallop (1975), and Jowett (1892)) adopted the narrow scope view that demands the term 'most likely' be placed *inside* of the that-clause when they translated the *Phaedo*. Consequently, the objection is not in accordance with the standard construal of the grammatical structure of *Phaedo* 78c6–8.

In what follows, let us consider the second passage that is commonly taken to support the indivisibility thesis, the passage from the *Timaeus*:

The component from which he [the father] made the soul and the way in which he made it were as follows: In between [(a)] the Being (*ousia*) that is indivisible and always changeless, and [(b)] the one that is divisible and comes to be in the corporeal realm, he mixed [(c)] a third, intermediate form of being, derived from the other two (*Timaeus*, 35a2–5).

According to the standard reading, (a), (b), and (c) refer to the Form of Being, the sensible particular, and the soul, respectively. Based on this reading, Rickless (2007b) regards the passage as the strongest evidence upholding the indivisibility thesis. I will call this passage *35a2*. In his view, the indivisibility thesis cannot be discarded given the traditional reading of *35a2*. In other words, he claims that we cannot put a price on the value of the indivisibility thesis, given *35a2*.

If the thesis has a high value, then nobody would be willing to trade it. However, I believe that the value of the thesis is set too high by *35a2*. So, I will attempt to lower the value *to the point* where it can be exchanged for some theoretical benefits.

To begin with, it could be pointed out that *35a2* is inconsistent with what Socrates says in the *Phaedo* if we interpret it as I did above. However, this inconsistency *per se* is not decisive for two reasons.¹⁶ First, it might seem rather unfair if I reject that *35a2* supports the indivisibility thesis

¹⁶ Thus, it would be worth noting that the passage in the *Phaedo* shouldn't be regarded as my main reason to deny the claim that *35a2* supports the indivisibility thesis.

solely based on my preferred reading of 78c6–8 in the *Phaedo*. Second, and more importantly, it is easy for developmentalists to assume that Plato decided to make his claim about the indivisibility of Forms much stronger in his later dialogues.

Luckily, there are three other passages in the Platonic corpus (in the *Sophist*, *Theaetetus*, and even the *Timaeus*) that show that the *Timaeus* passage just quoted presents an anomaly and give us good reason to adopt LID. Let me start with the *Sophist*, where we can easily find the idea of blending of Forms. In reply to the question about a good man, Plato says that man is one Form and good is another. The idea is simple and straightforward. Some Forms partake of other Forms.

This idea is mentioned explicitly in the Stranger's conversation about five kinds of Forms, Change, Rest, Being, Sameness, and Difference. It is worth noting that not only Change partakes of Being, but also Being partakes of Change. In the *Sophist*, Plato says that change is necessary for intelligence. Since Forms are intelligence-bearers (or truth-bearers), it follows that certain Forms like Being should partake in the Form of Change. Thus, if what Plato says in the *Sophist* is true, it is highly doubtful that Forms are changeless. The problem is that 35a2 states that what is assumed to be the Form (i.e. the Form of Being) is changeless. Therefore, the *Sophist* casts doubt on whether we should accept 35a2 in a literal sense.

Second, the *Theaetetus* explicitly indicates the possibility of Forms being divided. In the *Theaetetus* (204a1), Socrates says, "Let the complex be a single form resulting from the combination of the several elements when they fit together." This clearly indicates composite Forms. Here, it is worth paying attention to Owen's (1953) claim that the *Timaeus* ought to be dated before the *Theaetetus* on various grounds. If Owen is right about this, then it could be argued that Plato was reluctant to discard the indivisibility thesis until the *Timaeus*. I admit that it is suspicious that Owen's thesis is indeed right. However, it is apparently sufficient to show that his claim can play the desired role—that is, the role of moderating the value of the indivisibility thesis.

Third, one could still be unsatisfied at this point and may argue that we should determine whether to accept the indivisibility thesis on the basis of the *Timaeus* alone. I do not see any reason why. Plato would likely not

want this either because it is nearly impossible to construct the theory of Forms with just the *Timaeus*.

However, let us grant for the sake of argument that we should ascribe to Plato the indivisibility thesis on the basis of *Timaeus* alone and ignore all other texts in the Platonic corpus that would favor adopting LID instead. Even in this case, there is evidence within the *Timaeus* that challenges the view that we ought to adopt the indivisibility thesis in its traditional form. In fact, the paragraph immediately following 35a2 reads as follows:

Similarly, he [the Father] made a mixture of the Same, and then one of the Different, in between their indivisible and their corporeal, divisible counterparts. And he took the three mixtures and mixed them together to make a uniform mixture, forcing the Different, which was hard to mix, into conformity with the Same. Now when he had mixed these two together with Being, and from the three had made a single mixture, he *redivided* the whole mixture into as many parts as his task required (*Timaeus*, 35a5–11).¹⁷

In the above passage, the term ‘redivided’ is worth noting. My question is this: If (a) in 35a2 is indeed indivisible, how could the Father (namely, God) *re*-divide the whole mixture, which includes (a), into many parts? The way I see it, this requires that (a) be the sort of thing that can be divided in principle at least. If not, then the term ‘redivided’ is unsuitable in this context.¹⁸ This strongly suggests that the indivisibility thesis in its traditional form is too strong and the less stringent version (LID) seems more plausible, even in the context of the *Timaeus* alone. Thus, I suggest that we should not take 35a2 literally. Specifically, I suggest that there is

¹⁷ Italics added. Here’s the Greek for reference: μεγνός δὲ μετὰ τῆς οὐσίας καὶ ἐκ τριῶν ποιησάμενος ἕν, πάλιν ὅλον τοῦτο μοίρας ὅσας προσήκεν διένειμεν, ἐκάστην δὲ ἕκ τε ταύτου καὶ θατέρου καὶ τῆς οὐσίας μεμειγμένην.

¹⁸ It might be argued that Forms are indivisible, but the mixture of Forms and other elements can be divided. However, I do not see how this should work. Socrates says that the Father redivided the mixture into *many* parts. Given this, the divided parts should contain some part of Forms. As a result, we should accept an uncanny view about mereology to endorse this move.

a hidden phrase like ‘most likely’ in 35a2. Accordingly, the result will be as follows: “the Being that is most likely indivisible and changeless...”

At this point, let me summarize my argument again. I have attempted to moderate the value of the indivisibility thesis by casting doubt on the passages in the *Phaedo* and *Timaeus* that are said to uphold the indivisibility thesis. Specifically, in regard to the passage in the *Phaedo*, I argued that the passage does not actually support the thesis. In regard to the passage in the *Timaeus*, I argued that the passage does not accommodate some passages in the *Sophist*, *Theaetetus*, and *Timaeus* as well as the *Phaedo*. Therefore, I conclude that the indivisibility thesis is not the sort of thing that can never be sacrificed fully.

Before proceeding further, it is worth noting that LID also has some theoretical foundations. First, we may see that the passage in the *Phaedo* literally supports LID. Moreover, other, additional passages in the *Timaeus* may be used as well to support it. The passages involve the conversation between Timaeus and Socrates about giving an account. In the *Timaeus* (29b3–29c7), Socrates agrees with Timaeus’s point that we can only give a sort of *likely account* (*eikos logos*) of a certain subject. This is because we are no more than human in nature (*Timaeus* 29c–d). If so, Timaeus’s point can provide a theoretical basis for LID in that LID has a good fit with the notion of a likely account; LID not only allows for cases in which a Form is not yet divided into parts, but also for cases in which a Form has been divided.¹⁹

Now, I shall obtain several theoretical advantages at the cost of the indivisibility thesis. While it is true that the indivisibility thesis has been supported by the orthodox reading of 35a2, it may also be that 35a2 is controversial enough to contemplate the adoption of a weaker form of the thesis, namely, LID. If this is the case, then several potential benefits will motivate us to adopt LID. In other words, I believe that the theoretical

¹⁹ Please note that I am not arguing that the notion of a likely account contradicts the indivisibility thesis. Rather I am merely attempting to provide some theoretical foundation for LID. The reason to cast doubt on the indivisibility thesis is based on my discussion of the two passages that are commonly said to support the thesis, not the notion of a likely account.

benefits of LID will act as a tiebreaker in the decision of whether or not to weaken the indivisibility thesis.

First, since LID is a less stringent thesis than the original one, it can improve the coherence of Plato's dialogues and provide developmentalists with an adequate explanation. Regarding the debate on indivisibility of Forms, the aforementioned inconsistencies between dialogues—especially between the *Timaeus* and later dialogues—can be accounted for if we replace the indivisibility thesis with LID. Second, Since PPM does not violate LID, LID can play a key role in giving a compositional account of the participation relation between sensible particulars and Forms by providing a theoretical base for PPM. Again, according to the compositional account, the receptacle possesses a Form part that is a constituent of a Form, and thereby, the sensible particular resulting from the combination of the Form part and the receptacle participates in that Form. Third, this line of reasoning may be the best suited for Platonic trope theorists. Trope theorists typically owe their explanatory power to the notion of exact resemblance. And the best-well known strategy for dealing with this notion is to treat it as primitive. Many trope theorists maintain that there is no further explanation for the notion of exact resemblance because it is the notion that constitutes our conception of tropes. They just take it for granted that there are some property instances that are qualitatively the same but numerically different. However, even if Platonic trope theorists can follow this strategy, they have one more task than non-Platonic trope theorists. The task is that they need to elucidate how the notion of exact resemblance can be related to Forms. The final reward of LID is the simple answer it provides through PPM; the fact that some tropes exactly resemble each other can be explained by the fact that they compose a Form. That is, the reason why, let say, some red tropes resemble each other perfectly is that they compose the Form of Redness.²⁰ Thus, the exact resemblance is not a primitive concept in this model. Rather, the notion of Form is a primitive one.

Lastly, I shall conclude this paper by evaluating the costs. A big expenditure is that endorsing LID takes us away from the traditional reading

²⁰ Plato denies the existence of a Form of a color. This is just mentioned as an example.

of Plato. The indivisibility thesis will no longer be preserved in its traditional form. In addition, it makes it harder to explain the perfection of Forms. According to LID, Forms can be divided in principle. If so, how can we explain the perfection thesis, which states that Forms are perfect? This could be a burden for someone opting for the trade-off and will require further work on another occasion.

Another cost is that surrendering the indivisibility thesis and endorsing PPM forces us to accept the claim that there are tropes (or property instances) in Plato's metaphysical view. However, certainly some scholars might not want to be a trope theorist even in a loose sense. Thus, if one disagrees with the key idea of tropes and wants to remain an orthodox Platonist, one is better off not giving up the indivisibility thesis. On the contrary, if one has some of the intuitions that trope theorists have, I would strongly recommend to reap several theoretical benefits at the cost of the indivisibility thesis. Adopting LID and taking the relation between shares of Forms and Forms to be compositional would be one of the most attractive options for Platonic trope theorists.

6. Conclusion

The whole-part dilemma begins with Parmenides's question of what the participation relation is. And this question led us to an investigation of the relation between Forms and shares of Forms. From my perspective, the dilemma is the device that is designed to initiate the thought that the relation in question *might be* compositional.

In this paper, I pushed the mentioned thought to the greatest degree. By doing so, I suggested a compositional understanding of Plato's theory of Forms, and argued that the whole-part dilemma can be resolved by this understanding. Again, the compositional approach is not arbitrary. Taking the relation between shares and Forms to be compositional is the most natural way to grab the second horn of the whole-part dilemma.

While I am not too concerned about the oneness of Forms, I think the plausibility of this paper depends on how convincing my argument regarding the indivisibility thesis was. I hope one finds the argument persuasive and contemplates the option of adopting PPM. At the cost of the indivisibility

thesis, we can not only clarify the core notion of Plato's theory of Forms, but also eliminate inconsistency between dialogues. Finally, my attempt will also help Platonic trope theorists carry their own burden by providing a simple account of the notion of exact resemblance without invoking an additional primitive concept.

Acknowledgments

I would like to thank an anonymous referee, Song Ee Baek, and Yong Sung Kim for helpful comments and suggestions. I am especially grateful to Jan Maximilian Robitzsch whose detailed comments and suggestions resulted in many significant improvements in this paper.

References

- Baxter, Donald L. M. 1988. "Many-One Identity." *Philosophical Papers* 17 (3): 193–216. <https://doi.org/10.1080/05568648809506300>
- Baxter, Donald L. M., and Aron J. Cotnoir. 2014. *Composition as Identity*. Oxford University Press.
- Brown, Eric. 2004. "On Harte on Plato on Parts and Wholes." presented in Eastern APA.
- Buckels, Christopher. 2018. "Triangles, Tropes, and Τὰ Τοῦαυτᾶ: A Platonic Trope Theory." *Plato Journal: The Journal of the International Plato Society* 18: 9–24. https://doi.org/10.14195/2183-4105_18_1
- Cherniss, Harold F. 1932. "Parmenides and the Parmenides of Plato." *American Journal of Philology* 53 (2): 122–138. <https://doi.org/10.2307/289804>
- Cooper, John M (ed.). 1997. *Plato: Complete Works*. Hackett.
- Fine, Kit. 2010. "Towards a Theory of Part," *Journal of Philosophy* 107 (11): 559–589. <https://doi.org/10.5840/jphil20101071139>
- Gallop, David (ed.). 1975. *Plato's Phaedo*. Clarendon Press.
- Grube, George M. A (ed.). 1997. *Plato's Phaedo*. Hackett.
- Harte, Verity. 2002. *Plato on Parts and Wholes: The Metaphysics of Structure*. Oxford University Press.
- Jowett, Benjamin (ed.). 1892. *Plato's Phaedo*. eBooks@Adelaide.
- Maurin, Anna-Sofia. 2018. "Tropes." *Stanford encyclopedia philosophy*. <https://plato.stanford.edu/entries/tropes>
- Mertz, Donald W. 1996. *Moderate Realism and Its Logic*. Yale University Press.
- McDaniel, Kris. 2004. "Modal Realism with Overlap." *Australasian Journal of Philosophy* 82 (1): 137–52. <https://doi.org/10.1080/713659792>

- McPherran, Marc L. 1988. "Plato's Particulars." *The Southern Journal of Philosophy* 26 (4): 527–553. <https://doi.org/10.1111/j.2041-6962.1988.tb02163.x>
- Owen, Gwilym E. L. 1953. "The Place of the Timaeus in Plato's Dialogues." *Classical Quarterly* 3 (1–2): 79–90. <https://doi.org/10.1017/s0009838800002652>
- Panagiotou, Stavros. 1987. "The Day and Sail Analogies in Plato's Parmenides." *Phoenix* 41 (1): 10–24. <https://doi.org/10.2307/1088599>
- Peck, Arthur L. 1953. "Plato's Parmenides: Some Suggestions for its Interpretation." *Classical Quarterly* 3 (3–4): 126–150. <https://doi.org/10.1017/S0009838800003074>
- Priest, Graham. 2013. "The Parmenides: a dialethic interpretation." *Plato Journal* 12: 1–63. https://doi.org/10.14195/2183-4105_12_3
- Rickless, Samuel. 2007a. "Plato's Parmenides." *Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/entries/plato-parmenides>
- Rickless, Samuel. 2007b. *Plato's Forms in Transition: A Reading of the Parmenides*. Cambridge University Press.
- Sayre, Kenneth M. 1996. *Parmenides' Lesson: Translation and Explication of Plato's Parmenides*. University of Notre Dame Press.
- Wallace, Meg. 2011. "Composition as Identity: Part 2." *Philosophy Compass* 6 (11): 817–827. <https://doi.org/10.1111/j.1747-9991.2011.00430.x>

Conditional Uniqueness

Erhan Demircioğlu*

Received: 21 January 2021 / Accepted: 15 February 2022

Abstract: In this paper, I aim to do three things. First, I introduce the distinction between the Uniqueness Thesis (U) and what I call the Conditional Uniqueness Thesis (U*). Second, I argue that despite their official advertisements, some prominent uniqueness theorists effectively defend U* rather than U. Third, some influential considerations that have been raised by the opponents of U misfire if they are interpreted as against U*. The moral is that an appreciation of the distinction between U and U* helps to clarify the contours of the uniqueness debate and to avoid a good deal of talking past each other.

Keywords: Rationality; rational belief; evidence; uniqueness; permissivism.

Is there any slack between the evidence and what is rational to believe given the evidence? According to the Uniqueness Thesis (U), the answer is no:

U: Necessarily, there is at most one rational doxastic attitude one can take towards a proposition P, given a particular body of evidence E.

* Koç University

 <https://orcid.org/0000-0002-1579-7505>

 Department of Philosophy, Koç University, Rumelifeneri Yolu, 34450, Sarıyer – Istanbul, Turkey

 erdemircioglu@ku.edu.tr

© The Author. Journal compilation © The Editorial Board, *Organon F*.



This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International Public License (CC BY-NC 4.0).

U has been defended notably by Feldman (2006), White (2005), Christensen (2007) and Matheson (2011). Permissivism (P) is the denial of U. According to P, there are some possible cases in which there is more than one rational doxastic attitude that one can take towards a proposition, given the same body of evidence. P has been defended notably by Douven (2009), Titelbaum (2010), Kelly (2013), and Schoenfield (2014).

An influential charge against U voiced by permissivists is that U can only be true if a simplistic picture of how rationality is determined by evidence is taken for granted. According to what I call *the simplicity objection*, rationality is determined only *partially* by evidence, while U can only be true if it is determined *fully* by evidence.¹ The main tenet of this objection is that “the evidence all by itself leaves underdetermined whether it is rational for you to believe one or the other (proposition)” (Douven 2009, 352). The simplicity objection rests on the idea that *if* there is a factor other than evidence that is a determinant of rationality, *then* there might be more than one rational doxastic attitude one might take towards a proposition given the same evidence, and a defender of the simplicity objection holds that the antecedent of this conditional is true.

What might those extra factors be in addition to evidence that determine what is rational to believe for a subject? There are various alternatives here, one of which derives from the subjective Bayesian conception of prior probability distribution. According to subjective Bayesianism, what determines what is rational to believe are both the subject’s total evidence and her prior probability distribution, where the only constraint on the latter is consistency with probability calculus. On this view, there is no unique starting point for all subjects even when their total evidence is the same. So, as Kelly (2013, 308) puts it, “even if specifying an agent’s total evidence and her prior probability distribution suffices to pin down some doxastic attitude as the uniquely reasonable one, it does not follow that merely specifying her total evidence suffices to do the same”. So, if subjective Bayesianism is true, then the simplicity objection is sound and U is false.

¹ Endorsing U commits one, as Ballantyne and Coffman correctly note, to the thesis that “whatever fixes your rational attitudes can do so only by fixing what evidence you have” (2011, 1).

It is clear that the uniquer must argue, in response to the simplicity objection, that there is no extra factor, any factor other than evidence, which contributes to the determination of what it is rational to believe—that evidence is the sole determinant of rationality. However, interestingly, what we typically find in works that purport to defend U is not an argument for the thesis that evidence is the sole determinant of rationality but an argument for the thesis that *assuming that evidence is the sole determinant of rationality*, U is true. Here is a case in point, a passage from the opening paragraph of White's (2005) seminal paper the main advertised aim of which is to defend U:

A rational person doesn't believe just anything. There are limits on what it is rational to believe. How wide are these limits? That's the main question that interests me. But a second question immediately arises: What factors impose these limits? A first stab is to say that one's *evidence* determines what it is epistemically permissible for one to believe. Many will claim that there are further, non-evidentiary factors relevant to the epistemic rationality of belief. I will be ignoring the details of alternative answers in order to focus on the question of what kind of rational constraints one's evidence puts on belief. (White 2005, 445)

Given these remarks, it is plausible to take the main question White raises as this: How wide are the limits on what it is rational to believe given the evidence, on the assumption that evidence is the sole factor that determines those limits?

The case is even clearer in Matheson's (2011) defense of U. Matheson explicitly takes evidentialism for granted, according to which what is rational for a subject to believe is determined only by the subject's evidence. Matheson writes:

The falsity of evidentialism would spell trouble for U... However, I will not be examining indirect attacks to U via criticisms of evidentialism, though such critiques do affect the plausibility of U... Rather, I will be assuming the truth of evidentialism and will proceed to assess the prospects of U given that assumption... (Matheson 2011, 364)

So, there are two different questions that might be of interest here. One is whether U is true (*the uniqueness question*), and the other is whether assuming that evidence is the sole determinant of rationality, U is true (*the conditional question*). The permissivist answers the uniqueness question “no”, and the uniquer effectively argues for a “yes” answer to the conditional question. That is, while the permissivist rejects U, the uniquer effectively defends what one might call the *Conditional Uniqueness Thesis* (U*), according to which if evidence is the sole determinant of rationality, then U is true. It is clear, however, that this by itself does not mean that there is any disagreement between the uniquer and the permissivist. The permissivist can consistently agree that an answer to the conditional question is a “yes”, and a commitment to a “yes” answer to the conditional question does not entail any commitment with regard to the uniqueness question.

A failure to make a clear distinction between U and U* has led to an exaggeration of differences and given rise to a good deal of talking past each other. For instance, Ballantyne and Coffman offer a “general recipe” for constructing counter-examples to U, the first step of which is this: “Begin with a possible thinker, who accepts an approach to rationality that allows something other than one’s evidence all by itself to help determine which attitudes are rational for one” (Ballantyne and Coffman 2011, 12). This sounds like a good start for developing the simplicity objection against U. However, Ballantyne and Coffman (2011, 12) also claim that this appeal to “extra-evidential features” relevant to rationality undermine White’s defense of U. However, this latter claim rests on overlooking passages from White like the one quoted above and is false. White’s defense of U is in effect a defense of U*, and as such Ballantyne and Coffman’s strategy against it is bound to misfire.

Here is another example. Kelly (2013) criticizes White’s defense of U by noting that it rests on a failure to distinguish between intrapersonal and interpersonal readings of U. U taken as having only intrapersonal import—Intrapersonal Uniqueness—is a thesis about how a particular body of evidence constrains the number of rational doxastic attitudes for a *single* subject; and thus taken, it is “silent on whether some other individual with the same total evidence might take up a different attitude towards the same proposition that’s fully reasonable” (Kelly 2013, 304). Interpersonal Uniqueness is the

thesis that there is just one rational doxastic attitude *any* individual having a body of evidence might take towards a proposition. Interpersonal Uniqueness is stronger than Intrapersonal Uniqueness. And, as Kelly (2013, 305) notes, it is clear that the uniqueness debate is about whether Interpersonal Uniqueness is true. The simplicity objection, if sound, undermines Interpersonal Uniqueness but is consistent with Intrapersonal Uniqueness given that the extra factor that purports to contribute to rationality (e.g., prior probability distribution) is presumably fixed by the facts about the single subject in question. And, Kelly (2013, 305–6) argues that White’s defense of U is best construed as a defense of Intrapersonal Uniqueness because, if interpreted as a defense of Interpersonal Uniqueness, it fails drastically.

However, if Kelly’s diagnosis about White’s defense of U is on the right track, then we are left with an unsettling question: if it is clear that the uniqueness debate is about Interpersonal Uniqueness, and if White’s defense of U is best construed as a defense of Intrapersonal Uniqueness, then how does White fail to see that his defense of U is simply irrelevant to the uniqueness debate? There are two possible answers to this question. The less charitable answer is that White conflates Intrapersonal with Interpersonal Uniqueness and thus does not recognize the slide from the former to the latter. However, I don’t think even Kelly (2013, 309) would wholeheartedly endorse this answer because, as he openly acknowledges, he does not “imagine that any of [the points he makes] is news to [White]”. The more charitable (and textually supported) answer is that White’s defense of U is in effect a defense of U* and there is no substantive intra/inter-personal distinction that applies to the latter. There is an intra/inter-personal distinction that applies to a uniqueness thesis just in case there is an extra factor other than evidence that contributes to rationality but might not be shared by two different subjects having the same evidence. However, whether there is such an extra factor is irrelevant to a defense of U*: if evidence is the sole determinant of rationality, then if Intrapersonal Uniqueness is true, then Interpersonal Uniqueness is true. So, when it comes to defending U*, the slide from intrapersonal to interpersonal considerations is not fallacious.

The following seems to me a fair description of a portion of the current dialectic. The uniquer’s “official” aim is to give an affirmative answer to

the uniqueness question (that is, the question whether the Uniqueness Thesis is true) but perhaps he has not been as clear as he could have been in signaling the fact that that answer requires a defense of the thesis that evidence is the sole determinant of rationality. Rather, the uniquer typically moves directly to a defense of a “yes” answer to the conditional question (that is, the question whether if evidence is the sole determinant of rationality, U is true), sometimes even without an explicit indication that that answer falls short as an answer to the uniqueness question. The permissivist (qua a proponent of the simplicity objection), on the other hand, gives a negative answer to the uniqueness question by arguing that evidence is not the sole determinant of rationality. However, the permissivist has not been as clear as he could have been in signaling the fact that a negative answer to the uniqueness question does not entail any commitment with regard to the conditional question, which is what the uniquer attempts *in effect* to answer anyway. The upshot is that some confusion surrounding the uniqueness debate might have been avoided if the distinction between the Uniqueness Thesis and the Conditional Uniqueness Thesis were clearly appreciated.

References

- Ballantyne, Nathan and E.J. Coffman. 2011. “Uniqueness, Evidence, and Rationality.” *Philosophers’ Imprint* 11 (18): 1–13.
- Christensen, David. 2007. “Epistemology of Disagreement: The Good News.” *Philosophical Review* 116 (2): 187–217. <https://doi.org/10.1215/00318108-2006-035/>
- Douven, Igor. 2009. “Uniqueness Revisited.” *American Philosophical Quarterly* 46 (4): 347–361.
- Feldman, Richard. 2006. “Epistemological Puzzles about Disagreement.” In *Epistemology Futures*, edited by Stephen Hetherington, 216–236. Oxford: Oxford University Press.
- Kelly, Thomas. 2013. “Evidence Can Be Permissive.” In *Contemporary Debates in Epistemology*, edited by Matthias Steup, Hoboken: John Wiley & Sons.
- Matheson, Jonathan. 2011. “The Case for Rational Uniqueness.” *Logos & Episteme* 2 (3): 359–373. <https://doi.org/10.5840/logos-episteme20112319>
- Schoenfield, Miriam. 2014. “Permission to Believe.” *Nous* 48 (2): 193–218. <https://doi.org/10.1002/9781119420828.ch19>

-
- Titelbaum, Michael. 2010. "Not Enough Evidence There There." *Philosophical Perspectives* 24 (1): 477–528.
- White, Roger. 2005. "Epistemic Permissiveness." *Philosophical Perspectives* 19 (1): 445–459. <https://doi.org/10.1002/9781119420828.ch18>

The Personite Problem and the Stage-Theoretic Reply

Harold Noonan*


Received: 1 July 2021 / Revised: 28 August 2021 / Accepted: 15 October 2021

Abstract: Personites are shorter-lived, person-like things that extend across part of a person's life. Their existence follows from the standard perdurance view of persons. Johnston argues that it has bizarre moral consequences. For example, it renders morally problematic spending time learning a difficult language in anticipation of going abroad. The crucial thought is that if persons have moral status so do personites. Johnston argues for this claim. Kaiserman responds, on behalf of stage theory, that this only works on a perdurantist account. This is a conservative response to the problem. It seeks to show that retaining the ontology of perdurantism one can resolve the difficulty by a semantic change. I show that the personite problem can be reworked as an argument against stage theorists. The stage theorist can respond by rejecting an assumption of the reasoning. But if it is acceptable for him to do so the perdurantist can reject this assumption too, which is enough by itself to block Johnston's argument. Thus, for all it helps with the personite problem, stage theorists might as well be perdurantists.

Keywords: Personites; perdurance; stage-theory; moral status; Johnston.

* University of Nottingham

 <https://orcid.org/0000-0001-5538-5444>

 University of Nottingham, Department of Philosophy, Room C41 Humanities, University Park, Nottingham, NG7 2RD, UK

 Harold.Noonan@nottingham.ac.uk

© The Author. Journal compilation © The Editorial Board, *Organon F*.



This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International Public License (CC BY-NC 4.0).

1. Introduction

Personites are shorter-lived, very person-like things that extend across part, but not the whole, of a person's life. The term is introduced by Johnston (2016; 2017). That there are personites is a consequence of the standard perdurance view of persons championed by David Lewis (1986). Johnston argues that the existence of personites has bizarre moral consequences. For example, if there is a personite now coinciding with me which will not exist tomorrow this renders morally problematic my planned visit to the dentist today, since the personite, unlike me, will suffer pain today but not live long enough to experience any gain. The same reasoning renders morally problematic spending time learning a difficult language in anticipation of going abroad. Again, in accordance with this reasoning, taking a child to the dentist or making her do her homework become morally problematic actions.

The crucial thought in the background of these reflections is that *no relation (e.g., being a child of, being the wife of, being the creation of, being part of) a sentient being has to another can deprive it of the right to be counted (as a patient) in the moral calculus.*¹ Hence the relation a personite has to a person cannot do so. Personites, if they exist, have moral status.

Johnston gives a more precise argument that personites have moral status. Kaiserman (2019) argues, on behalf of the stage theory (Sider 2001; Hawley 2001), that this only works on a perdurantist account. On a stage-theoretic account, he argues, it fails. This is a conservative response to the problem. It seeks to show that one can retain the ontology of the perdurantist and resolve the difficulty by a semantic change. I show that the personite problem for perdurantism can be reworked as an argument against stage theorists. The stage theorist can respond by rejecting an assumption of the

¹ This entails that other things being equal, the suffering of two sentient beings (which considered individually have moral status) is worse than the suffering of one, however they are related. (Hence it rules out a hedonic moral calculus which treats suffering as like mass, and so not additive in the case of coinciding things. Perhaps this most primitive form of hedonism, to which 'the most telling objection has been regarded as being that it treats persons as mere receptacles of good-making features', is the only refuge for the believer in personites [Johnston 2017, 642, fn. 5]).

reasoning. But if it is acceptable for the stage theorist to do so the perdurantist can reject this assumption too, and this enough by itself for him to block Johnston's argument. Thus, stage theorists are no better placed than perdurantists to deal with the personite problem. For all it helps with the personite problem, stage theorists might as well be perdurantists.

2. The personite problem

First, I state Johnston's challenge to perdurantists more carefully in the form Kaiserman discusses (in the following 'x in w is a duplicate of y in v' means 'x in w instantiates all the same intrinsic properties as y in v'):

- (1) For all possible worlds w and possible objects x, if x is a person in w then x has moral status in w.
- (2) For all possible worlds w and v and possible objects x and y, if x in w is a duplicate of y in v, then x has moral status in w if y has moral status in v.
- (3) For all personites x, there are a possible object y and possible world w such that y is a person in w and y in w is a duplicate of x in the actual world.
- (4) Therefore, all personites have moral status.

3. The stage-theoretic response

Kaiserman objects to premiss (3) on behalf of stage theory. The objection is obvious. Stage theory has the same ontology as, but a different semantics from, perdurantist theory. It gives an account of temporal predication in terms of temporal counterpart relations. Hence, according to the stage theorist persons are instantaneous person-stages; that is just a matter of what 'person' means. So, in fact, no *non-instantaneous* proper part of a maximal sum of person-stages linked pairwise by personal unity is a duplicate of any (even merely possible) person since no non-instantaneous thing can be a duplicate of any instantaneous thing. So no personite can be a duplicate of a person. Premiss (3) is false.

Unfortunately, this reply to Johnston does not prevent a reworking of the argument for the possession of moral status by personites which

threatens the stage theorist as much as the original threatens the perdurance theorist.

To see this, we need only recall that stage theory is the temporal analogue of Lewisian modal counterpart theory and what follows from that. As Kaiserman explains, according to stage theory where I am now, i.e., where the instantaneous stage denoted by the token of ‘I’ I am now uttering is, there is not “a multiplicity of entities with different counterpart relations” (Kaiserman 2019, 220). There is just one object coincident with me—*me*.² But there are a variety of different counterpart relation in which that one object stands to others. ‘I will be in Hungary next year’ is true (if it is) if there is a personal temporal counterpart of me which is in Hungary next year. This may be true even if there is here no human animal/body which is in Hungary next year. For example, this will be so if I have a brain transplant before I depart and the psychological continuity account is the correct account of personal identity, i.e., of the personal temporal counterpart relation. Thus, even though every human animal is a person-stage, and ‘the human animal here’, as uttered by me now, denotes the person stage ‘I’ denotes, ‘the human animal here will be in Hungary next year’, uttered by me now may not be true, since the term ‘the human animal’ may evoke the human animal counterpart relation. This is, of course, exactly analogous to what Lewis says about *de re* modal claims.

So consider the following scenario (elaborated from Johnston). I am going to Hungary next year (at time t_2). Before, at time t_1 , I will have a brain-transplant (so when I say ‘I am going to Hungary next year’ I mean ‘the composite of my brain and new body is going’). My old body will be disposed of. So the animal here now will be no more after t_1 . Before t_1 you, who have my best interests at heart, will have to choose whether to make me learn Hungarian before the brain transplant, knowing that it will be unpleasant for me to do so, but aware that I will benefit greatly once I am in Hungary.

Now the reworked argument against stage theory can be given. According to the stage theorist: I am a person, so I have moral status. The only

² In fact, this proposition is the main focus of Kaiserman’s (2019, 219-20) replies to objections. That he is exactly right about this is also precisely the crucial premiss in my objection to him.

thing I am coincident with is me. I am coincident with the human animal here. So it is me. Leibniz's Law holds. So the human animal here has moral status. You have the ability to cause me, before my departure to Hungary and also before the brain transplant, pain (by making me learn Hungarian), i.e., you have the ability to ensure that there is a future personal counterpart of me existing before the brain transplant which suffers pain.³ That will also be an animal counterpart of me, since no brain-transplant will have taken place when it exists. So it will be an animal counterpart of the animal here. So the animal here will exist at that time, just as I will, and will be in pain. Now suppose the painful future existence of that counterpart of me will ensure that the person-stages in Hungary related to me-now by the personal counterpart relation will be pain-free (I will be able to follow the lessons in school and mix freely with the Hungarian children). Then if I say now, 'I will suffer pain before my brain transplant but will benefit by being pain-free when I am in Hungary' I will speak truly. But although I will speak truly if I say, 'Before my brain transplant the human animal here will suffer', it will not be true for me to say, 'the human animal here will benefit subsequently'. But I am the human animal here and I have moral status. So the human animal here has moral status. So choosing to inflict the future pain on me before my brain transplant in order to prevent subsequent suffering in Hungary is morally problematic, since it will ensure the infliction of pain on the animal here from which it will never benefit. If you make that choice you are choosing to make it true that something existing now which is endowed with moral status will suffer pain in the future from which it will never benefit.

So goes the argument that the stage theorist as well as the perdurantist faces the personite problem. Of course, corresponding to every personite the perdurantist recognises and must regard as a duplicate of a possible person, the stage theorist must recognise a temporal counterpart relation. For, as noted, the perdurantist and stage-theorist have the same ontology. So in this argument 'the animal here' can be replaced by any singular term which

³ Why do you have this ability? Perhaps because I am a young child, and you are my parent and for family reasons I am being sent to Hungary next year to live with my grandparents. Understand the scenario in this way.

according to the perdurantist refers to an appropriately short-lived personite.

The response the stage-theorist must make is obvious. He must channel his inner Lewis and deny that I can infer from the joint truth of ‘I have moral status’ and ‘I am the animal here’ that ‘the animal here has moral status’ is true. He must say that ‘has moral status’ is inconstant in denotation (in Lewis’s sense [Lewis 1971; Lewis 1986, 248ff]).⁴ When a token of ‘has moral status’ is attached to a token of a subject term (e.g., ‘I’), and/or uttered in a context, which evokes (to use Lewis’s language) the personal counterpart relation, it denotes the class of person-stages, i.e., the class of persons, so the token sentence utterance is true. When a token of ‘has moral status’ is attached to a co-designating token which evokes some other, morally insignificant, temporal counterpart relation (like the animal or body temporal counterpart relation) it denotes the empty class, so that token sentence utterance is false, despite the co-designation. So, although ‘I have moral status’ is true, ‘this animal here has moral status’ is false, even though ‘I am the animal here’ is true.

What if the stage-theorist does not respond in this way? Then he is committed to saying that all three of the following propositions are true (expressed by sentences uttered in a single context where the only temporal counterpart relation evoked is the one for animal persistence): (a) this animal here has moral status, (b) if tutoring in Hungarian goes ahead this

⁴ Kaiserman does not speak in Lewisian terms of ‘inconstancy’. But he does say that the stage theorist should relativized the predication of temporal properties to a choice of counterpart relation. Nor does he enquire whether the (crucial) predicate ‘has moral status’ is inconstant in denotation. He does, however, say that the stage theorist should insist that there is a particular counterpart relation which is such that what I ought to do depends on what is true of me relative to that counterpart relation—this is the one that matters. Thinking all this through in Lewisian terms and responding to the reworking of the personite argument I gave leads, I argue, to the conclusion that the stage theorist should say that ‘has moral status’ is inconstant in denotation. But if he can say this so can the perdurantist. Note that in the modal case Lewis does not think that modal predication is inconstant because counterpart theory is correct; rather, he thinks the inconstancy is a fact that any account of modal predication must accommodate; acceptance of inconstancy does not require acceptance of counterpart theory.

animal will be caused to suffer before the brain-transplant, (c) this animal will never benefit. So, to conform to the common-sense view that in insisting on the tutoring in Hungarian before the departure to Hungary you (my parent) are acting wholly unproblematically morally,⁵ the stage-theorist must deny that it *follows* from these three propositions that this animal's suffering is in any way morally problematic. He must say that if we are told that something that possesses of moral status has had suffering inflicted on it from which it has not benefited and will not benefit, we cannot *infer* that that action has thereby any moral cost.⁶ Whereas if we deny the constancy of 'has moral status' we can endorse this inference.

But, of course, if the stage theorist can deny that 'has moral status' is constant in denotation so can the perdurance theorist. And if one cannot infer from the truth of 'X has moral status' and 'X=Y' that 'Y has moral status' is true, a fortiori one cannot infer from the truth of 'X has moral status' and the truth of 'X is a (mere) duplicate of Y' that 'Y has moral status' is true. So the perdurance theorist can acknowledge the existence of personites and deny their moral status, i.e., deny premiss (2) of Johnston's argument.

If this is deemed unsatisfactory a more drastic response to the personite problem is needed, as Johnston argues: perhaps the perdurantist/stage-theoretic ontology must be rejected, perhaps even the ontology of liberal endurantists (Kaiserman 2019, 219) along with it, and perhaps any ontology consistent with naturalism. That discussion is for another place.

⁵ Which must be so in this case unless all education is somehow morally problematic! (So, of course, to give the argument against the stage theorist there is no need to consider Johnston's Hungarian language learning scenario. Just consider taking a child to the dentist or making her do her homework.)

⁶ Note that to say such suffering is a moral cost is not, of course, to say that it must be immoral to inflict it. It is no part of ordinary moral thought that this follows. It is no part of ordinary moral thought that it cannot in any circumstance be morally justified, on balance, to inflict suffering from which it will not benefit on a possessor of moral status. Rather, it is part of common-sense morality that such circumstances are common (for example, "the needs of the many outweigh the needs of the few, or the one" [Mr Spock, Star Trek]).

References

- Hawley, Katherine. 2001. *How Things Persist*. Oxford: Oxford University Press.
- Johnston, Mark. 2016. "Personites, Maximality and Ontological Trash." *Philosophical Perspectives* 30(1): 198–228. <https://doi.org/10.1111/phpe.12085>
- Johnston, Mark. 2017. "The Personite Problem: Should Practical Reason Be Tabled?" *Noûs* 51(3): 617–44. <https://doi.org/10.1111/nous.12159>
- Kaiserman, Alex. 2019. "Stage Theory and the Personite Problem." *Analysis* 79(2): 215–222. <https://doi.org/10.1093/analys/any074>
- Lewis, David. 1971. "Counterparts of Persons and Their Bodies." *Journal of Philosophy* 68(7) 203–11. <https://doi.org/10.2307/2024902>
- Lewis, David. 1986. *On the Plurality of Worlds*. Oxford: Blackwell.
- Sider, Theodore. 2001. *Four Dimensionalism: An Ontology of Persistence and Time*. Oxford: Oxford University Press.

BOOK REVIEW

Catarina Dutilh Novaes: *The Dialogical Roots of Deduction: Historical, Cognitive and Philosophical Perspectives on Reasoning*
Cambridge: Cambridge University Press, 2021, xiii + 271 pages


Jaroslav Peregrin*

What is reasoning and what is it good for? An almost self-evident explanation may run as follows: reasoning helps us extend our knowledge by equipping us with new pieces of knowledge drawn out of our older pieces. And as such, it is clearly useful and therefore it is obvious why the human brain has developed to support it. An individual capable of reasoning—and hence capable of extending her knowledge—is clearly superior to one who is not, and hence no wonder the former overtakes the latter in the evolution race. Reasoning, viewed from this perspective, is an individual matter; a matter that has to do with the maintenance of information that is stored in one’s mind/brain. Any kind of interpersonal reasoning, aka argumentation, is then the outcome of the individual reasoning coming into the open—for once we are capable of reasoning, it may be useful to make one’s reasoning known to one’s peers and to confront one’s own ways of reasoning with those of others.

This plausible sounding explanation, however, has been challenged in recent years by several experts. Hugo Mercier and Dan Sperber (Mercier & Sperber, 2011; Sperber & Mercier, 2012) put forward the thesis that public argumentation is more basic than individual reasoning—that rather than the former being an externalization of the latter, the latter is an internalization of the former. Mercier & Sperber (2017) went on to develop a comprehensive theory of the origins, the nature, and the evolutionary rationale of reasoning.

Catrina Dutilh Novaes’ book pursues a similar goal. Her particular focus is on *deductive* reasoning, and she strives to show that the basis of any deduction

* Institute of Philosophy of the Czech Academy of Sciences

 Institute of Philosophy of the Czech Academy of Sciences, Jilská 1, 110 00 Prague 1, Czech Republic

 peregrin@flu.cas.cz



lies in dialogue and hence that reasoning in the public sphere is more fundamental than any private ruminations.

Part I of the book is called *The Philosophy of Deduction*. In the first chapter the author describes her understanding of deduction. She characterizes it as having three attributes:

- necessary truth-preservation (a deductive argument can not lead us from true premises to a false conclusion);
- step-wise structure (a deductive argument is a chain of perspicuous steps); and
- bracketing belief (a deductive argument cannot be influenced by collateral beliefs).

Given these, Dutilh Novaes considers three fundamental questions concerning deduction:

- *Where Is Deduction to Be Found?* Here the author's answer is straightforward: deductive reasoning is not something ubiquitous, it is "predominantly instantiated in mathematics and in some other regimented contexts of argumentation, such as philosophy." (p. 12)
- *What Is the Nature of Deductive Necessity?* Here the author does not reach an unambiguous answer: "We may never come to a fully convincing account of the necessity involved in deductive arguments." (p. 17)
- *What Is the Point of Deduction?* Here the author looks first at what it is not, and then defers her positive answer to the rest of the book: "Deduction does not seem to be a particularly suitable way to produce new information, given that it is non-ampliative, and it does not seem to be a reasonable guide for managing our beliefs and thoughts either." (p. 21)

In the next chapter Dutilh Novaes explains her motivation for exploring the "roots of deduction"—and she stresses the necessity to distinguish ontogenetic, phylogenetic and historical roots. She also foreshadows what will govern the upcoming investigation of the book.

In Chapter 3 the author analyzes the kind of dialogue that she holds must underlie deduction. She surveys the existing attempts at capturing deduction via dialogic (or game-theoretic) means, in Hintikka's game-theoretic semantics and especially in Lorenzen's dialogic logic. She concludes that the kind of dialogue she is after is best characterized as that between characters she calls

a “Prover” and a “Sceptic”, whose roles display, in her view, an optimal mixture of cooperation and adversariality.

In Chapter 4 the author checks whether the notion of deduction that grows out of these dialogical roots displays the three key features of deduction she identified earlier. She concludes that while the necessary truth-preservation grows out of the adversarial dimension of dialogue (the Sceptic persists in challenging the Prover until even the most marginal cases have been covered), the perspicuity grows out of the cooperative dimension (the Prover tries to make the deduction as transparent as possible for the Sceptic). The belief bracketing is then connected with the ability to assume perspectives different from one’s own. At the end of this chapter the author sketches the route from dialogue proper to deduction proper—via the internalization of the Sceptic by the Prover. In the same chapter Dutilh Novaes considers some of the most basic philosophical problems related to deduction: proof-theory vs. model theory, the normativity of logic, logical paradoxes, structural rules of deductive systems and logical pluralism.

Part II of the book is devoted to topics from the history of logic. Here Dutilh Novaes attempts to show that our facility of deduction originated out of various kinds of dialogues as a matter of fact. As in the rest of the book, a positive quality of her exposition is the breadth of literature she makes use of. In Chapter 5 she considers deduction in the context of what we know about ancient Greek mathematics and dialectics. In Chapter 6 she concentrates on Aristotelian Syllogistic (as the first complex logical system), also touching upon the contexts of ancient India and China. Chapter 7 then covers the role of deduction in the long period from the Middle Ages to our Modern times. Everywhere she finds deductive and logical practices to be underlain by dialogical and interactional ones.

Part III of the book, by my lights, is the most interesting. Here Dutilh Novaes considers the problem of deduction and dialogue from the viewpoint of human cognition. In Chapter 8 the author surveys some well-known ways in which human reasoning tends to fail systematically. This evidence poses problems for those who maintain that deductive reasoning is a direct evolutionary adaptation—for if it were, it would be hard to imagine that it would be so awkward. A possible explanation might be that it is not itself an adaptation, but rather the by-product of an adaptation. A significant fact is that the results of reasoning are significantly improved if the reasoning proceeds interactively in a group of individuals.

In Chapter 9 the author argues that on the ontogenetic level, the skills needed for deductive reasoning are originally acquired by means of dialogues and are significantly bolstered precisely by the dialogical nature of the contexts in which they are acquired.

In Chapter 10 Dutilh Novaes investigates the phylogenetic roots of deduction, to conclude, as the reader would expect, that here again the skills for deductive reasoning derive from those for dialogic communication. A great deal of this chapter is devoted to the author's polemic with the concurrent view of Mercier and Sperber. Like Dutilh Novaes, Mercier and Sperber claim that reasoning proper originates on an interpersonal level, because it requires the confrontation of two different human discursive abilities. We are very good, Mercier and Sperber insist, at finding reasons for our own view - not in an impartial manner, but with a strong 'my side bias'. On the other hand, we are also very good at checking and challenging reasons put forward by others - and when these two abilities are played against each other in a dialogic situation, reasoning as the search for truth is likely to be the result.

This appears to be nearer to Dutilh Novaes' theory than she herself is willing to acknowledge. The difference is that, for Mercier and Sperber, the abilities which yield reasoning as a by-product are genetic adaptations—they have to do with social coordination. In contrast, Dutilh Novaes insists that deductive reasoning is what Heyes (2018) dubbed a cognitive *gadget*—the corresponding abilities are not anchored in our genes, but are a matter of cultural learning. However, both sides agree on what I take as the most central message—namely that deductive reasoning is not itself an adaptation.

In the final chapter of the book the author strives to show that even mathematical practice shows clear signs of its dialogical origins.

The whole book is interdisciplinary in the best sense of the world: it brings together—and interconnects—relevant results of logic, philosophy, psychology, evolution theory and history in a way that casts fresh light on the relationship between argumentative and deductive reasoning. I think that the case for the dialogical roots of deduction has been made quite persuasively.

The only argument that I find less convincing is the author's repudiation of the view of Mercier and Sperber. Concerning their view Dutilh Novaes writes: "Reason must be an adaptation, but if conceived as having the function of supporting the cognitive processes of the lone reasoner, it does not seem to perform this function very well. So, there must be a different function that reason is in fact responding to, given that it cannot be anything other than an adaptation"

(p. 193). And further: “Prima facie, to argue for the adaptive nature of reason seems like a tall order in view of the numerous empirical findings suggesting that human reason is ‘biased and lazy’. ... But Mercier and Sperber go on to argue that these two features are in fact advantageous for the function of reason as socially conceived” (*ibid.*)

I think that we must distinguish between two varieties of “reason as socially conceived” (or argumentation). Let me call the first of them “Socratic”: this is the argumentation which aims at an impartial seeking of truth, where reason acts “as a judge” (to use the metaphor of Haidt, 2001). The other variety of argumentation is “sophistic”: this aims at defending one’s pre-given views cost what it may (it acts as a “defense lawyer”).

In my view, Dutil Novaes wrongly portrays Mercier & Sperber as claiming that Socratic argumentation is an adaptation. As I read them, what they claim is that it is a kind of sophistic argumentation that is an adaptation, and that Socratic argumentation is its by-product. Defending one’s position (cost what it may) is an adaptation; and seeking flaws in another’s defense is a counter-move to this adaptation. It is only when these two adaptations are played against each other that “Socratic” argumentation may arise.

Nevertheless, the book is definitely a valuable contribution to current discussions both about the nature of logic and mathematics and about the nature of human reason.

References

- Haidt, Jonathan. 2001. “The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment.” *Psychological Review* 108 (4): 814–34.
<https://doi.org/10.1037/0033-295X.108.4.814>
- Heyes, Cecilia. 2018. *Cognitive Gadgets: The Cultural Evolution of Thinking*. Cambridge (Mass.): Harvard University Press.
- Mercier, Hugo and Sperber, Dan. 2011. “Why Do Humans Reason? Arguments for an Argumentative Theory.” *Behavioral and Brain Sciences* 34 (2): 57–111.
<https://doi.org/10.1017/S0140525X10000968>
- Mercier, Hugo and Sperber, Dan. 2017. *The Enigma of Reason*. Cambridge (Mass.): Harvard University Press.
- Mercier, Hugo and Sperber, Dan. 2012. “Reasoning as a Social Competence.” In *Collective Wisdom: Principles and Mechanisms*. Edited by Hélène Landemore and Jon Elster, 368–92, Cambridge: Cambridge University Press.
<https://doi.org/10.1017/CBO9780511846427.016>