ACTA
PHILOSOPHICA
FENNICA
VOL. 97 2021

Reasons and Responsibilities

Proceedings of the Philosophical Society of Finland Colloquium 2020

Edited by
INKERI KOSKINEN
& TEEMU TOPPINEN

SOCIETAS PHILOSOPHICA FENNICA

ACTA PHILOSOPHICA FENNICA

Acta Philosophica Fennica is published by the Philosophical Society of Finland. Since the inception of the series in 1935 it has been the forum for much of the best philosophical work in Finland. In 1968−1981 it was distributed by the North-Holland Publishing Company (Amsterdam), and since 1981 by the Academic Bookstore and Bookstore Tiedekirja (Helsinki).

Information for Authors

The Acta series publishes shorter and longer monographs as well as collections of articles in philosophy. All subfields of philosophy and all philosophical traditions fall within its intended scope. Authors should send their contributions to

Acta Philosophica Fennica,
Department of Philosophy,
P.O. Box 24 (Unioninkatu 40 A),
FI-00014 University of Helsinki,
Finland.

Before submission, authors are requested to contact the Editor,
e-mail: leila.haaparanta@helsinki.fi

Subscription Information

Permanent subscriptions can be placed directly with Bookstore Tiedekirja, Snellmaninkatu 13, FI-00170 Helsinki, Finland, tel. +358−9−635 177, email: tiedekirja@tsv.fi, www.tiedekirja.fi. Other orders can be placed online with Bookstore Tiedekirja, www.tiedekirja.fi.

ISBN 978−951−9264−94−3
ISSN 0355-1792

Hansaprint Oy
Helsinki 2021

Table of Contents

Preface

In January 2020 the Philosophical Society of Finland organised its yearly colloquium, which for the second time was held in a new format. As the philosophical community in Finland has become linguistically much more diverse than it once used to be, the important task of maintaining our national languages as languages of philosophy must be balanced with the need to include everyone. Therefore, every other year the presentations in the society's colloquium can now be given also in English, in addition to Finnish and Swedish. As we have seen both in 2018 and in 2020, the new format not only better reflects our philosophical community as it is today, but also attracts contributors from many other countries to which Finnish philosophers have always had close ties.

FiPhi 2020 was held in Helsinki on 9–10 January – that is, about two months before our lives changed in ways we were completely unable to anticipate when listening to presentations in small rooms without wearing face masks, or continuing our discussions in crowded restaurants and bars. Now that we are writing the preface to this collection, the vaccination coverage in Finland is getting near to 80%, and we are cautiously returning to face-to-face teaching. However, academic meetings have most likely changed permanently, as we have all been forced to get used to online and hybrid conferences and workshops. Fewer of us will be taking intercontinental flights just to attend a conference, and this must be greeted as a positive change, as the kind of conference travel that was the norm before the pandemic was not ecologically sustainable. However, in mostly local conferences such as FiPhi, we do hope to return to a new normal that resembles the old one. With some luck Finnish philosophers and some

of our international friends will be able to gather in Oulu mostly in person for FiPhi 2022.

During the state of emergency, and before the vaccinations started, we opened a call for papers based on talks given at FiPhi 2020. This collection is the result of that call. The diversity of the topics it covers – from environmental ethics to philosophy of physics – reflects the diversity of philosophical research in Finland.

Ninni Suni addresses the issue of doxastic responsibility. On one hand, we do seem to hold each other as responsible for our beliefs ("You shouldn't believe everything that's on YouTube!"); on the other, beliefs do not seem to be directly under our control, and we standardly take being responsible to imply being in control. Suni proposes that we can do justice to our intuitions without adopting any control condition on doxastic responsibility. She outlines an "attributionist" view on which someone is responsible for a belief when the belief reflects what Suni calls their epistemic perspective (a set of dispositions to notice, explain, and respond to evidence).

Ilkka Pättiniemi, Rami Koskinen, and Ilmari Hirvonen review current epistemology of modality. They argue that some of the major accounts in this literature would work only if we had access to the kind of knowledge that we seem to be unable to produce. As a satisfactory epistemology of modality should be applicable by us as the limited beings we are, Pättiniemi, Koskinen and Hirvonen find the major accounts lacking. They then defend a framework of relative modality as a partial remedy to the situation.

Jaakko Reinikainen examines the problem of semantic objectivity – the problem of explaining how it is that our linguistic practices are suitably constrained by how the world is – as it arises in the context of Robert Brandom's work (especially in his *Making it Explicit*). Having highlighted some relevant tensions in Brandom's account of semantic objectivity, Reinikainen suggests that these tensions may be alleviated by articulating the way in which the idea that both facts and attitudes are conceptually structured, more in focus in Brandom's later work, is implicitly operative also in *Making it Explicit*.

Heidi Haanila argues that dream research offers valuable tools for the study of self. She discusses the distinction between the experiential and the reflective self, and focuses on the ways in which the dream self differs from the waking self in terms of both experiential and reflective self. She then defends the idea that the differences between the ways in which aspects of self are organized in dreaming and in typical waking self-consciousness enable us to distinguish between different aspects of self, and to identify some necessary features of self.

Joonas Martikainen examines aphonia, that is, the loss of a political voice of one's own, in the light of Maurice Merleau-Ponty's existential phenomenology. He argues that aphonia is qualitatively different from both a lack of opportunities for democratic participation and a lack of communicative capabilities needed for effective participation, and calls for alternative ways to make political participation possible for marginalized groups through a "therapeutic" approach to political inclusion.

Agostino Cera focuses on the notion and idea of anthropocene, the proposed geological epoch characterised by significant human impact on Earth's geology and ecosystems. His aim is to prove that it is a threshold concept that capable of shaking the foundation of sciences such as geology. He contrasts his arguments about the anthropocene as an epistemic hyperobject to Carlos Gray Santana's recent work and argues that geology struggles to fully grasp the idea of anthropocene.

Simo Kyllönen defends a harm-based solution to the Non-Identity Problem, a much-discussed problem in ethics, which concerns making sense of cases where a certain way of acting (e.g., an action that results in a future person's quality of life being extremely low) seems wrong, but arguably no one is made *worse off* (because without the action the future person would not exist). Kyllönen articulates what the calls the *Additional reasons* account of harm, which he argues allows for a promising solution to the problem.

Teemu Tauriainen critically examines truth pluralism. While different forms of truth pluralism have been suggested, most of them commit to domain reliance, according to which different ways of being true are tied to discourse do-

mains rather than individual sentences. As a result, the truth of different types of sentences is accounted for by their domain membership. Tauriainen argues that in the standard domain reliant pluralist frameworks some sentences end up being both true and false, thus conflicting with both standard laws of non-contradiction and identity.

Jan Hauska addresses the metaphysical underpinnings of the principle of the composition of forces, which plays a prominent role in Newtonian physics. As assigning reality to all the forces mentioned by the principle leads to serious difficulties, various ideas have been put forward about which of the forces exist and how they are related. After critically examining a recent suggestion due to Olivier Massin, Hauska proposes a construal of the principle in terms of powers.

It is an honor to be able to publish this collection in *Acta Philosophica Fennica*. We are grateful to Eero Kaila, the secretary of the Philosophical Society of Finland, for preparing the layout of the volume. We would also like to thank Ilkka Niiniluoto for his long service as the editor of the series, and welcome the new editor, Leila Haaparanta.

Inkeri Koskinen & Teemu Toppinen

Doxastic Perspective and Responsibility for Belief

NINNI SUNI

1 Introduction: The problem of doxastic responsibility

Our everyday practices reveal a commitment to the idea of doxastic responsibility. For instance, we say things like: "You shouldn't believe everything that's on YouTube!" or, "You ought to believe the best scientific evidence about climate change." The idea of a responsible epistemic agent seems to be in the background when we say things like: "She gets all her news from the yellow press," or "He's read multiple studies on the subject." These statements refer to epistemic norms, which in turn identify credible sources, good epistemic practices and trustworthy epistemic agents. Given that we are social beings that rely on each other for much of the information that we acquire, this makes sense. A practice for tracking whom to trust is vital. But simply evaluating beliefs as true of false is not enough; rather, we are interested in whether the belief is *attributable* to the agent, not merely an outcome of circumstantial luck. The idea of doxastic responsibility thus implies that beliefs are in some sense agentive and that they are subject to a prescriptive doxastic ought. [1]

However, responsibility as commonly understood applies to actions which seem to require voluntary control, and this prompts skeptical arguments against doxastic responsibility. The classical formulation of the skeptical position is by William Alston. He argued that deontological concepts such as

[1] It is important to emphasize at the outset that this does not mean *moral* responsibility for beliefs, but a distinctively epistemic assessment that targets one's epistemic agency, e.g., an assessment of one's credibility (cf. Kauppinen 2018).

duty, obligation, and responsibility apply only when we have direct voluntary control, and beliefs, at least in paradigmatic cases, are not under direct voluntary control (Alston 1988). Alston's original argument was directed at a specific view in epistemology, epistemic deontology, which seeks to give the meaning of epistemological concepts such as justification and warrant in terms of deontic concepts like duties and obligations. However, the worry generalizes beyond this specific view. Insofar as responsibility requires voluntary control, beliefs do not seem to be the kinds of things for which we can be responsible. The implication is that a natural reading of our everyday language regarding doxastic oughts and responsibility must be revised, resulting in an error theory of such language.

The pull of Alston's argument can be illustrated by what Chrisman (2008) calls the *no rewards principle* (NRP):

> NRP: No matter how large the reward, S cannot simply decide to believe that *p* in order to collect that reward.

Suppose I offered you 100 million euros to believe that Bernie Sanders is the President of the United States. The offer is tempting, but try as you might, you cannot change your belief simply by deciding to do so. By contrast, were I to offer a large reward for doing something you ordinarily would not do, say, eating rotten meat, it is up to you to decide to do so and collect the reward. Moreover, McHugh (2012) points out that the problem is not just with beliefs that you know to be false, but also beliefs for which you have no evidence whatsoever. Suppose that you were offered a large reward for believing that it rained on Aristotle's 30th birthday. You have no reason to believe this, and the reward cannot provide you with one, being of the wrong kind. It is impossible to form the belief solely for a reward. In other words, the no rewards principle seems to show that doxastic involuntarism (DI) is true:

> DI: Doxastic states are not under effective voluntary control.

The skeptical position can then be summarized in the Anti-Deontology Argument. It is usually formulated in terms of the *ought implies can principle* (OIC), which has independent plausibility.

The Anti-Deontology Argument:

P1 If Doxastic Responsibility (DR) is true, then Doxastic Voluntarism (DV) is true

P2 DI

P3 DV is false (from P2)

C DR is false (from P1, P3)

The argument seems valid, and both P1 and P2 have strong intuitive appeal, but there are still many ways to resist it. Some deny P2, that is, endorse some form of doxastic voluntarism (Ginet 2001, Ryan 2003, Steup 2000, 2008, 2017), while others resist the ought-implies-can principle backing P1, arguing that the principle does not always hold (Sinnott-Armstrong 1984). The most headway can be made, however, by unpacking P1, which relies on an implicit premise. The missing premise is this:

P1* Voluntary control is necessary for doxastic responsibility

This paper assesses the different ways P1* has been resisted by relating the solutions to the conditions of responsibility operating in the background, sometimes implicit, sometimes explicit. I will start with a view which denies the premise by arguing that the doxastic ought does not require that the agent can follow it, moving then to views which deny only the 'voluntary' part of the premise, arguing that there is a distinctively epistemic kind of control which is necessary to doxastic responsibility. I will argue that both of these views run into problems which can be averted by adopting a specific view of responsibility which rejects control as a necessary condition for responsibility, and I sketch a way towards such a view. The idea is that beliefs are agentive because they reveal the doxastic perspective from which they were formed and they are therefore attributable to the agent in the responsibility-implying sense. I will conclude by considering a possible objection.

2 The No-Ought-Implies-Can strategy

The first strategy of arguing against P1* is to deny that ought implies can in epistemic context (Feldman 2001, Kornblith

2001, Chrisman 2008). I will call it the No-Ought-Implies-Can strategy, or NOIC for short. The main insight within this strategy is that in epistemic context 'ought' refers to standards of evaluation.

The idea can be fleshed out in different ways. Feldman (2001) argues that doxastic oughts are akin to role-oughts that flow from the roles we play in social life, as teachers, parents, friends, and so on. What is important in this context is that these oughts do not seem to imply can: an incompetent teacher still ought to explain things clearly, but she cannot, and a bad parent ought to take better care of his children, even if he cannot. Feldman suggests that we should understand doxastic oughts as flowing from our roles as believers (Feldman 2001, 675): we ought to form our beliefs according to our evidence, rather than wishes or fears, even if we cannot help forming wishful beliefs from time to time.

The problem with Feldman's proposal is that unlike role-oughts, epistemic oughts seem to be categorical, like moral oughts (Kornblith 2001). Many of the roles we occupy are optional: an incompetent teacher can quit her job in order to pursue a career that brings out her talents; some people refrain from having children because they believe they would not be able to take good care of them. By contrast, Kornblith argues, in the epistemic case we are not only making the conditional claim that if someone wants to be a good believer, she ought to believe so-and-so; we want to endorse an even stronger claim, that every individual ought to believe according to the evidence. Take a parallel case in the moral sphere: suppose that Kelly is a thief. Does it follow that as a thief she ought to steal as much as possible? Of course not. The moral obligation not to steal is not conditional on the role she plays, nor on the reasons she might have for occupying that role. Similarly, epistemic oughts are retained even if some of us underperform as an epistemic agent.[2] Kornblith argues that

[2] Note that Kornblith's argument seems to suppose that role-oughts are conditional on our goals—the norms of being a good teacher apply to me because I aim to be a good teacher—and that this is the crucial difference between role-oughts and categorical oughts. But this need not be so: role-oughts do not seem to require goals in order to apply. Suppose that John does not care at all whether he is a good teacher or not. We would still say that as long as he actually is a teacher, the teacher-role-oughts apply to

the only way for Feldman to explain this contrast is to appeal to the fact that being a believer is not something we can escape. Unlike being a teacher, we cannot simply step out of our roles as epistemic agents. But inescapability is also an unsatisfactory explanation of the categoricity of certain obligations. To see why, Kornblith asks us to suppose that Kelly is not just a thief, but a kleptomaniac, who cannot escape being a kleptomaniac. It still does not follow that she has a moral obligation to steal.

Kornblith's argument is unsatisfactory as it stands because it seems plausible that kleptomania is not the kind of role which issues oughts or norms.[3] But the argument can be revised to make it more robust. Inescapability has also been proposed as an explanation for the categoricity of obligations in the context of constitutivism, which seeks to draw substantial moral norms from the constitution of agency. For example, Christine Korsgaard argues that being an agent is a "necessary identity," and that it follows from this that we are subject to certain substantial demands by virtue of the inescapability of our roles as agents (Korsgaard 1996, 100-102). In short, the argument is that agency issues substantial norms or reasons, and those norms or reasons are categorical because the role of an agent is inescapable. It could be argued, then, that our inescapable roles as agents include epistemic agency, and that this is the source of the categoricity of epistemic obligations.

However, aside from the question of whether substantial norms can be derived from the thin notion of agency, inescapability as a source of categoricity has been refuted. Inescapability is the wrong kind of necessity on which to ground categorical obligations; what the constitutivist needs is normative necessity, not essential necessity. This is the crux of David Enoch's famous 'shmagency' objection (Enoch 2006, 187-188). The difference between normative and essential ne-

---

him. For this reason, role-oughts might seem categorical after all—unlike Kornblith claims—and just like moral and epistemic oughts are. However, this does not help Feldman, for he still needs to account for the plausible difference between teacher-oughts and epistemic oughts. In the next move in the dialectic, Kornblith supplies the inescapability argument to do just this, then goes on to reject it.

[3] Thanks to Teemu Toppinen for pointing this out.

cessity is well exemplified by Michael Smith (2015, 194, 198). He points out that besides being agents, we are also necessarily human, but the fact of that necessity clearly does not establish any analytic connection between our function as human beings and reasons for action. To function optimally as human beings, we have a reason to stay alive, be healthy, and produce offspring, but these reasons are not categorical; it is perfectly possible that there are no reasons to have children, to aim for health, and even to stay alive. It is conceivable, even if dismaying, that there are no reasons for human beings to exist, regardless of the (putative) fact that our whole biology is wired up for sustaining the existence of the human genome. So, even if we are necessarily human, no categorical obligations follow from this. Mere inescapability does not therefore suffice to account for the categorical nature of moral oughts, and therefore the categoricity of epistemic oughts is not explained by the inescapability of our roles as believers either.[4]

Kornblith (2001) has an alternative suggestion. He retains the idea that some oughts flow from evaluations of what counts as good performance, but proposes that instead of roles, ideals are the source of epistemic oughts. For ideals to be able to guide our actions, they must take human limitations into account—an unreachable ideal loses its action-guiding power. At the same time, ideals should not be so closely connected with capabilities of individuals that we lose sight of the fact that sometimes people are incapable of reaching the ideal. According to this view, then, the duty of not stealing is a moral ideal, which is not undermined by the existence of kleptomaniacs, and the duty to believe according to evidence is similarly an epistemic ideal which is not undermined by the fact that sometimes people engage in wishful thinking. Reasonable ideals lie somewhere in the large mid-

[4] Of course, one may dispute the categoricity of epistemic norms. While Kornblith takes it simply as a given, others deny that epistemic norms are categorical at all (Heathwood 2009, Cowie 2014, 2016). By contrast, Cuneo (2007) and Rowland (2013) argue that epistemic discourse is essentially committed to the categoricity of epistemic norms. If epistemic error theory is right, then of course much of the present discussion is fundamentally misguided, so the arguments and conclusions here should be read as tentative in that respect.

dle ground between the superhuman and the all-too-human. Kornblith argues that once we recognize this, we see that doxastic oughts do not require the level of voluntary control that the anti-deontology argument demands.

But ideals, too, fail to explain doxastic oughts. As stated at the beginning of this section, the main insight within the strategy is that, in epistemic context, 'ought' refers to standards of evaluation. This is not a sufficient solution to the problem, however, because standards of evaluation do not presuppose agency. It is common to distinguish between an ought that applies to actions and one that applies to states of affairs (Humberstone 1971, Harman 1977). Using ought language, we can either prescribe actions or evaluate states of affairs, where evaluation can target things such as cars and apples, without thereby attributing to them responsibility for being the way they are. Moreover, we evaluate things like weather and natural scenery without attributing to *anyone* responsibility for being the way they are. The problem with the simple NOIC view is thus that it equivocates between the evaluative and prescriptive readings of 'ought.'

Chrisman (2008) has a more sophisticated take on the NOIC strategy, one that avoids unobtrusively equivocating between evaluative and prescriptive oughts. His solution relies on Sellars's (1969) distinction between rules of action and rules of criticism, or ought-to-do's and ought-to-be's. Only the former kinds presuppose voluntary control, whereas the latter apply to states, how things ought to be. However, Chrisman, following Sellars, holds that these two kinds of oughts are importantly connected: rules of criticism materially imply rules of action. In other words, statements of the form:

> X ought to be in state S.

materially imply that:

> (Other things being equal and where possible) one should bring it about that X is in state S (Chrisman 2008, 360).

And, according to Chrisman (ibid., 364), doxastic oughts are of the form:

> X ought to be in doxastic attitude A towards proposition *p* under conditions C.

Even though the ought here implies agency, it does not imply that the *subject* of the ought is capable of voluntarily following the rule in question. Compare with:

> The beds ought to be made every morning

which materially implies that:

> (Other things equal and where possible) one should bring it about that the beds are made every morning,

but does not imply that the subject—the bed--is the one responsible for bringing it about. According to Chrisman, this solution manages to respect both doxastic responsibility and doxastic involuntarism because it allows believers to be open to criticism even if they do not exercise voluntary control in believing what they believe.

As stated above, Chrisman successfully avoids the problems of equivocation that Feldman and Kornblith are vulnerable to. Moreover, in trying to maintain the connection between rules of criticism and rules of agency, Chrisman explicitly attempts to retain the prescriptive reading of 'ought.' The problem with Chrisman's solution is, however, that it loses sight of the relevant agent. Consider:

(1) Anne ought to write the report.

If analyzed as:

(2) One ought to bring it about that Anne writes the report,

it becomes ambiguous regarding *who* ought to bring it about. In ordinary language the meaning is rather clear: it is Anne herself who ought to bring it about that the report is written. Of course, there remains a possibility that the statement refers to someone else, perhaps Anne's secretary, whose job is to make sure that Anne does everything she ought to do. But consider then an everyday-language statement:

(3) You ought to believe *p*.

When formulated as:

(4) One ought to bring it about that you believe *p*

the proposition makes very little sense. There is no one whose job it is to see to your beliefs but yourself. Other people have even less control over your beliefs than you do—evil demons aside. Chrisman's solution therefore fails to properly respect the prescriptive sense of 'ought.'[5]

Is there any way to avoid the ambiguity? Chrisman (2012) hints that the scope of the relevant agents could be restricted by context. This would get rid of the ambiguity in (2): unless Anne really has a secretary, it is reasonable to interpret the statement as referring to Anne herself. However, (4) remains as puzzling as ever. The only reasonable scope of agents here is whoever is referred to as 'you' in a given context. That would be equivalent to replacing 'one' with 'you' in (4), thereby getting:

(5) You ought to bring it about that you believe *p*.

This formulation, however, looks alarmingly like doxastic voluntarism, since it is very natural to read the phrase "to bring it about" as involving voluntary guidance. Remember that we wanted a formulation that only materially implies agency, without requiring that the subject of the proposition is the agent in question. But perhaps there is a way to read it otherwise. Here is a suggestion:

(6) You ought to be in a state such that you bring it about that you believe *p*.

But this formulation is not helpful either. It means that what we are doing is again merely evaluating the agent according to some standard, which is not the same thing as prescribing an action.

To see why evaluating agency is not equivalent to prescribing, it is helpful to consider an example. Imagine an agent, call him Derek, who has deuteranopia: his eyes do not perceive the color red. Because of this inability, his color judgments are systematically more or less off, depending on the amount of red in the color in question. For instance, he perceives purple and blue as equally blue. We can evaluate most of his color judgments individually as faulty: he just does not get them right. The fault also seems to derive from his agency

[5] Thanks to Joanna Klimczyk for helpful discussion on this point.

in the sense that the fault is not due to any outside factor, poor lighting conditions, colored lights, or some such thing. Instead, the faulty judgments stem from Derek's own doxastic system. Still, it would seem unfair to hold him accountable for the mistakes.

Now let us consider a contrasting case. Suppose that Derek works in a laboratory where his job is to make diagnoses. One disease, examplisis, is diagnosed by adding a few drops of an indicator liquid to the sample. The sample will then turn purple if it is positive, and blue if it is negative. Unfortunately, due to his deuteranopia, Derek is unable to tell the difference between blue and purple, rendering him unable to make the correct diagnosis. By contrast, his colleague Jerek has no physical incapability. He has just never been interested in colors and never bothered to learn to distinguish but the most basic of them. Think of the proverbial husband who, when his wife asks him to buy a can of fuchsia-colored paint, comes back with a can of magenta wondering what the difference between them is; are they not both sort of pink? Expand this example a bit and we get Jerek. Now, both Derek and Jerek are, evaluatively speaking, unreliable in their diagnoses of examplisis, and their unreliability is grounds for not asking their advice when the goal is to get a correct diagnosis. But in addition, Jerek seems responsible for his mistake in a sense that does not apply to Derek.

Derek's mistakes may be attributable to him, but there is no way to reason him out of his condition. There is a certain range of epistemic reasons that he is not capable of responding to, even with the help of the second-order reasons he might recognize when he realizes that he is unable to do his job well. His inability blocks the prescriptive sense of 'ought.' We cannot expect him to change when presented with reasons to do so. Jerek, on the other hand, has no such excuse. There is a sense in which we could reasonably expect him to improve his judgments because there seems to be a gap between his abilities and his actual performance. This gap is where the prescriptive ought finds its grip; it is where we place responsibility attributions. So, in order to distinguish between Derek and Jerek, we need more than just an evaluation of agency.

In conclusion, the No-Ought-Implies-Can strategy fails to solve the problem of doxastic ought because it fails to capture the prescriptive sense of 'ought.' The challenge is to find a distinctively epistemic agency which does not require voluntary control, but which does not collapse into mere evaluation either—a type of agency that resides somewhere between voluntary control and evaluation of states of affairs. I will turn to views that aim at precisely this in the next section.

3 The Process View

The NOIC strategy shares with doxastic voluntarism an implicit assumption that the 'can' in 'ought implies can' requires voluntary control (Shah 2002, Chuard & Southwood 2009) and tries to avert it by giving up the prescriptive ought. Resisting this assumption therefore opens up new theoretical space for spelling out conditions for doxastic responsibility.

A popular move is to model doxastic agency after the more familiar practical agency but identify a type of control which is distinct from voluntary control and which is exercised in cognitive activity such as reasoning or inquiry. According to these views, belief itself is a non-agentive, static state, but doxastic agency is located in the belief-forming processes or in the possibility of consciously reflecting and endorsing the belief after it is formed. These processes can be understood as a form of control themselves, and the notion of control as something distinctively epistemic, such as cognitive control (McHugh 2013), evaluative control (Hieronymi 2008, 2009), or rational control (O'Brien 2007).[6] For example, in McHugh's view, conscious control through inquiry is necessary for doxastic responsibility, but often only dispositionally so: the agency that we have over automatic beliefs is such that we *would* consciously endorse them when prompted, or we may reject them when considering the issue more carefully. An agent can therefore be held responsible for the act of reasoning or judging, or for being in a position to exercise such reasoning, whether or not it was actually exercised (McHugh 2011, 2013).

[6] Something along these lines has also been proposed by Peacocke (1998), Shah and Velleman (2005), Soteriou (2005), and Cassam (2010).

The process view seems at first a happy compromise that acknowledges both the role of reasoning and that beliefs themselves are not active. It would allow us to agree with the skeptics that beliefs are not subject to voluntary control, that beliefs themselves are not an exercise of agency, and that many of our beliefs are automatic, something we cannot help but have, in the way that sensory beliefs often seem to be. This way, the only kinds of norms that apply to beliefs themselves are standards of correctness. Yet we could insist that there is a genuine, distinct type of doxastic agency which is exercised in the act of belief-formation—in consciously deliberating, evaluating, reasoning, or judging whether *p* is the case—and that this is the domain of the prescriptive ought that Alston wrongly assumed to require voluntary control.

Unfortunately, the process view comes with implausible commitments. The insistence that beliefs are not themselves agentive is problematic because it means that the process view cannot explain what has come to be called *the transparency of belief*. In short, the problem starts from the observation that when asked whether I believe *p*, the most straightforward way to answer the question is to consider directly whether *p* is the case (Moran 2001, 2012). The puzzle here is to explain how a question concerning one's mental states is apparently transparent to a question concerning something else entirely. This poses a problem for the process view because in that view beliefs are just the outputs of one's deliberative processes, stored somewhere in one's memory like jars on a shelf. Therefore, it seems puzzling that in order to answer the question whether I believe *p*, the most straightforward thing to do is to engage in the process of asking whether *p* is the case, rather than just checking the shelves to see whether there happens to be a belief that *p* in there somewhere. By contrast, if someone asks you whether you have strawberry jam in your cupboard, the most straightforward thing to do is not to start preparing strawberry jam right away, but rather to go and check the shelves.

Moran argues that the transparency is made intelligible only if I can reasonably assume that whether I come to believe *p* is somehow determined by my considering the question of whether *p* is true:

> I *would* have a right to assume that my reflection on the reasons [for P] provided an answer to the question of what my belief (…) is, if I could assume that *what* my belief here is was something determined by the conclusion of my reflection on those reasons. (Moran 2003, 405)

Thus, in Moran's view, what explains the transparency is an agent's capacity for making up her mind—her doxastic self-determination. Doxastic agency must be such that it puts me in a position to know, on the basis of drawing the conclusion *p,* that I believe *p* (Boyle 2011, 8). Boyle argues, furthermore, that a related point must apply to the grounds on which my conclusion is held: "If I reason 'P, so Q', this must normally put me in a position, not merely to know *that* I believe Q, but to know something about *why* I believe Q, namely because I believe that P and that P shows that Q" (ibid.).

Initially it may seem that the process view can easily accommodate this datum, but Boyle argues that this is not the case. This is because, as both Moran and Boyle emphasize, transparency and self-knowledge hold for beliefs generally, but the cognitive control view must see them as holding only on those occasions when one consciously deliberates whether *p*, whereas the bulk of our beliefs would not come within the sphere of self-knowledge. Boyle considers a response by Shah and Velleman which makes precisely this distinction:

> If the question is *whether I already believe that P*, one can assay the relevant state of mind by posing the question *whether P* and seeing what one is spontaneously inclined to answer. In this procedure, the question *whether P* serves as a stimulus applied to oneself for the empirical purpose of eliciting a response. (…) By contrast, the question *whether I now believe that P* is potentially transparent to the question *whether P* in the capacity of just such an invitation [to reasoning]. (Shah and Velleman 2005, 506-507)

Shah and Velleman emphasize that the procedure of eliciting a spontaneous answer from oneself must be a brute stimulus and requires one to refrain from any reasoning because the reasoning might accidentally alter the state of mind one is trying to assess. They propose a strict distinction between an occurrent act of reasoning, in which doxastic self-governance is exercised, and the stored results of previous acts of such

reasoning, which are of course the products of doxastic self-governance but the recollection of which is not agentive.

Boyle (2011, 10) argues that the distinction is not tenable. First, if the only point of asking myself "Do I believe *p*?" is to elicit a spontaneous response, then it seems that it should be an open question for me as to whether I believe *p*, just as it would be an open question whether there is strawberry jam in the cupboard. But that seems to leave commitment to the truth of *p* out of the picture. If I recall what I believe about *p*, surely I must also recall what I think about the truth concerning *p*, that is, what I call to mind cannot be only the past assessment of the question, but also the present.

Second, Boyle argues that when an agent is questioned about her existing beliefs, we also expect her to be able to provide her grounds for those beliefs, whether or not she has consciously deliberated on the issue. Of course, an agent may fail to provide reasons, or her reasons may be inadequate, but the point is, no one questions the *applicability* of such questions even concerning one's automatic beliefs. Rather, the agent is held criticizable for holding beliefs on inadequate grounds, and this criticism is directed at what she *presently* believes, not only at how she formerly reasoned.

To sum up, the distinction between two different temporalities is implausible because we tend to interpret the products of an agent's past assessments as something she presently actively believes, something the truth of which she is committed to, and for which she sees (some) grounds for being committed to. The emphasis that Shah and Velleman place on the empirical observation of one's own responses is suspect because it seems that it would leave automatic perceptual beliefs outside doxastic responsibility. Since perceptual beliefs are not formed by deliberation, it is unclear how I could access the grounds on which the belief is held merely by self-observation. McHugh includes perceptual beliefs under doxastic agency via dispositional control. He argues that even though perceptual beliefs are in general automatic, we are able to exercise dispositional control over them by pausing and reconsidering when given some higher-order evidence (McHugh 2013, 134-135). But this is precisely the type of reasoning that Shah and Velleman place in the second cat-

egory, that of actively reasoning whether I now believe *p*. The original, automatic belief remains opaque.

The problem stems from the strict distinction between active cognitive processes and static doxastic states over which the processes govern. The connection between a belief and agency must be more intimate than the process view allows. While it is plausible that cognitive activities such as reasoning, inquiry, and deliberation are important for the responsibility one bears for one's beliefs, they cannot be mere external governing forces. I will argue for a view with such an intimate connection at its heart in the next section.

4 The Doxastic Perspective View

The process view resists P1* by rejecting volition and arguing that there is a distinctively doxastic type of control at work in doxastic responsibility. The third way to reject P1* is then to reject control as a necessary condition of responsibility altogether and embrace some form of attributionism about doxastic responsibility. In this section I will offer a brief sketch of one such account, the details of which must be left for future work.

Attributionist views of responsibility hold, roughly, that agents are responsible for those actions that can in some way be attributed to the agent: actions that somehow reflect who she is as a person, her identity as a moral agent, her moral character, or her evaluative judgments (e.g., Arpaly 2003, Sher 2009, Smith 2005, 2008, 2015). The views differ from each other in many respects, but what unites them is their denial of voluntary control as a necessary condition of moral responsibility. The arguments usually start from the observation that we often tend to hold each other responsible for things that are clearly not under voluntary control, things such as forgetting, omission, neglect, certain emotional reactions, and crucially, doxastic attitudes (Sher 2009, Smith 2005).[7]

[7] To clarify, it is not clear whether these authors mean moral responsibility for beliefs or a distinctively epistemic assessment. I hold that moral and epistemic assessments are distinct, but the basis of responsibility attributions is the same. That is, whether we can credit or criticize an agent for her belief depends on whether that belief is the upshot of her agency in the same way that whether we can praise of

In short, I suggest that beliefs are agentive because they are formed within the agent's doxastic perspective, which is attributable to the agent in a responsibility-implying sense. Unlike the process view holds, beliefs are intrinsically agentive and not merely static products of cognitive processes: they reveal the agent's doxastic perspective. I will unpack this below.

The first step is to recognize that beliefs are never just simple reflections of evidence. An agent's take on available evidence is in part a function of her prior pre-doxastic attitudes which manifest in the comparative weight she puts on various pieces of evidence. This is a familiar idea presented in different ways across various philosophical disciplines, from the theory-ladenness of observation to cognitive penetration of perception. Recently, Babic (2019) has argued that an agent's attitude towards epistemic risk in part determines how she ought, rationally speaking, to update her credences in light of new evidence. That is, how an agent ought to update her credences depends in part on such factors as how she evaluates the risks involved in different types of mistakes. Because such evaluations are not evidential, two agents can have the same evidence and still arrive at different beliefs without either of them being (subjectively) irrational. This is why beliefs are personal in the sense that they reflect the agent's doxastic perspective.

But what is doxastic perspective? The notion of perspective is borrowed from Elisabeth Camp (2017). This is how she describes it:

> On my view, a perspective is an open-ended disposition to notice, explain, and respond to situations in the world — an ability to "go on the same way" in assimilating and responding to whatever information and experiences one encounters. As such, perspectives differ from propositional attitudes -- in at least two related ways. First, a perspective determines no truth-conditions of its own […]. Second, having a perspective is a matter of cognitive action rather than cognitive content: it involves actually noticing, explaining, and responding to situations in a certain way, and not just representing situations as 'to be interpreted' in

---

blame an agent for her behavior depends on whether that behavior is the upshot of her agency.

> that way. In slogan form, perspectives are tools for thought, not thoughts in themselves. (Camp 2017, 78-79)

Camp's notion of perspective is wider than mine because she also wants to include non-cognitive attitudes under its rubric. As I am only interested in belief-forming methods, I will narrow it down to doxastic perspective:

> **Doxastic perspective**: A disposition to notice, explain, and respond to evidence in a manner which forms a systematic and (more or less) coherent view of the world. A doxastic perspective does not have truth-conditions; it is a tool for thought rather than a thought in itself.

A doxastic perspective is thus not merely about representing evidence as to be interpreted in a certain way, but *actually* noticing, explaining and responding to evidence in a way which results in an agent's personal view of the world. An agent's doxastic perspective is formed by the pre-doxastic, non-evidential attitudes that in part determine what she pays attention to, how she assigns weight to various pieces of evidence, and how she evaluates the risks involved in different kinds of mistakes.

There are various ways to argue for the attributability of an agent's doxastic perspective depending on the view of attributability conditions one endorses. On George Sher's view (2009), doxastic perspective would be part of an agent's psychophysical constitution, and as such grounds for responsibility. On a Frankfurt-inspired real-self view, doxastic perspective could perhaps be construed as the set of second-order evaluations that form an agent's doxastic real self (cf. Frankfurt 1987). On Angela Smith's (2005) rational relations view, doxastic perspective would be agentive because it consists of attitudes that are part of the agent's web of evaluative attitudes. I am going to suggest that doxastic perspective is attributable to an agent because it consists of an agent's goals, cares, and values, which are deeply personal. The agent's goals, cares, and values determine how the agent evaluates the risks involved in various possible mistakes, which in turn affect how she evaluates, for instance, the credibility of a source and the relative weight of various pieces of evidence,

and ultimately how she combines the available evidence with her pre-existing beliefs to form or revise a belief.

How does the doxastic perspective view handle the problems that affect the process view? First, because all beliefs are formed within an agent's doxastic perspective, there is no distinction between automatic perceptual beliefs and beliefs formed by deliberation. Doxastic perspective affects what one notices and how one responds to it, so even automatic perceptual beliefs come under its blanket. Similarly, doxastic perspective determines in part which pieces of evidence enter the deliberation and how they are weighted within the deliberation. Second, because there is no strict distinction between non-agentive beliefs and the agentive processes which govern them, transparency does not pose a problem. Beliefs are not static, but dynamic in that they reflect the agent's continued take on the relevant reasons. Thus, when asked whether I believe *p*, it makes no difference whether I rely on my memory of my previously formed belief, or whether I start assessing the truth of *p* at that moment: the result will in either case reflect my doxastic perspective. When asked for reasons for believing *p*, I can access the grounds on which I hold the belief because they are also part of my doxastic perspective.

5 Objection to and clarification of the view

One possible worry for this kind of view is whether it can really account for the prescriptive doxastic ought. It might seem that we are after all just evaluating an agent's beliefs against some standard. Where is the *agency* in all of this? If beliefs are simply determined by my doxastic perspective, how can I in any sense be responsible for them?

Answering this objection requires drawing a distinction between an agent whose belief is truly compelled—say, by paranoid delusions or deuteranopia—and one who is merely biased by her perspective. So, let us go back to Derek and Jerek. How does the doxastic perspective view explain the difference between them?

Derek's deuteranopia makes him unable to respond to a certain range of reasons, namely, the red pigment in purple. His inability is grounds for exemption from the sphere of responsibility practices concerning color judgments: the mis-

take does not reflect his pre-doxastic attitudes, his evaluations, or doxastic perspective. Rather, his agency is compromised with respect to color judgments. He is unable to respond to the relevant reasons, and thus we cannot expect him to change his view. The most we can do is to evaluate his beliefs from an objective point of view. This is a broadly Strawsonian take on things (Strawson 1962, 9). To evaluate an agent from an objective point of view is to refrain from treating her as a member of the epistemic community, that is, to refrain from treating her as a credible source, one whose judgments ought to be taken into account or who has authority concerning some issue. Needless to say, to exclude someone from the epistemic community would be a terrible injustice if done without a good reason (Fricker 2008), but an impairment in judgment would be a right kind of reason. A good reason is one that explains why the agent is unable to participate in the community (with respect to a specific question, or more generally), in other words, why the prescriptive ought is not applicable to her. An agent whose mistaken belief is due to a cognitive impairment can be evaluated as a poor doxastic agent without implying that she is responsible for the mistake because her mistake does not reflect her doxastic perspective, that is, her goals, cares, commitments, or values. Rather, the mistaken belief only reflects the cognitive impairment.

The doxastic perspective thus does the work in doxastic responsibility that quality of will does for Strawson's notion of moral responsibility: an agent is responsible for those beliefs that reflect the agent's doxastic perspective. Unlike Strawson, though, I hold that exemptions may concern only specific areas, such as Derek's color judgments, without having to compromise any other areas of doxastic agency (cf. Kauppinen 2018). Derek may still be a highly reliable source concerning, say, mathematics.

We can use Strawson's distinction between exemption and excuses to further spell out what constitutes an excuse in this picture. In Strawson's view, excuses do not invite us to modify our attitudes towards the agent, but only to view the mistake as one in respect of which reactive attitudes are inappropriate, that is, the mistake is not incompatible with the agent's attitude and intentions being just what they ought

to be (Strawson 1962, 7-8). In parallel, a doxastic excuse is such that the fact of the mistake is not incompatible with the agent's perspective and doxastic agency being just what they ought to be. Thus, the mistake does not reflect the agent's doxastic perspective, but, say, missing evidence or a temporary lapse of attention.

In sum, the difference between Derek and Jerek is that Derek's deuteranopia constitutes an exemption from doxastic responsibility regarding color judgments. Jerek's mistake, on the other hand, reflects his doxastic perspective, the fact that he does not find colors interesting and thus fails to notice important differences between them. It is fair to hold him responsible for his judgments: we can expect him to improve, perhaps with the help of the second-order reasons provided by his boss.

Conclusion

I have discussed three different ways to refute the skeptical anti-deontology argument against doxastic responsibility. The first is to deny that ought implies can in the doxastic sphere. The second is to deny voluntary control in favor of a distinctively epistemic form of control. I have argued that both of these approaches run into trouble. The third solution is to adopt an attributionist approach to responsibility and argue that no kind of control is necessary for doxastic responsibility. I sketched the outlines of a possible solution along these lines, one that ties doxastic responsibility to an agent's personal doxastic perspective, constituted by her pre-doxastic attitudes which determine the relative weight an agent gives to various pieces of evidence. I closed by arguing that we can then give a broadly Strawsonian analysis of responsibility practices which identifies excuses and exemptions.

Acknowledgements

The paper benefitted from comments by two anonymous referees as well as the audience members at the 2020 Philosophical Society of Finland Colloquium. I would also like to thank Jani Hakkarainen for insightful comments and helpful discussion on an earlier version of the paper, Teemu Toppinen for well-placed comments on the early drafts, as well as the

audiences at the 2018 Congress for Doctoral Students in Philosophy at the University of Tampere, the MLAG conference at the University of Porto, the Future of Normativity conference at the University of Kent, and the Philosophical Perspectives conference at the University of Eastern Piedmont, Novara. I am grateful to the Finnish Cultural Foundation for continuing support for my research.

University of Helsinki

References

Alston, W. P. (1988), "The deontological conception of epistemic justification," *Philosophical Perspectives* 2, pp. 257–299.

Arpaly, N. (2003), *Unprincipled virtue: An inquiry into moral agency,* Oxford University Press, Oxford.

Babic, B. (2019), "A Theory of Epistemic Risk," *Philosophy of Science* 86 (3), pp. 522-550.

Camp, E. (2017), "Perspectives in Imaginative Engagement with Fiction," *Philosophical Perspectives*, 31, pp. 73-102.

Cassam, Q. (2010), "Judging, Believing, and Thinking," *Philosophical Issues* 20, pp. 80-95.

Chrisman, M. (2012), "'Ought' and Control," *Australasian Journal of Philosophy* 90 (3), pp. 433-451.

Chrisman, M. (2016), *The Meaning of 'Ought': Beyond Descriptivism and Expressivism in Metaethics*, Oxford University Press, Oxford.

Chrisman, M. (2018), "Epistemic Normativity and Cognitive Agency," *Noûs* 52:3, pp. 508-529.

Chuard & Southwood (2009), "Epistemic Norms without Voluntary Control," *Noûs* issue 4, pp. 599-632.

Cowie, C. (2014), "Why Companions in Guilt Arguments Won't Work," *The Philosophical Quarterly* 64, pp. 407-422.

Cowie, C. (2016), "Good News for Moral Error Theorists: A Master Argument Against Companions in Guilt Strategies," *Australasian Journal of Philosophy* 94, pp. 115-130.

Cuneo, T. (2007), *The Normative Web.* Oxford University Press, Oxford.

Enoch, D. (2006), "Agency, Shmagency: Why Normativity Won't Come from What Is Constitutive of Action," *Philosophical Review*, Vol. 115, No. 2, pp. 169-198.

Fairweather, A. and Zagzebski, L. (eds.) (2001), *Virtue Epistemology: Essays on Epistemic Virtue and Responsibility*, Oxford University Press, Oxford.

Feldman, R. (2001), "Voluntary belief and epistemic evaluation," in M. Steup (2001), pp. 77-92.

Frankfurt, H. (1987), "Identification and Wholeheartedness," in *The Importance of What We Care About*, Cambridge University Press, Cambridge.

Fricker, M. (2007), *Epistemic Injustice: Power and the Ethics of Knowing,* Oxford University Press, Oxford.

Ginet, C. (2001), "Deciding to believe," in M. Steup (2001), 63-76.

Harman, G. (1977), *The Nature of Morality,* Princeton University Press, Princeton.

Heathwood, C. (2009), "Moral and Epistemic Open-Question Arguments," *Philosophical Books* 50, pp. 83-98.

Hieronymi, P. (2008), "Responsibility for Believing," *Synthese,* 161, 357-373.

Hieronymi, P. (2009), "Two Kinds of Agency," in L. O'Brien & M. Soteriou (2009), pp. 138-162.

Humberstone, L. (1971), "Two Sorts of Oughts," *Analysis* 32, 8-11.

Kauppinen, A. (2018), "Epistemic Norms and Epistemic Accountability," *Philosophers' Imprint* vol. 18, no. 8.

Kornblith, H. (2001), "Epistemic Obligation and the Possibility of Internalism," in A. Fairweather and L. Zagzebski (2001), pp. 231-48.

Korsgaard, C. (1996), *The Sources of Normativity*, Cambridge University Press, Cambridge.

McHugh, C. (2011), "Judging as a Non-Voluntary Action," *Philosophical Studies* 152, pp. 245-269.

McHugh, C. (2012), "Epistemic Deontology and Voluntariness," *Erkenntnis* 77, pp. 65-94.

McHugh, C. (2013), "Epistemic Responsibility and Doxastic Agency," *Philosophical Issues* 23, pp. 132-157.

Moran, R. (2001), *Authority and Estrangement*, Princeton University Press, Princeton.

Moran, R. (2003), "Responses to O'Brien and Shoemaker," *European Journal of Philosophy* 11, pp. 402-419.

Moran, R. (2012), "Self-Knowledge, Transparency, and the Forms of Activity," in D. Smithies and D. Stoljar (2012), pp. 211-238.

O'Brien, L. (2007), *Self-Knowing Agents*, Oxford University Press, Oxford.

O'Brien, L. & Soteriou, M. (eds.) (2009), *Mental Actions,* Oxford University Press, Oxford.

Peacocke, C. (1998), "Conscious Attitudes, Attention, and Self-Knowledge," in C. Wright, B. Smith, and C. MacDonald (1998).

Rowland, R. (2013), "Moral Error Theory and the Argument from Epistemic Reasons," *Journal of Ethics and Social Philosophy* 7, pp. 1-24.

Ryan, S. (2003), "Doxastic compatibilism and the ethics of belief," *Philosophical Studies* 114 (1-2), pp. 47-79.

Sellars, W. (1969), "Language as Thought and as Communication," *Philosophy and Phenomenological Research,* XXIX, pp. 506-527.

Shah, N. (2002), "Clearing Space for Doxastic Voluntarism," *The Monist* Vol 85, No. 3, pp. 436-445.

Shah, N. and Velleman, D. (2005), "Doxastic Deliberation," *Philosophical Review*, 114, 4, pp. 497-534.

Sher, G. (2009), *Who Knew? Responsibility Without Awareness,* Oxford University Press, Oxford.

Sinnott-Armstrong, W. (1984), "'Ought' Conversationally Implies 'Can,'" *Philosophical Review* XCIII, pp. 249-261.

Smith, A.M. (2005), "Responsibility for Attitudes: Activity and Passivity in Mental Life," *Ethics* 115 (2), pp. 236-271.

Smith, A.M. (2008), "Control, responsibility, and moral assessment," *Philosophical Studies* 138, pp. 367-392.

Smith, A.M. (2015), "Responsibility as Answerability," *Inquiry*, Vol. 58, No. 2, pp. 99-126.

Smith, M. (2015), "The Magic of Constitutivism," *American Philosophy Quarterly* Vol 52, No. 2, pp. 187-200.

Smithies D. and Stoljar, D. (eds.) (2012), *Introspection and Consciousness*, Oxford University Press, Oxford.

Soteriou, M. (2005), "Mental Action and the Epistemology of Mind," *Noûs*, 39: 1, pp. 83-105.

Steup, M. (2000), "Doxastic Voluntarism and Epistemic Deontology," *Acta Analytica,* XV, 24, pp. 25-56.

Steup, M. (ed.) (2001), *Knowledge, truth and duty: Essays on epistemic justification, responsibility and virtue,* Oxford University Press, Oxford.

Steup, M. (2008), "Doxastic freedom," *Synthese*, 161, pp. 375–392.

Steup, M. (2017), "Believing intentionally," *Synthese*, 194, pp. 2673-2694.

Strawson, P.F. (1962), "Freedom and Resentment," *Proceedings of the British Academy* 48, pp. 1-25.

Wright, C., Smith, B., and MacDonald, C. (eds.) (1998), *Knowing Our Own Minds,* Oxford University Press, Oxford.

Epistemology of Modality: Between the Rock and the Hard Place

ILKKA PÄTTINIEMI, RAMI KOSKINEN &
ILMARI HIRVONEN

We review some of the major accounts in the current epistemology of modality and identify some shared issues that plague all of them. In order to provide insight into the nature of modal statements in science, philosophy, and beyond, a satisfactory epistemology of modality would need to be suitably applicable to practical and theoretical contexts by limited beings. However, many epistemologies of modality seem to work only when we have access to the kind of knowledge that is at least currently beyond our reach. Or, in the extreme case, it is argued that even if we knew all the relevant information about the respective domain – or even the entire state of the world – there would still remain a special class of modal truths that would be left unaccounted for. Neither picture bodes well for practical applicability, nor for the philosophical justification of these epistemologies. This is especially the case as we hold that one of the main motivations for modal inquiry typically arises in cases of imperfect information and limited cognitive resources. We close by providing a partial remedy to the situation by suggesting an overall framework of relative modality (RM) that can be used to both unify some existing modal epistemologies and, at the same time, make them more metaphysically modest.

1. Introduction

In this paper, we review and criticize some popular approaches to the epistemology of modality. These include essentialism (e.g., Lowe 2012; Hale 2013), conceivability-based accounts (e.g., Yablo 1993; Chalmers 2002; 2010), and certain

philosophical uses of the framework of possible worlds (e.g., Lewis 1986; Nozick 1981, 128–137). Our treatment is by no means complete; there are also epistemological accounts of modality that fall outside the scope of our discussion, including, but not limited to, variations and combinations of the aforementioned positions. However, our purpose is to highlight what we see as a general trend amongst the standard philosophical answers to the epistemological challenge of modalities. We think that the approaches are far too often driven by background assumptions that lack adequate epistemic justification. As a result, instead of giving us tools to tackle puzzling cases of modalities in science, philosophy or ordinary life, these theories rather lead to further philosophical problems. In a nutshell, they tend to either explain our modal access by positing explanantia that are themselves epistemically highly problematic or, failing to or not attempting to explain our modal access, they lead to forms of modal skepticism.

We think this situation is problematic for many reasons, chief among them the fact that modalities are an integral part of our scientific and everyday reasoning. Ideally, we would like to see an epistemological theory of modalities that is at once without obvious philosophical problems and can do justice to our actual epistemic practices. Indeed, the existing accounts of modality seem to be in stark tension with the pragmatic rationale behind modal reasoning. Moreover, since modal language is often invoked in the context of limited knowledge (e.g., Dray 1957, 165; Wimsatt 2007, 130–131), it would be good if our epistemological theory could also say something about these situations. That is, something other than that they are *all* unjustified. Surely some of these modal claims are still epistemically more (or less) warranted than others?

After reviewing the standard answers in the epistemology of modality, we close with a short account of our own that should provide a partial remedy to the situation. More precisely, we sketch an overall framework of relative modality (RM) that can be used to unify some existing modal epistemologies and, at the same time, make them more metaphysi-

cally modest.[1] RM is concerned about what is kept fixed in publicly evaluable systems, scenarios, models, theories, and other vehicles of inference-making (for a precursor of this kind of view, see Quine 1982). The epistemology of modal statements thus becomes an internal question of the features and boundary conditions imposed by the system in question. These system features are then typically justified externally through experiments, manipulations, and so on. In certain contexts, they may even be simply stipulated. Our picture complements a parallel line of argumentation developed recently by Fischer (see Fischer 2016; 2017). However, contrary to Fischer's Theory-Based Epistemology of Modality (TEM), our view does not impose strong veridicality conditions or elements of accompanying mental models in modal justification. RM is also very flexible because its basic principles can be applied without much modification to science, philosophy and ordinary cases of modal reasoning.

The structure of the paper is as follows. In the next section, we introduce the problem of epistemic modal access. The following three sections then review and criticize the standard answers to this challenge. Section 3 focuses on essentialism, while Section 4 investigates conceivability-based accounts. In Section 5, we discuss some basic features of the framework of possible worlds and argue that they are often misapplied to give a false sense of epistemic justification for modal claims. Building on these criticisms, we then discuss the overall situation and provide our partial answer by introducing the relative modality (RM) framework in Section 6. Finally, Section 7 concludes the paper.

2. Epistemologies of Modality: Gaining Access

Central among the questions about modal statements are the following: (i) are there modal facts or truths?, and (ii) if there are, how do we come to know, or gain access, to them? The second of these questions will be our primary concern here – the epistemology of modality. But we will also say something about (i), since if there are no modal facts, the epistemology of modality will be rather useless.

---

[1] For a more comprehensive account of the proposed epistemology of modality, see Hirvonen, Koskinen and Pättiniemi (forthcoming).

In his book *Modality* (2003), Joseph Melia gives the following argument for the indispensability of modal facts. Let us assume that we have a theory of the world that contains only and all facts about the state of every object, past, present, and future (Melia 2003, 1). It might state, say, that the chair at the southwest corner of an attic weighs exactly 5.6735 kg, and that ten years hence it will be in the basement instead of the attic. Both these statements will be true. But does this theory contain all truths? Melia does not think so. It would encompass truths about epistemic modalities, but other kinds will be left out; namely, truths "that go beyond the merely actual and tell us something about how things might be, or must be, or would be had things been other than they actually are" (Melia 2003, 3). Indeed, the theory will not contain any informative facts about possibilities, necessities, counterfactuals, and so on. Here we are led to quite a strong version of modal skepticism. Nevertheless, it is worth asking: would there even be any need for modalities under Melia's scenario? We are not convinced that there would be. To see this, let us take a look at the indispensability of modal statements.

So, do we really need modal statements for anything? Indeed, we do. They allow for prediction and control and assignment of causes and culpability. What, after all, is prediction if not the determination of possible (or necessary, if we are lucky!) future states of a system? And control is just prediction combined with an intervention aiming at serving our goals. As an illustration, we will briefly consider a timely example from the science of climate models.

A climate model is basically a set of equations that characterize the dynamical and thermodynamical processes in the atmosphere and the oceans, with a set of initial conditions and parameters that characterize the state of the atmosphere and the oceans, and of differing 'drivers' of climate change, such as forcing caused by the increase of carbon dioxide in the atmosphere (Neelin 2011, chs. 3 & 7).[2] Such models are built to facilitate a better understanding of the Earth's climate

[2] Also, many processes, such as cloud formation, will be added as parameters due to their complexity. The whole nature of climate models need not concern us here, but for those wishing to learn more, Houghton (2005), in addition to Neelin's book (2011), is a good starting point.

and of climate change. Climate models allow us to determine the causes of, say, past and present warming events and to compare the differences in the drivers of such events. This is accomplished through a counterfactual analysis; what would have had to have been different to cause a different outcome?[3] The models can also be used to predict, or make *projections*, of future climate, given differing interventions on factors such as greenhouse gas production. Indeed, the most interesting output from climate models is not what will happen if things stay as they are now, but how things *can* be if we change the current situation. That is, (1) what are the possible future states of the climate, and (2) how can we bring these about? (Meehl et al. 2007; Neelin 2011, ch. 7.) In other words, climate models tell us not just how things are, but *why* they are as they are (the causes of climate change) and, further, they allow us to predict and, hopefully, to control the climate. Therefore, climate models are *modal* in an interesting and indispensable way.

The example of climate models illustrates a more general pattern across science and in more ordinary matters: modal statements are indispensable. Their indispensability comes from the fact that we do not have Melia's grand theory; that is, we are not all-knowing. An all-knowing being would have no need to know whether something will happen out of necessity or only contingently: it simply will happen. The same holds for counterfactuals, causes, culpability, and so on. The theory will tell us what has happened, what is happening, and what will happen, even our (futile) attempts to change events. A world with such a theory will be a necessitarian one.[4] Because of this, Melia is wrong in thinking that such a theory would leave something out: it would not. However,

---

[3] This is not to say that a counterfactual analysis of the *metaphysics* of causation is necessarily correct, but rather that we need it to pick out causes in our systems of interest.

[4] Is this saying too much since one cannot reason from actuality to necessity? The problem that omniscience brings is that if one *knows* the future state of a system, then that state will occur; otherwise, one would *not* have known it. Whether one chooses to call this "necessary" will be a matter of taste. Formally it seems to bear all the hallmarks of necessity. A world with Melia's grand theory will be practically indistinguishable from a necessitarian one.

we have good reasons to think that such a theory is not to be had. For us limited beings, modalities are not a thing to be excised from a mature science, but rather the very *point* of science. They are that which allows for explanation, understanding, control, and prediction. Scientific theories are modal to their *core*. Now we get to our main question: given that modal statements are needed, do we then know any modal facts? And even more importantly, how do we gain access to them?

Current epistemologies of modality are often built up from metaphysical theories concerning modalities. They try to get from what modalities *are*, in some metaphysical sense, to how we come to have knowledge of them. We take this to be quite wrong-headed, especially given that we do not have an agreed-upon epistemology of metaphysics. Moreover, if we cannot know the correct metaphysics, we can hardly use it to find out about modalities. So, we take that an epistemology of modality has to start *epistemology first*.

There have, fortunately, been approaches that respect an epistemology first approach. Examples include Bob Fischer's (2016; 2017) Theory-Based Epistemology of Modality (TEM), and Sonia Roca-Royes' (2017) approach that reasons from actuality and similarity to possibility, at least in the case of *de re* possibilities. According to these approaches, one way of gaining (ampliative) modal knowledge is through what actually is the case, combined with manipulation and reasoning from similarity. A second way is based on what we call relative modality (RM): for any system, modal claims are evaluated *relative* to said system. The simplest case will be using classical logic. Simply put, if a claim leads to a contradiction, it will be impossible (relative to the system); if a claim does not lead to a contradiction, it will be possible; if the negation of a claim leads to a contradiction, the claim will be necessary.

If the kind of epistemology characterized above is viable, it goes a long way to show that in the context of science and everyday matters, a metaphysically based epistemology of modality is unnecessary. Further, it seems that many such metaphysical theories can be taken to be instances of relative modality, where the systems in question are not always well

justified. To show this, we will take a look at some contenders for an epistemology of modality.

3. Essentialism and Counterfactuals

The most well-known essentialist accounts of modal knowledge come from E.J. Lowe (2012) and Bob Hale (2013). According to Lowe, our knowledge of (metaphysical) modality is based on our ability to grasp the essences of entities. These essences can be expressed through real definitions, and essence is simply what the entity in question is. "Grasping" the essence of something is to *understand* what that thing's real definition is. (Lowe 2012.)

Hale's story of modal knowledge is quite similar to Lowe's. He also starts from the essences of entities and their real definitions (Hale 2013, 133n, 254). Some real definitions can be known a priori. Such cases include analytic truths, like "a cob is a male swan", and our explicit grasping of some relevant concept like "a natural number" or "a square". (id., 255–256.) This a priori way of knowing essences is familiar to us already from Lowe's view. Some essences, however, are not accessible to us a priori through mere conceptual reflection. In these situations, essences are known via empirical investigation together with general essentialist principles, such as "any object is essentially an object of a certain general kind" (id., 259–260, 270). Given our empirical knowledge and knowledge of the general principles, we can obtain knowledge of facts concerning essences covered by the general principles (id., 269). However, in a posteriori cases our knowledge of essences might remain incomplete: perhaps we have not yet been able to figure out all essential facts of an entity but only a subset of them (Vaidya 2018, 235).

The problem here is that the essentialist move merely changes the epistemology of modality to the epistemology of essences. This way of passing the buck does not appear to present a satisfactory answer to our conundrum, for there seems to be less agreement about what properties are essential compared to what sort of claims are necessary. Lowe (2012, 940) even explicitly admits that "philosophers can have honest disagreements about questions of essence." Moreover,

he also states that sometimes we do not fully adequately grasp the essences of things that we are thinking (ibid.).

However, it seems that we can know necessary modal truths without knowing their essences. Consider, for instance, the ellipse. According to Lowe, even though an ellipse can be defined as a type of conic section, such a definition would not capture its essential features. Among the reasons that Lowe offers for the conclusion is that cones cannot be essential for ellipses since ellipses can exist without cones. (Lowe 2012.) Irrespective of whether Lowe is right about this, there seems to be something wrong with his reasoning. After all, we can infer all of an ellipse's properties from the cone-section definition, even those that Lowe considers essential. Thus, if someone does not know the *real* definition of an ellipse, she can nevertheless deduce the same necessary truths from this non-essential definition as from the real one. To take stock, knowledge of essences is not required for inferring modal knowledge, and there are "honest disagreements" about which properties are essential. Thus, we can know necessary truths even if we do not know the essences of things.

Hale's situation is similar to Lowe's. Besides the fact that we might not need the real definitions of entities to have modal knowledge, Hale's account also requires knowledge of general essentialist principles for a posteriori knowledge of essences. It appears to be relatively safe to assume that at this point, there is no agreement about what those principles should be, since there is no agreement among philosophers whether essences exist in the first place. And still, we do seem to agree about modal claims and have modal knowledge.

The situation is similar in the case of Williamson's counterfactual account of modal knowledge. Williamson's conception of the epistemology of modality is founded on our ability to evaluate counterfactual conditionals in our imagination while keeping some "constitutive facts" fixed (Williamson 2007, 164, 170). Even though Williamson does not discuss essences but "constitutive facts", in practice, the constitutive facts play the same role as Lowe's or Hale's essences. In addition, Williamson does not give a detailed account of how we get to know which facts are constitutive (Roca-Royes 2011; Fischer 2016). Still, he does say something about which things should be kept fixed when we are talking about nomic mo-

dalities: what is necessary, possible, and so on, according to the laws of nature under specific circumstances (Williamson 2016).

In Williamson's view, nomic modality requires that the laws of nature – which are discovered abductively – are kept fixed along with "all true claims of identity and distinctness" and "true claims of kind membership and non-membership" (Williamson 2016, 463). But this, in his mind, would already force us to the domain of metaphysical modality. Claims like "Hesperus is Phosphorus" and "Hesperus is not a quark" are not something that natural laws can tell us (ibid.). Hence, Williamson claims, metaphysical modalities are needed to make nomic modalities consistent to avoid blatant inconsistencies like "Hesperus is *not* Phosphorus" or, by the same token, "Hesperus is *not* Hesperus" (ibid.).

The problem with Williamson's approach is that for nomic modality, the relevant claims of identity and kinds are either already fixed through similar scientific research as the laws of nature or it is not clear how the additional claims should be fixed. This presents us with two options. On the one hand, either nomic or natural modality does not require additional metaphysical information besides the ontological commitments that scientists have already made. On the other hand, we need a separate epistemology for the metaphysical claims, and there does not appear to be agreement about what that epistemology should be like. However, it seems evident that the first option is right: we have adequate ways of evaluating natural modalities based on scientific research. Indeed, Williamson's troubles look very similar to those that Lowe and Hale have to face.

4. Conceivability as the Modalist's Guide

Deriving metaphysical possibility from conceivability has an illustrious history. Among the famed defenders of this line of thinking is no lesser a figure than David Hume:

> 'Tis an establish'd maxim in metaphysics, that whatever the mind clearly conceives includes the idea of possible existence, or in other words, that nothing we imagine is absolutely impossible. We can form the idea of a golden mountain, and from thence conclude that such a mountain may actually exist. We

> can form no idea of a mountain without a valley, and therefore regard it as impossible. (*Treatise*, I, ii, 2)

However, we will concentrate on newer proponents of the "conceivability entails possibility" principle, namely Stephen Yablo (1993) and David Chalmers (2002; 2010). Still, the remarks made here will also apply to more classic defenders of the principle such as Hume and, arguably, Descartes.

According to Yablo (1993, 29), p is conceivable for a subject S if S can imagine a world that S takes to verify p. And, respectively, p is inconceivable to S if S cannot imagine any world that S does not take to falsify p. Chalmers' conception of conceivability shares much with Yablo's account, but he makes additional requirements on the capabilities of the subject S. Or more specifically on the *type* of conceivability, but it turns out that this, in turn, requires much from S, more indeed than can be expected from limited cognitive beings.

Like Yablo, Chalmers divides conceivability into several different types, two of which pretty much coincide with Yablo's conceptions and thus are amenable to the same treatment. Unfortunately, the rest are rather technical, and their full explication would take up more space than the present work allows for. What we can say, however, is that the remaining types of conceivability call for "ideal rational reflection" (Chalmers 2010, 143) and thus for ideal rational reflectors; these, in turn, seem to be in a rather short supply.

In Chalmers' parlance, positive conceivability means that a subject can imagine a situation where p would hold. On the other hand, negative conceivability means that a subject does not find a contradiction in a situation where p would hold. (Chalmers 2010, 144.) Chalmers also makes a distinction between prima facie and ideal conceivability. Roughly, prima facie conceivability is something that limited beings can conceive, whereas ideal conceivability requires ideal rational reflection. (Chalmers 2002, 147; 2010, 143.) As an example, squaring the circle was, at least, negatively prima facie conceivable because those who tried to achieve it did not see a contradiction in the endeavor. But it is not ideally conceivable because squaring the circle is impossible with a finite number of operations.

Last but not least, Chalmers separates primary conceivability from secondary conceivability (Chalmers 2002, 157; 2010,

146). This distinction is based on his version of two-dimensional semantics. Primary conceivability is connected to a proposition's primary intension and the secondary conceivability to its secondary intension. This is best illustrated with an example. Take the question: "Could it have turned out that water is not H_2O?" If one considers the primary intension of the question, then the answer is yes. One can imagine a scenario where it would have turned out that the "watery stuff" in the actual world was something other than H_2O, say, XYZ. However, from the perspective of the secondary intension this is impossible because the term 'water' refers to H_2O in all counterfactual situations, given that water is necessarily H_2O. Since we cannot know a priori that water is H_2O, it is in some sense – the primary sense – conceivable that water is not H_2O. Still, in another sense, due to Kripkean a posteriori necessities, it is inconceivable that water would not be H_2O. After all, if water is necessarily H_2O, then water is H_2O in all possible worlds. (Chalmers 1996, 57–59; 2002, 157; 2010, 146; Vaidya 2015; Feng 2017, 21–23.)

However, here the question arises of why we should use either Yablo's or Chalmers' approach. Presumably, one would not use either method to find out about, say, physical or mathematical possibilities. Let us return to squaring the circle as a simple example to illustrate this.

For centuries mathematicians tried to find a method for squaring the circle, that is, transforming a circle into a square of an equal area through finite steps using only a compass and a ruler. Clearly, these mathematicians did not *consider* their task impossible or *inconceivable*, for if they had, they doubtless would have discontinued their efforts.[5] But, as it turns out, squaring the circle *is* impossible. (Schubert 1891.) This seems to imply that all of those mathematicians who tried to accomplish it, and thought they had conceived of it, were mistaken. Hence, one can err in taking something to be

[5] Descartes famously distinguished conceivability from imaginability when he pointed out that *imagining* the difference between a thousand-sided and a thousand-and-one-sided polygon would be quite difficult if not impossible. Still, as Descartes points out, it clearly is possible to make a conceptual distinction between the two, and thus, their difference is nevertheless *conceivable*. (Descartes 1984, 50–51.)

conceivable. Note that there are external, intersubjectively evaluable criteria for determining whether a circle can actually be squared. Now, we are left with two options: (1) Claim that the mathematicians who tried to square the circle had not, in fact, conceived of squaring the circle. They merely thought they had. So, then, the problem will be knowing when one has indeed conceived of something. If external, intersubjective criteria are lacking, this task seems impossible to undertake; there will be no intersubjective way of justifying whether one has indeed conceived of something or merely thinks that one has. (2) Claim that the mathematicians had conceived of squaring the circle, but the task just happens to be impossible. Then the link from conceivability to possibility will be severed. Therefore, the conceivability-to-possibility principle is either incorrect or limited in its scope because it requires less limited beings than mere humans.[6] If the principle is not reliable in mathematics, why would we take it to be reliable in a field where justification is even harder to come by, namely metaphysics?

Furthermore, Peter van Inwagen (1998) has argued that if conceivability is a guide to possibility, then we need to conceive all the required steps for really conceiving the thing. His examples are transparent iron and purple cows. If someone indeed claims that these things are (metaphysically) possible because they are conceivable, then they should actually con-

[6] In an unpublished manuscript, "The Unsoundness of Arguments From Conceivability", Andrew Bailey has presented this very same argument, namely, that as cognitively limited creatures, we are unable to determine whether something is ideally conceivable or not. Chalmers has responded to him by citing instances of clearly prima facie conceivable or inconceivable things that are also ideally conceivable or inconceivable: "Although we are non-ideal, we can know that it is not ideally conceivable that 0=1 and that it is ideally conceivable that someone exists. We know that certain things about the world (say, that all philosophers are philosophers) are knowable a priori and that certain things about the world (say, that there is a table in this room) are not so knowable even by an ideal reasoner." (Chalmers 2010, 155.) However, even if we can know that 0=1 is not ideally conceivable, that does not yet, in itself, give us good reason to think that some metaphysical ideas (such as philosophical zombies) are ideally conceivable. Perhaps such ideas are more alike with squaring the circle: they seem conceivable even if they are not.

ceive the things in question on the physical and chemical levels. That is, what things in the DNA of the cow make its color possible, or what in the structure of the iron could make it transparent. Similarly, what steps are required in squaring the circle. If one really considers it conceivable, one should conceive *all* the appropriate steps needed for the squaring. But this would entail actually squaring the circle or giving a mathematical proof of its possibility. What role would be left for conceivability?

5. Possible Worlds

The last philosophical approach to the epistemology of modality that we examine concerns the logico-semantic framework of possible worlds. This is not so much a specific epistemic theory, but more of an amalgamation of approaches and strategies that refer to a common formalism. The most classical account of modality in terms of logic is through the idea of non-contradictoriness: possible propositions consist simply of all those things that can be asserted without contradiction. Necessities, in turn, are such that their denial would lead to a contradiction, and so on. However, this classical logical treatment of modality is ambiguous because it, in a crucial way, depends on the domain of investigation and how it is being represented. In order to apply classical logic to any material modalities, choices have to be made as to how to interpret and formalize them, what to include in the domain of the logical calculus, and so on.

In contemporary philosophy, modalities are typically investigated in specially devised modal logics of which there are many axiomatizations. The reigning semantics for these formal systems is provided by the framework of possible worlds (e.g., Kripke 1959; see also Hintikka 1957). Assessment of possibility and necessity is made based on a set of worlds (typically sets of propositions) and accessibility relations between the worlds. So, for example, if Tuomas happens to find himself in a situation where it is raining, the proposition that it is possible for Tuomas to be in a situation where it is not raining is dependent on a few things. Let us say that in our scenario, Tuomas is in a world w. Then, for the alternative possibility to hold, there needs to be another world, call it w',

in which (i) it is not raining and (ii) it is accessible from w. Furthermore, we would also like to know that the identity of Tuomas stays the same across these two worlds.

Possible world semantics provides a powerful tool to tackle modal scenarios of various kinds in philosophy and elsewhere. Some philosophers also use possible worlds as a metaphysical theory, the *locus classicus* being Lewis' theory of modal realism (Lewis 1986). However, what is noteworthy is that all the aforementioned basic facts that are required for the complete assessment of modal statements need to be stipulated on a case-by-case basis. Thus, even though it provides a richer representational framework for various purposes, possible worlds semantics does not really go any further than classical predicate logic to explain or ground modalities. All the epistemologically crucial steps happen when the particular stipulations are being made.

What does this mean in practice? Let us look at an example. Typically, possible worlds are evoked to explain why one alternative state of affairs is philosophically more plausible than others. This is manifested in the way philosophers speak about "close" or "nearby" possible worlds. Elaborate arguments are invoked in the context of the analysis of knowledge, for example, where various modal conditions are applied to determine what kind of changes to our actual circumstances we should regard as epistemically relevant (e.g., Nozick 1981, 172–178; Pritchard 2005, ch. 6). Using the framework of possible worlds, philosophers can thus sometimes rule out certain scenarios as far-fetched or irrelevant in the context of their argument. The basic idea here is often quite intuitive. For example, the scenario in which unicorns exist is closer to the actual world than the scenario where both unicorns *and* centaurs exist. This seems to be valid logical reasoning based on the properties of the conjunction connective. But what if we simply compared worlds in which unicorns exist and worlds in which centaurs exist. Which of these possible worlds is closer to the actual world? What is the metric used here, and how could it be justified?

The problem is, unfortunately, that it is precisely the questions of the metric that is often not explicated in philosophical arguments that refer to the closeness of possible worlds (e.g., Nozick 1981, 172–178; Pritchard 2005, ch. 6). Notice that we

are not implying that a sensible metric is not to be had in these kinds of situations, but instead that there are likely to be multiple (formally definable) metrics that could be used in the context of possible worlds semantics. Here, it is the very *choice* of the particular metric that is doing the heavy lifting, not the semantic framework of possible worlds itself. However, as in the case of sciences more generally, no model or representation can justify its utility in isolation from its purpose and application. Thus, possible worlds seem to face similar challenges as the two previous routes to modal knowledge.

6. Discussion

We have argued that all of the above theories face epistemological challenges individually. However, we have not yet considered whether they (and further variations based on them) can also conflict with each other. This is clearly a problem since they aim to describe the correct set of modal facts and our epistemic access to them. Interestingly, however, their possible agreement could also be seen as a problem. For then, the question arises concerning what makes any particular theory of modality special. If a conceivability-based theory of modality gives all the same answers as a counterfactual one, which of these is doing the grounding of our epistemic access? It is also considerations like this that urge us to move more towards the justification of these systems as a whole.

It behooves us now to give a more detailed account of relative modality (RM). Recall from section 2 that according to RM, modal claims are evaluated *relative* to a system. At its simplest, this will be done through classical logic, where statements are possible if they do not lead to a contradiction with the system, necessary if their negation leads to a contradiction, and so on. So, what RM allows for is good reasoning about modal claims *relative* to a given system. What it does *not* give are criteria for the *choice* of a system.

At first blush, relative modality would appear to offer a friendly ground for metaphysical modality. After all, we can evaluate the modal claims of any metaphysical system using RM. But here the modal knowledge gained is only knowledge about a system. If the goal of metaphysics is to say something

about the world, then such knowledge is otiose unless one can show that the system is a good match for reality. Of course, one can construct a system of rules as one sees fit. It is possible to build a system (or theory, model etc.) that does not correspond to reality. So a system can be based on, say, what kinds of rules individuals find amusing, as is the case with games like chess, or on the intuitions of individual thinkers.

Problems arise when one claims that their system describes how things are in the real world. How are we to evaluate whether such a system is any good as a description? One can do this in science by checking whether the predictions of the system match our empirical findings. But this only tells us about natural or empirical modalities. Insofar as empirical testing of certain claims is not possible, how can one check whether the system in question tells us anything about the real world? Hence, one faces the challenge of how to make modal claims non-arbitrary. We claim that one is not justified in accepting the claims until this challenge is met.

There is another problem if metaphysical modality is understood through relative modality. Which metaphysical claims should one fix? According to our relative modality account, metaphysical modalities are founded upon fixing certain claims – claims like Hesperus is Phosphorus, gold is the element with the atomic number 79, cats are animals, Elizabeth II is the daughter of George VI, water is H_2O, and so on. These claims, as themselves, are not yet modal claims. Nevertheless, their fixation as a part of a system is what makes them necessary. But why should these specific statements be fixed as axioms of our ontological system? Why can they not merely correspond to, say, a particular state of play in chess, a certain arrangement of pieces on the board? Why do they instead have to be analogical to the rules of chess?

Kripke, for instance, has argued that the special metaphysical status of these statements comes from the fact that their parts refer to the same entities or substances in all possible worlds. In other words, they are rigid designators. (Kripke 1980.) Indeed, they refer to the same target in the actual world, but to say that they refer to the same target in all possible worlds is to merely state – from the point of view of relative modality – that this is something we should keep fixed.

It does not tell us *why* we should do so. Now, Kripke offers several thought experiments to prove his point. This is all fine, but do not these thought experiments only tell us that we intuitively keep certain things fixed or that we keep them fixed for other than metaphysical reasons, like physical or historical reasons? Thus, the justification for their special metaphysical status remains still unclear.

Now, we are in a position to see that both Lowe's and Hale's essentialist accounts are based on RM. The basic idea behind them is that entities have essentialist properties that are kept fixed. Furthermore, from our knowledge of essences, we can deduce necessary truths. After all, as Lowe states, "any *essential* truth is *ipso facto* a *metaphysically necessary* truth" (Lowe 2012, 938 italics in original). So, here, (metaphysically) necessary truths are derived from the essential properties of our target of inquiry. From these necessary truths, further modal truths can be deduced. If *X* does not contradict any necessary statement, then *X* is possible. If *X* does contradict such statements, then it is impossible, and so on. But this is precisely the way RM deals with modal inferences, only here the system is fixed to be the essential features of the target, or domain, of inquiry. But, as stressed above, why choose either of these systems? The interesting epistemological question is not "what is possible/necessary given a system?" but rather "how to choose a system in which to evaluate modal claims?". Lowe and Hale do not answer this latter question. Thus, given that the machinery through which modal inferences are made is RM and that it is not clear which, if any, claims concerning essences are justified, this essentialist route to modal knowledge is questionable at best.

Similar reasoning holds for both Yablo's and Chalmers' use of the conceivability-to-possibility principle. That is, they are both based on relative modality. Again, possibility is relativized to an individual's ability either to imagine scenarios or infer contradictions. In either case, there is a system, although not one explicitly spelled out, in relation to which a proposition is considered to be possible. And again, Yablo and Chalmers seem to have very little in the way of justification for their preferred system. This, again, leaves Yablo's and Chalmers' approaches questionable.

The above was not an argument for the falsity of any or all metaphysical modal claims. So, we are not saying that metaphysicians advancing metaphysical claims are mistaken. We have merely argued that, at least thus far, they do not have a good justification for such claims.

As a final note, are we not guilty of moving the interesting epistemic questions from the modal claims onto the choice of a system? Indeed we are, but this does not have to be a bad thing. For, insofar as we can justify our choice of a system, we will at the same time gain a way of justifying modal claims. So, when is a choice of a system justified? In the case of, say, scientific models and theories, they are justified by empirical corroboration, consistency with other theories, and so on. In other cases, like the rules of chess, such external justification is not needed. But subjecting our justification of a system to ampliative reasoning will make claims based on RM *epistemic*, at least in some sense. Here lies a risk that we will end up having provided an epistemology only for *epistemic* – and thus subjective – modalities. First, imagining, conceiving and appeals to intuition are also subjective in nature. So, these ways of justifying a system or a claim will be subjective. Second, justification of scientific, mathematical and logical theories is done in an *intersubjective* way. Reasons, results, inferences, and so on, have to be presented in a way that is accessible to others for the scientific community at large to be able to evaluate them.

7. Conclusions

We have criticized some popular approaches to the problem of epistemic access to modal knowledge. These included essentialism, conceivability-based accounts, counterfactual reasoning, and the use of possible worlds as an epistemic grounding of modal claims. We argued that all of these epistemologies seem to work only when we have access to the kind of knowledge that considerably surpasses what can be expected from our scientific, yet piecemeal and cognitively limited, accounts of the world. They then solve this situation through strong metaphysical assumptions or succumb to modal skepticism. Thus, instead of guiding our modal access,

they lead us astray or function as overly officious gatekeepers. This, to us, is untenable.

We argued that some of these worries could be eased if we adopt the framework of relative modality (RM). RM is concerned with what is kept fixed in publicly-evaluable systems of modal inquiry. The epistemology of modal statements thus becomes an internal question of the features and boundary conditions imposed by the system in question. These system features are then typically justified externally through experiments, manipulations, theoretical derivations, or they may even be stipulated. The primary motivation behind this move is not to rule out any particular theory of modalities but rather to make the epistemology of modality methodologically honest.

University of Helsinki
University of Vienna
The Helsinki Circle

References

Bailey, A. (manuscript), "The Unsoundness of Arguments from Conceivability", unpublished manuscript, URL = <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.591.8106&rep=rep1&type=pdf> 17.6.2021.

Chalmers, D. (1996), *The Conscious Mind: In Search of a Fundamental Theory*, Oxford University Press, Oxford.

Chalmers, D. (2002), "Does Conceivability Entail Possibility?" in T. Szabó Gendler and J. Hawthorne (2002), pp. 145–200.

Chalmers, D. (2010), *The Character of Consciousness*, Oxford University Press, Oxford.

Cottingham, J., R. Stoothoff, and D. Murdoch (transl.) (1984), *The Philosophical Writings of Descartes*, Volume II, Cambridge University Press, Cambridge.

Descartes, R. (1984 [1641]), *Meditations on First Philosophy*, in J. Cottingham, R. Stoothoff, and D. Murdoch (1984), pp. 1–62.

Dray, W. (1957). *Laws and Explanation in History*, Clarendon Press, Oxford.

Feng, S. (2017), *Conceivability and Possibility: Conceivability as an Epistemic Guide to Possibility*, doctoral dissertation, Ruprecht-Karls-Universität Heidelberg, Heidelberg.

Fischer, B. (2016), "A theory-based epistemology of modality", *Canadian Journal of Philosophy* 46(2), pp. 228–247

Fischer, B. (2017), *Modal Justification via Theories*, Springer, Cham.

Fischer, B. and F. Leon (eds.) (2017), *Modal Epistemology After Rationalism*, Springer, Cham.

Fred-Rivera, I. and J. Leech (eds.) (2018), *Being Necessary: Themes of Ontology and Modality from the Work of Bob Hale*, Oxford University Press, Oxford.

Hale, B. (2013), *Necessary Beings: An Essay on Ontology, Modality, and the Relations Between Them*, Oxford University Press, Oxford.

Hintikka, J. (1957), "Modality as referential multiplicity", *Ajatus* 20, pp. 49–64.

Hirvonen, I., R. Koskinen and I. Pättiniemi. (Forthcoming). "Modal inferences in science: A tale of two epistemologies." *Synthese*.

Houghton, J. (2005), "Global Warming", *Reports on Progress in Physics* 68(6), pp. 1343–1403.

Hume, D. (2007), *A Treatise of Human Nature: A Critical Edition*, Volume 1: Texts, D. Norton and M. Norton (eds.), Oxford University Press, Oxford. [*Treatise*]

Kripke, S. (1959), "A completeness theorem in modal logic", *Journal of Symbolic Logic* 24(1), pp. 1–14.

Kripke, S. (1980), *Naming and Necessity*, Harvard University Press, Cambridge, MA.

Lowe, E.J. (2012), "What is the source of our knowledge of modal truths", *Mind* 121(484), pp. 919–950.

Meehl, G.A., T.F. Stocker, W.D. Collins, P. Friedlingstein, A.T. Gaye, J.M. Gregory, A. Kitoh, R. Knutti, J.M. Murphy, A. Noda, S.C.B. Raper, I.G. Watterson, A.J. Weaver, and Z.-C. Zhao (2007), "Global Climate Projections", in S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor, and H.L. Miller (2007), pp. 747–845.

Melia, J. (2003), *Modality*, Acumen, Chesham.

Neelin, J. D. (2011), *Climate Change and Climate Modeling*, Cambridge University Press, Cambridge.

Nozick, R. (1981), *Philosophical Explanations*, Harvard University Press, Cambridge, MA.

Nozick, R. (2003), *Invariances: The Structure of the Objective World*, Harvard University Press, Cambridge, MA.

Pritchard, D. (2005), *Epistemic Luck*, Oxford University Press, Oxford.

Roca-Royes, S. (2011), "Modal knowledge and counterfactual knowledge", *Logique et Analyse* 54(216), pp. 537–552.

Roca-Royes, S. (2017), "Similarity and possibility: an epistemology of de re possibility for concrete entities", in B. Fischer and F. Leon (2017), pp. 221–245.

Schubert, H. (1891), "The Squaring of the Circle: An Historical Sketch of the Problem from the Earliest Times to the Present Day", *The Monist* 1(2), pp. 197–228.

Szabó Gendler, T. and J. Hawthorne (eds.) (2002), *Conceivability and Possibility*, Oxford University Press, Oxford.

Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor, and H.L. Miller (eds.) (2007), *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, Cambridge.

Vaidya, A. (2017), "The Epistemology of Modality", in E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2017 Edition), URL = <https://plato.stanford.edu/archives/win2017/entries/modality-epistemology/>.

Vaidya, A. (2018), "Analytic Essentialist Approaches to the Epistemology of Modality", in I. Fred-Rivera and J. Leech (2018), pp. 224–244.

van Inwagen, P. (1998), "Modal Epistemology", *Philosophical Studies* 92(1), pp. 67–84.

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

Williamson, T. (2016), "Modal science", *Canadian Journal of Philosophy* 46(4–5), pp. 453–492.

Wimsatt, W. C. (2007), *Re-engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*, Harvard University Press, Cambridge, MA.

Yablo, S. (1993), "Is conceivability a guide to possibility?" *Philosophy and Phenomenological Research* 53(1), pp. 1–42.

Brandom and the Pragmatist Quest for Semantic Objectivity

JAAKKO REINIKAINEN

1 Introduction

Ever since their contemporary origins in the writings of C.S. Peirce and William James, pragmatist approaches to language have had problems (at least according to the critics) with objectivity. The basic problem of the pragmatist, who eschews commitments to substantial metaphysics such as the correspondence theory of truth, is to show how our epistemic, linguistic practices can be suitably *constrained* by how the world actually is in order for the practices to successfully, at least some of the time, *represent* the world, as opposed to merely "frictionlessly spinning in the void," to borrow John McDowell's famous phrase. The reason why spinning is generally considered to be a bad thing is that it leaves the door open for foundational skepticism regarding the veracity of our assertions and beliefs, with the close alternatives falling on the spectrum of anti-realism, deflationism, and quietism.

The aim of this paper is to critically examine the concept of semantic objectivity inherent in Robert Brandom's works, most importantly (1994/MIE). The reason for focusing on Brandom is that his ambitious aim is to combine the pragmatist preoccupation with our epistemic, justificational, linguistic practices with a robust enough account of objectivity to meet at least some desiderata of traditional realist intuitions. His "deontic scorekeeping model" therefore offers a particularly fruitful theoretical crossroads where the more abstract ideas above can break lances.

The main interest of this paper is exegetical, namely to clarify the aims, arguments, and problems of the account of semantic objectivity that Brandom presents in MIE. Concern-

ing Brandom's theoretical aims, I shall argue that there is some discrepancy between his formal and informal characterizations of the criteria by which his account is to be judged as adequate. In particular, it is not clear whether Brandom's idea of semantic objectivity as a "structural feature" of the scorekeeping practice suffices to cash out his claim that what determines semantic contents in a practice are the objects that claims made in the practice purport to represent.

After expounding on the discrepancy, I shall propose to reconstruct a mostly implicit line of argument in MIE, highlighted by the more recent developments of Brandom's work, which I think suffices to smooth it over. The missing piece for Brandom's pragmatist quest for semantic objectivity is *conceptual realism*, or the idea that both subjects and objects can be understood as conceptually structured. While conceptual realism only comes into explicit focus in Brandom's later works, I shall show that the essential idea is already operative in MIE.

Lastly, I shall note that although including conceptual realism in the theory is arguably the best way to fix its internal discrepancies, the inclusion is problematic insofar as conceptual realism as a metaphysical thesis is in no way motivated independently in MIE. While I remain neutral in this paper as to the independent plausibility of conceptual realism, I will argue that it represents an important watershed between MIE and Brandom's later works.

The paper's order of presentation starts with an outline of Brandom's pragmatist project in the philosophy of language, with a focus on the problem of semantic objectivity and the internal discrepancy mentioned above (2). In section 3, I shall further specify the discrepancy and what it would take to overcome it. Section 4 will argue that the task is best left for the conceptual realism that Brandom further develops in his later works. Finally, in section 5, I will argue against certain alternative ways to secure semantic objectivity and representational purport in the scorekeeping practice that do not appeal to conceptual realism.

2 The Core Architecture

The most important technical and philosophical contribution of MIE is what Brandom calls "the deontic scorekeeping practice." There are several ways one might approach the apparatus, and the secondary literature already offers several thorough presentations (Wanderer 2008; Loeffler 2017). For the purposes of this paper, two of Brandom's key claims are worth noting. The first is that assertions are primarily knowledge claims (MIE, 201). The second is that all three traditional main ingredients of knowledge—justification, propositional content, and truth—can be understood in terms of the deontic scorekeeping practice.

The natural place to start is with propositional contents, "of the sort that we express by the use of declarative sentences and ascribe by the use of 'that' clauses" (MIE, 5). Brandom contrasts two major contemporary strategies of coming to grips with such contents, namely by their truth conditions or by their inferential roles, and opts squarely for the latter. Thus conceptual contents at large, including subsentential and unrepeatable token expressions, are to be explained in terms of their contribution to inferential relations, which are divided into three classes: *commitment*-preserving, *entitlement*-preserving, and *material incompatibility* relations.

These semantic relations are in turn offered a pragmatic explanation in terms of what it is for an interpreter (a "scorekeeper") to take or treat herself and other subjects to be doing in drawing the aforementioned inferences expressed in assertions, where the appropriate doings are rendered in a normative, deontic idiom of sanctions. Intertwined in his strategy are what Brandom has later distinguished as the doctrines of *semantic* and *methodological pragmatism* (2011, 58, 61). Briefly, the claim of methodological pragmatism is that the theoretical point or purpose of postulating "meanings" (i.e., propositional, conceptual contents) is to explain proprieties of use, or why is it that certain uses of a word are correct while others are incorrect.[1] The main claim of semantic pragmatism in turn

[1] For Brandom, "proprieties of use" primarily concerns the business of drawing material inferences, not applications, e.g., in an ostensive setting. Nonetheless, to simplify the terminology, for the purposes of this paper I shall use "application," "use," and "drawing inferences" as synonyms.

is the foundationalist one that what conjoins a token expression with its meaning is the use which the speaker (and her community) makes of the expression.[2]

The paradigmatic move within the scorekeeping practice is to *attribute* a commitment to a claim (proposition) *p*. When attributing commitment *p* to a subject, the scorekeeper treats the subject not only as disposed to assert expressions of *p*, but also as obliged to uphold the claim in circumstances where *p*'s *truth* or the subject's *entitlement* to it comes into question. Furthermore, along with *p* itself, the scorekeeper also attributes to the subject commitment to all the claims that he takes to be the *material inferential* entailments.[3] These commitments are said to be *undertaken* by the subject, which is to say the subject herself may not *acknowledge* commitment to the same claims or their material entailments as her scorekeeper does. So, if the scorekeeper attributes to the subject commitment to the claim that "grass is green," and if the scorekeeper treats "grass is green" as materially entailing the claim "grass is colored," she will also treat the subject as committed to the claim that "grass is colored" whether the subject herself acknowledges commitment to either claim or not.[4]

To be committed to a claim is one thing, and to be entitled (i.e., justified) to it is another. Two facets of entitlement are worth noting here: on the one hand, *the default and challenge structure*, and on the other, the two mechanisms by which one may become justified to a commitment. First, one may become entitled (in the eyes of a scorekeeper) to a commitment by the *intercontent* mechanism of showing the committed claim as a material inferential consequence of commitments one already enjoys entitlement to. Second, one may become entitled to a commitment by the way of an *interpersonal* mechanism of deferring to another scorekeeper's commit-

---

[2] Though Brandom does not distinguish between methodological and semantic pragmatism in MIE, the claims are independent of each other, yet clearly fit well in the same picture.

[3] The concept of material inference that Brandom inherits from Sellars means "the kind of inference whose correctness essentially involves the conceptual contents of its premises and conclusions" (MIE, 97).

[4] This is a simplified example due to the fact that material inferences are non-monotonous, or sensitive to the context of the background claims available.

ment to a claim as a justification for endorsing it oneself (MIE, 175). In both cases, the status of being entitled to a commitment is social and normative in nature, i.e., relative to a scorekeeper, whose attributions of entitlements themselves are similarly open to evaluation by other scorekeepers.

The default and challenge structure's primary purpose is to stave off the justificatory regress that threatens both of these mechanisms. Since only tautological claims can justify themselves, any non-trivial claim must appeal to other claims for justificatory support, which then leads to the familiar dilemma where either appeal is made to premises that have already been used, or the chain of justification becomes infinite, with analogical worries facing the interpersonal mechanism. Brandom's solution is to admit that although every claim is in principle subject to a potential epistemic challenge, as a matter of social fact, some claims in the practice are treated as being such that everyone is by default entitled to make them, and that challenging them requires justification in order to be legitimate (MIE, 177).

The defining idea of MIE is to explain how a community of scorekeepers operating on these (simplified) principles can come to *institute* discursive, pragmatic norms sufficient for *conferring* propositional, conceptual, objective semantic contents on their token expressions. My focus here will be on the conferral half of the project, and more particularly on the semantically objective status of conceptual contents. Assuming that there are discursive norms governing what inferences it would be correct and incorrect for the scorekeepers to undertake and to attribute to each other, what guarantees that these norms deserve a specifically semantic interpretation, i.e., that it is *meaning* that these norms confer? In particular, why should we think that the norms are in any way related to the world of objects which the practice supposedly purports to represent, as opposed to being set by a malicious Cartesian demon, say?

A common way to explain how propositional contents come to represent the world is by allusion to truth in some way, e.g. by the correspondence theory. However, as already mentioned, Brandom opts out of a mixed ("two-factor") approach to propositional contents, which means he cannot appeal to truth as an explanans anywhere in his project. In fact,

he is a deflationist about truth, and sees the primary task as explaining what it is that we *do* in applying the truth locution to sentences rather than as giving a substantial semantic account about the truth predicate itself (MIE, 325-326).

Another initially promising candidate for explaining semantic objectivity that Brandom rejects is dubbed by him as the "I-we" sociality account. According to the I-we account, which bears some resemblance to Peirce's thinking on the end of inquiry as pointed out by Vitaly Kiryushchenko (2021), the epistemic subject proper is the whole community of rational interlocutors understood as a regulative ideal. The ideal community sets a kind of epistemic standard on truth in the sense that what would be held true by the ideal community (or some part of the community in epistemically ideal circumstances) would coincide with truth, or with what is correct according to the discursive norms. Brandom however famously denies that such a perspective exists:

> What is shared by all discursive perspectives is *that* there is a difference between what is objectively correct in the way of concept application and what is merely taken to be so, not *what* it is—the structure, not the content. (MIE, 600)

So, whence comes semantic objectivity and the anchoring of language in objects if not from truth or from an epistemically privileged collective perspective? I believe that here Brandom's answer—the structure, not the content—does not quite line up with the more informal phrasings of the criteria by which he thinks the project should be judged. But before contrasting those criteria, we must briefly clarify what semantic objectivity as a "structural feature" of scorekeeping amounts to.

Above I mentioned that the fundamental move within the scorekeeping practice is that of attributing a commitment (together with its material inferential entailments) to a claim. The set of claims which the scorekeeper treats the subject as having thus undertaken is contrasted with the set which the subject, both according to herself and according to the scorekeeper, acknowledges. So, in effect every scorekeeper keeps two sets of "books" on every other subject/scorekeeper: the set of commitments that the subject is disposed to assert (i.e.,

which she acknowledges) and the set that it would be correct for her to assert (i.e., which she has undertaken).

As already mentioned, according to the "official" answer of MIE, semantic objectivity is a "structural feature" of the scorekeeping practice. This means that the distinction between what is held correct by someone or everyone and what in fact is correct (even according to the scorekeeper herself) is made from within every individual perspective. But precisely because every perspective not only makes this distinction but is also subject to it, objectivity cannot be *defined* in terms of any single, epistemically privileged perspective.

Brandom supports this somewhat surprising claim by delivering objectivity proofs, the purpose of which is to ward off two threatening inferences:

No First-Person Ignorance (p) [p → (I claim that p)]

No First-Person Error (p) [(I claim that p) → p]

These inferences are threatening because, if true in the scorekeeping practice, they would make every scorekeeper take herself to be omniscient and incorrigible (MIE, 605). According to Brandom's proofs, which I won't be reviewing in depth here, semantic correctness does not collapse to the scorekeeper's perspective or to what she takes to be correct. In both cases, the pivot of the proofs is to show that the antecedents and the consequents of the threatening inferences are not incompatibility-equivalent, i.e., everything that is incompatible with the first is not incompatible with the second, and vice versa. The material incompatibility relation is defined in terms of commitment and entitlement: two claims are incompatible when commitment to one precludes entitlement to the other and vice versa (MIE, 160). Here, the distinction between being committed and entitled to a claim plays a major role, for although both threatening inferences are commitment-preserving, they are not entitlement-preserving (MIE, 606).

Brandom admits that passing the objectivity proofs is a "fairly weak" merit (MIE, 606). Nonetheless, he also claims that it is objectivity enough to meet the requirement for "a kind of correctness that answers to how things actually are, rather than to how they are *taken* to be, by anyone (including

oneself) or everyone" (MIE, 607). The kernel of Brandom's structural account of semantic objectivity is that it is not the semantic theorist's burden to formulate a set of criteria or a method by which we could find the claims that are correctly held to be correct within the practice – held correct in an ultimate sense, as it were. That matter is left solely to the practice itself, the "messy retail business of assessing the comparative authority of competing evidential and inferential claims" (MIE, 601).

My purpose now is to question whether the structural account of semantic objectivity and passing of the objectivity proofs suffices to fulfill Brandom's more informal characterizations of the criteria he sets for himself in MIE. To begin with, what Brandom thinks is important for semantic objectivity is the *source* of correctness for evaluating applications of conceptual norms, as he clearly states early on:

> The objectivity of conceptual norms requires that any attitude of taking, treating, or assessing as correct an application of a concept in forming a belief or making a claim be coherently conceivable as mistaken, *because of how things are with the objects the belief or claim is about.* (MIE, 63, my italics)

> The *objectivity* of representational content is a feature of the practices of assessing the correctness of representations. The status of representings as correct or incorrect, successful or unsuccessful, depends on how things are with what is represented, rather than on the attitudes of representers. What is distinctive of specifically *representational* correctness is this objectivity—the way in which assessments of representational correctness take representings to answer to what is represented, rather than to how what is represented is *taken* to be. It is the way in which the status being assessed outruns any particular attitude toward it. Understanding the objectivity of representational content requires understanding this particular structure of its authority and acknowledgement—what it is for those assessing the correctness of representings to cede authority over them to what is represented, to treat their correctness in practice as determined by those representeds. (MIE, 78)

In the next section, I shall argue that passing the objectivity proofs is only a necessary and not a sufficient condition to

meet the criterion that the source of semantic correctness is (at least partially) in the objects which claims made within the practice purport to represent. Moreover, in the next section I shall argue that MIE already contains the ingredients, if not the full argument, for a sufficient condition.

3 Three Levels of Semantic Objectivity

Above I noted a discrepancy between how Brandom characterizes the criteria of adequacy by which his account for semantic objectivity is to be judged and the objectivity proofs he delivers. As I initially pointed out, what Brandom aims for is an account where the source of correctness for evaluating representings (paradigmatically assertions and inferences, or more generally applications of norms) within the scorekeeping practice is at least partially in the objects that the claim-making practices purport to represent. What the objectivity proofs essentially achieve, however, is the merely negative point that claims about what is correct (in the sense of being true) do not collapse to (are not incompatibility equivalent with) claims about who is committed and entitled to what—not even in the case of the scorekeeper and her whole community. This is what Brandom wins by showing that the threatening inferences *No First-Person Ignorance* and *No First-Person Error* do not hold in the scorekeeping system. The problem is that this merely negative claim by itself leaves entirely open what, if anything, *does* determine which representings are correct and which are incorrect; in other words, it leaves entirely open the crucial question of the source of semantic correctness.

In order to make this distinction clearer, it is useful to distinguish between three levels of semantic objectivity that can be uncovered in MIE:

> *(AI)* A norm *n* is attitude-immanent for community *C* iff it is not possible for everyone in *C* to be mistaken about the correct applications of *n*.
>
> *(AT)* A norm *n* is attitude-transcendent for community *C* iff it is possible for everyone in *C* to be mistaken about the correct applications of *n*.

> *(PO)* A norm *n* is properly objective for community *C* iff the world of objects partially determines the correct applications of *n*.[5]

At the lowest tier of objectivity for norms, we find so-called *attitude-immanent* norms, prime examples of which are social norms such as greeting gestures and marriage institutions. In the case of these non-discursive norms, "it makes no sense to suppose that [the community] could be wrong about this sort of thing" (MIE, 53). A few specificational remarks should follow the biconditional definition. First of all, the principle is bound to incorporate an ineliminable measure of vagueness in regard to how finely the norm's content should be individuated, for it is typically the case that the community members do not have robust evaluative intuitions about all possible circumstances in which the given norm could be applied. So, I take it to be compatible with the (AI) status that a norm's content is not wholly transparent to the community in the positive sense that they could not find genuinely novel, as of yet unthought-of circumstances of application for the norm, although they could not then all be incorrect about how to apply it. Second, the collective judgment can be represented either by all the mature members separately or by some select, deferrable group of experts among them.

The class of norms the objectivity of which (AI) grading most readily befits is often called "social norms"; a slightly misleading term since all norms have a social character in some sense, at least for Brandom. While much more could (and should) be said about attitude-immanent social norms, e.g., how to distinguish them from mere conventions,[6] the

---

[5] As a reviewer pointed out, for Brandom the authority of objects to determine correctness of applications of norms can only ever be partial, not complete. The nominal reason for this is explained by his acceptance of phenomenalism about norms (see below), though in this instance I cannot go into the reasons that drive Brandom to endorse phenomenalism to begin with. I agree with the reviewer though that working in the background here is Kant's influence and also Sellar's (1956) criticism of the Myth of the Given, which broadly denies the possibility of non-conceptual epistemic access to objects.

[6] I refer the reader to Brennan et al. (2013) for a thorough conceptual study on social norms.

important point here is to contrast them with attitude-transcendent norms. In contrast to (AI), a norm that is (AT) has applications which are not necessarily and sufficiently determined as correct by the community's collective judgment. In particular, Brandom argues that we must understand conceptual norms as distinct from merely social ones precisely in that only conceptual norms are rightly called attitude-transcendent (MIE, 53-54).

The important point to realize now is that passing the objectivity proofs only amounts to semantic objectivity in the (AT) sense. The fact that everyone in a community may intelligibly (take themselves to) be mistaken about the correct application of a norm does not entail that it is the world of objects which determines what the correct applications—if any—are. This seminal point has already been appreciated at least by Bernd Prien (2010, 454). Importantly, Prien also proposes an interpretation of MIE according to which it does ultimately secure semantic objectivity, although as we shall see later, I think his argument does not work.

It may not be so intuitive to think of the difference between (AT) and (PO) as a question of *levels*, which implies a continuum, because they appear to answer different questions. As a helpful reviewer put the point to me, whereas (AI) and (AT) only concern the criteria for the *application* of norms, (PO) concerns the more fundamental issue of what source determines the very *content* of norms; a distinction the reviewer proposed to capture in "semantic" and "metasemantic" terms respectively. The metasemantic question is about the metaphysical issue of the source of semantic correctness, of which the semantic question about the criteria of application remains neutral.

There are indeed two distinct senses of "semantic objectivity" at play here, one concerning the criteria for applicability, the other criteria for determination of content. However, my purpose in squeezing the two onto one continuum is to clarify Brandom's claim about the *conferral* of objective semantic contents by discursive norms instituted by the attitudes mentioned above. The way I'm inclined to understand his thinking here is that the *original* source in the metasemantic or metaphysical issue—what determines the contents of norms—is solely with normative attitudes. This is a thesis

that he undertakes under the name *phenomenalism about norms* (MIE, 25).[7] Once the attitudes have "instituted" norms that fulfil the (AI) criteria of applicability, it becomes possible for them to fulfil the more demanding (AT) criteria as well. However, at this juncture, what is also supposed to change is the metasemantic or metaphysical issue concerning the source for determining the contents of norms. In effect, the source of authority is in a way extended from the attitudes to the objects such that the latter come to be "incorporated" in the practice, to exercise "mediated" authority of their own over the attitudes. This latter thesis goes by the name *normative phenomenalism* (MIE, 627).

The reason why I take it to be justified to situate (AI), (AT), and (PO) on a continuum of levels rests with my reading of Brandom's larger project in MIE that seeks to explain the conferral of semantically objective contents by discursive norms implicit in practices. The shift from institution to conferral is supposed to be a continuous process, which I take means that the criteria by which the shift itself is judged as successful should be the same as what are used to evaluate institution and conferral separately, even if the shift contains an implicit, important distinction. There is indeed a kind of a "jump" from (AT) to (PO), but one that purports to reflect the qualitative shift which Brandom pursues under the conferral thesis. What changes during the conferral is the metasemantic or metaphysical source of content, the discursive authority that is no longer solely with the attitudes but becomes shared with or passed on to the objects. While Brandom appears to think that passing the objectivity proofs suffices to cover the shift from (AT) to (PO), I side with Prien in that something else is required to turn the merely negative claim about the applicability of norms to the positive claim about the determination of content. To repeat, the reason why (AT) norms arguably do not suffice for representational purport is that, as was seen above, a norm being (AT) does not foreclose the possibility

[7] I won't seek to give a strict definition for phenomenalism here, for I believe its spirit in Brandom's works is primarily programmatic and thus strategically malleable according to the context. However, it is also true that the fact exposes Brandom's key claims to hindering polysemy, as noted, e.g., by Jeremy Wanderer (2008, 74, fn.).

that its content is indeterminate (i.e., determined by nothing) or then it is determined by a Cartesian demon. Were that the case, it becomes hard to argue that the norm purported to represent *anything*, much less the world, which is why the claim to (PO) status for conceptual norms is crucial for Brandom to achieve.

4 The Pragmatist Route to Semantic Objectivity of Contents Goes Via Conceptual Realism

Above I argued that Brandom's official account of semantic objectivity, which rests on the objectivity proofs, does not suffice to meet the informal yet clear criteria that he sets for himself elsewhere in MIE. What remains unclear with regard to the semi-metaphorical conferral claim is how norms that are (AT) by their objectivity level may explain the rise of semantic contents robust enough to meet the (PO) standard. In this section, I shall argue that the missing piece is already inherent in MIE, although Brandom started developing the details of the answer only in his later works.

As I already explained, the shape of the problem of semantic objectivity for Brandom is to explain how the objects of the world can come to be incorporated in or mediated by our discursive attitudes in the sense that the original authority of the attitudes is in a way extended to the worldly objects. The sense in which the world is "incorporated" into practices should be initially differentiated from the way in which sounds and marks merely *convey* intentionality. This text conveys the intentionality of my assertions to you, but in no way do the pixels (or the ink of the printer) exercise authority over the correctness of what I say, which is only to point out the familiar idea of the sign's arbitrariness. Brandom's idea of "lumpy practices" seeks to capture a more robust sense in which the world partakes in discursive practices, somewhat like bats and balls "partake" in baseball, where their purely material aspects, while in a sense contingent, are not *as* arbitrary as those of the signs we use in making assertions (MIE, 632).

How this works in practice can be appreciated by the (in Brandomian circles) hackneyed example of the litmus paper test. Consider the following causal chain of events:

1. The subject has a discursive attitude describable as a disposition to draw the inference "If some substance tastes sour, it will turn litmus paper red."
2. The subject has a perceptual experience of a substance that tastes sour.
3. The subject has a consequent perceptual experience of the substance turning litmus paper blue.
4. The subject loses her attitude-disposition to infer "If some substance tastes sour, it will turn litmus paper red."

In this example, we can see the causal entanglement of practices and the world. On the side of the practices, we have events (or states) (1.) and (4.), and on the side of the world, we have events (2.) and (3.). (Alternatively we could replace, in this instance, the term "practices" with that of "abilities," for although in MIE Brandom's official stance is that the relevant dispositions can only emerge in the context of intersubjective practices, elsewhere he is less committal about this point.) Of course, the whole chain of events is part of the same world, i.e., the distinction between discursive practices/abilities and the world is drawn from within the world when viewed in purely causal terms. A similar story on the side of action could be told where the subject's attitudes are the cause of changes in the world rather than themselves causally changed by how the world is (Brandom 2008a, 178; MIE, 332-333).

The chain of events (1.)-(4.) above gives us a rudimentary grasp of how the world *causally* constricts the practices/abilities paradigmatically by affecting our dispositions to draw inferences.[8] Of course, not all such causal effects should be counted as having anything to do with how the facts of the

[8] The relevant practices or abilities are algorithmic in kind, the core of which Brandom identifies as a four-step feedback loop of action and perception. In *Between Saying and Doing* (esp. Ch. 1-2) he develops a new type of regimented logical vocabulary to discuss how such relatively simple systems, which arguably can be taken to exhibit primitive forms of a practical, know-how type of intentionality, can give rise to the theoretical, know-that type of intentionality. I cannot here discuss the details of the project.

world *justify* moves in discursive practices.[9] How is it then that the world *normatively* constrains our practices/abilities? The key idea here is Brandom's commitment to *conceptual realism*, encapsulated by the notion inherited from Frege that facts just are true claims (i.e., what is *claimed* and not the *claiming* of it) (MIE, 327). Seen from the subject's own perspective, the claim to which she acknowledges commitment at (1.) turns out to be false in the transition from (2.) to (3.), i.e., in the face of the perceived fact that there is a sour-tasting substance that turns litmus paper blue instead of red. Here, the crucial difference between a claim merely *taken* as true and a claim that *is* true is made from within the practices/abilities as opposed to within the world: it is the difference between the subject attributing commitments (either to others or to her past self) and undertaking them herself (in the present). Since the subject-relative normative status of a claim as a fact depends on whether it is *only* attributed or *also* undertaken, and since the attitudes are already something involved in the causal realm of facts, the mechanism by which facts come to exercise authority over attitudes is given by the scorekeeping apparatus considered as causally integrated with the world in complex ways.

The key claim of conceptual realism is that both facts and attitudes are conceptually structured according to two different readings of the generic material incompatibility relation. On the side of the world, the concept of the object can be understood as "repelling" incompatible properties under an *alethic* sense of necessity. On the side of the practices, subjects can be understood as "repelling" incompatible commitments under a *deontic* sense of necessity. In Brandom's words:

> It is *impossible* for one and the same *object* to have incompatible *properties* at the same time. But it is merely *impermissible* for one and the same *subject* to have incompatible *commitments* at the same time. (2008a, 191)

---

[9] Brandom (2001, 107) is strongly critical of reliabilist theories of justification that take causal, probabilistically reliable processes as at least in some cases sufficient to justify beliefs and assertions. I cannot enter this debate here, but the important point is that for Brandom, purely causal relations are not sufficient to account for the justification of beliefs or assertions: the normative element is also required.

We can now better appreciate in what sense the world becomes "incorporated" in or "mediated" by discursive practices, following the litmus paper example above. Brandom's idea is that the succession of events (1.)-(4.) can be understood from two different *modal* perspectives, depending on whether it is described objectively as what does happen or subjectively as what ought to or may happen. The world and the practices are ontologically speaking two halves of the same event or process, structured in the generic modal sense of a material incompatibility relation, which Brandom takes to be the key conceptual notion.

However, at this point it seems that it would be equally correct to say that the practices are incorporated in or mediated by the world rather than the other way around. To make an already impressive amalgamation of theses more complicated, Brandom also pursues an explanatory order he attributes to Hegel, according to which the objective side of alethic modal incompatibility *relations* must be understood and explained in terms of the subjective side of deontic modal incompatibility *processes* (2002, Ch.6).

It is noteworthy that the term "conceptual realism" appears nowhere in MIE, and thus it is appropriate to wonder whether the idea really is relevant for the issue of semantic objectivity as opposed to a late-coming, separate topic. The impression is reinforced by the fact that MIE's primary pragmatist strategy centers its explanatory force on the score-keeping practice, which assumedly is supposed to be independent of ontological issues concerning the constitution of the world. Furthermore, there is an active reason for Brandom to avoid undertaking any unnecessary ontological commitments as a consequence of his semantic theorizing, namely his fundamental opposition to the truth conditional strategy and the correspondence theory of truth that goes with it. Brandom accuses the correspondence theorist of confusing acts of claiming that something is true with the content of what is thereby said in the sense that what is true – i.e., the facts—is supposed to explain what it is for a claim to be true, i.e., its content understood as truth conditions (MIE, 330).

That being said, when Brandom echoes Frege in claiming that "Facts just are true claims," a careful reading shows that he is not by that token merely making the deflationist nega-

tive claim that truth is not a semantically explanatory relation between language and world. Instead, towards the end of MIE he proposes an alternative way to construe that relation:

> Concepts conceived as inferential roles of expressions do not serve as epistemological intermediaries, standing between us and what is conceptualized by them. This is not because there is no causal order consisting of particulars, interaction with which supplies the material for thought. *It is rather because all of these elements are themselves conceived as thoroughly conceptual, not as contrasting with the conceptual*. (MIE, 622, my italics)

> The conception of concepts as inferentially articulated permits a picture of thought and of the world that thought is about as *equally*, and in the favored cases *identically*, conceptually articulated. (Ibid.)

Condensed here is the main thesis of what Brandom later on has dubbed conceptual realism, or the idea that the world as such is conceptually structured. There is no *ontological* category distinction between predicates and properties: instead there is *identity*. The nature of the identity is modal, split between the alethic and deontic sides (2019, 54). *How* exactly the sides are supposed to be combined is of course a massive question, one that Brandom does not tackle in MIE and which thus falls outside the scope of this paper.

But if Brandom does indeed espouse conceptual realism already in MIE as the key to the conferral thesis that is to patch over the jump from (AT) to (PO) objectivity, why does he not explicitly say so? One reason I can think of is that at the time he did not have a well-thought-out idea of how to connect conceptual realism as an independent metaphysical stance with the scorekeeping practice, or to give an encompassing enough of account of it. Yet the idea that the world and discursive practices are causally integrated with each other is clearly stated and important for securing the condition, which Brandom sees as central, that the world serves as a dual constraint (normative and causal) on practices, even if the point is never brought into detailed discussion (MIE, 331, 332, fn.).

5 Why Conceptual Realism Is Essential for Proper Objectivity

To conclude this paper, I shall argue against certain alternative ways to understand Brandom's claim that the scorekeeping practice is able to confer objective semantic contents on token expressions.

Andrea Clausen (2004) for one argues that conceptual realism is non-essential and in fact a distraction from Brandom's aim of accounting for objective contents in terms of discursive practices. The basic reason why she considers conceptual realism redundant is that she thinks Brandom's scorekeeping account alone can afford an explanation of how token expressions can come to exhibit representational purport. The problem, however, is that she does not adequately distinguish between attitude-transcendence and what I have called proper objectivity, namely between the negative claim that everyone could be incorrect in (some) of their assertions and inferences and the positive claim that it is the world that determines the semantic incorrectness of assertions and inferences. Again, the fundamental reason why attitude-transcendence does not amount to proper objectivity is that, even if every subject in practice necessarily *presumes* a difference between what is taken to be correct and what is correct, and that there is only one correct set of assertions and inferences everyone should acknowledge, it does not follow that it is the *world of objects* which determines the identity of the set, or even that there *is* such a set. Here's a telling excerpt of this *non-sequitur*:

> What we claim to be correct can always turn out to be incorrect. Put alternatively, this means that we rub ourselves against a resistant reality. Second, what is correct is supposed to be independent of what anybody or all take to be correct. Put alternatively, this means that we refer to one and the same world. (Clausen 2004, 217)

In fact, the reason our claims can always turn out to be incorrect, as far as the scorekeeping practice is concerned, may be that the contents are actually indeterminate or then determined by a Cartesian demon. And even if everyone agrees that what is correct is independent of what everyone takes to

be correct, it remains possible that there is no reference to one and the same world.

Ronald Loeffler (2017) sees the problem between deriving (PO) from (AT) without further argument more clearly. Returning to the litmus paper parable, what Brandom wants to say is that *by* treating two of her commitments as materially incompatible with each other, the subject *takes* her commitments to be purporting to represent a singular object, namely the natural kind acid, for objects are (in part) defined as those entities which repel incompatible properties in the alethic modal sense. Loeffler raises the question, however, of why we should interpret the subject as purporting to represent an *object* by taking two of her commitments to be incompatible, for on the face of it we might equally well interpret her taking the incompatibility to amount to nothing more than a prohibition against endorsing two given assertion types (Loeffler 2017, 147). In other words, how does the intra-practice matter of which assertions are taken to be incompatible translate into the extra-practice matter of representational purport?

Loeffler's answer on behalf of Brandom is that, although from *our* point of view as external theorists the subject of the acid parable is not yet definitely purporting to represent anything beyond her practices or abilities, from the subject's *own* perspective it appears that the acid itself serves as the external standard of her commitments, which hence purport to represent how things really stand with acidic substances (2017, 148).[10]

The distinction between the native subject's own perspective and that of the external theorist's cannot, however, offer a sufficient reason to claim that the scorekeeping practice includes norms with representational purport or (PO) objectivity grading. The reason is, again, that each of the predicates

[10] Note that saying this is compatible with Brandom's insistence that although the subject is from her own point of view purporting to represent objects, the purport may be completely *implicit* in her practices or abilities in that she may not be able to explicitly assert that her commitments represent something external (Loeffler 2017, 149). The distinction between an implicit ability to *do* something that is independent of the explicit ability to *say* what one is doing is as important to Brandom's pragmatist account of intentionality, though it is also largely orthogonal to the issues I'm addressing here.

"takes to purport to represent an object" and "purports to represent an object," or alternatively "takes to be correct" and "is correct," and the predicates have distinct extensions, and claiming one does not entail the other. In particular, since Brandom's final major statement in MIE is that we are in fact engaged in the scorekeeping practice ourselves (the move he calls "the collapse of perspectives"), any difference to the extent which so starkly distinguishes between the native scorekeeper and her external interpreter cannot hope to be adequate as an account of actual representational purport, if by "actual" we mean whatever it is that we do in purporting to represent objects. Applying Loeffler's response to our own case, even if it is true that it (necessarily) *appears* to us that we are responsive to objects of the world when encountering incompatible commitments, it does not follow that we *really* are purporting to represent such objects.[11]

To end this section, I wish to reject one further argument which seeks to establish representational purport in the scorekeeping practice without resorting to conceptual realism as an independent metaphysical theory. Bernd Prien (2010) argues that what is needed to ensure proper objectivity is a special norm called the "principle of rational rectification" (PRR). The principle of rational rectification, which Brandom introduces in *Between Saying and Doing*, states that subjects are obliged to rectify the incompatible commitments they have committed themselves to. Indeed, as we already saw in this section, the principle in part defines the concept of the discursive subject for Brandom (2008a, 193). Prien claims that

> [p]ractices that include such a norm of rational rectification warrant an interpretation according to which the conceptual norms and thus the deontic statuses of the speakers are not determined by the deontic attitudes present in a community, but rather by the way the world is. Whenever a speaker runs into incompatible commitments because of the way the world is (for example, because there are sour-tasting liquids that do not turn litmus paper red), she is obliged to modify some of the inferential relations she acknowledges. In order to make sense of this obligation, we have to assume that it is the world that determines

[11] Loeffler also sees conceptual realism as an important part of Brandom's later attempts to account for semantic objectivity (2017, 178-179).

> what follows from what, and not the individual subjects, the experts, or the community as a whole. For even inferential relations accepted by the community as a whole have to be modimodified if this is the best way to remove an incompatibility. (2010, 455)

Prien claims that the PRR is a sufficient condition to warrant the properly objective status to conceptual contents in the discursive practices, for it is the only way to make sense of this obligation. It is difficult to see how that follows however, for it is perfectly intelligible that everyone in the practice is obligated to rectify their incompatible commitments and that the world does not determine what commitments really are incompatible. Furthermore, it is not clear how precisely the world is supposed to *oblige* subjects to rectify their incompatible commitments other than in the metaphorical, causal sense of obligation (Brandom 2008b).

A similar point applies to another special norm also mentioned by Prien, which we might call the *inter*subjective principle of rational rectification as opposed to the *intra*subjective PRR. The intersubjective PRR, first proposed by Loeffler (2005), states that different subjects A and B are obligated to rectify their commitments that are incompatible with some commitments of the other. For one, the intersubjective PRR seems to complicate Brandom's claim that we can define subjects as units of accountability *qua* subjects to intrasubjective PRR. If PRR is extended from intra- to intersubjective incompatibility relations, are we to conclude that two distinct subjects can form a singular unit of discursive accountability?

More acutely though, it remains unclear how PRR in either its intra- or intersubjective versions is supposed to entail that subjects really are responsible to the world in what concerns the correctness of their commitments. For the issue of *in virtue of what* commitments really are incompatible is orthogonal to whether and in what sense subjects are obligated to rectify their incompatible commitments. Even if it is the world that somehow non-metaphorically obliges the subjects to rectify their incompatible commitments, something which Brandom explicitly denies (2008b), it is a different matter to establish whether the world also determines (and does not merely appear to determine) which commitments are incompatible. So

PRR alone does not entail that the scorekeeping practice that includes it also includes norms with representational purport.

Conclusions

To summarize, the crucial problem for Brandom's pragmatist project in MIE is to explain how the norms instituted by attitudes can confer propositional contents robust enough to be *about* worldly objects. In order to achieve this, he argues that the practice incorporates or mediates objects, somewhat like games "incorporate" physical objects into their rules. However, a prerequisite for the incorporation is that Brandom must undertake ontological commitments regarding the nature of the objects as such, namely that they too are conceptually structured. The essential idea of conceptual realism already operative in MIE is that the subject/object divide can be explained in terms of the modal divide between alethic and deontic halves. This, I have argued, is Brandom's best strategy in MIE for explaining why the scorekeeping practice should be interpreted as including genuinely representational properties.

The cost of embracing conceptual realism, however, is that it ultimately means expanding the base explananda with which Brandom operates in MIE. The official strategy of the book is to explain how norms instituted by attitudes may confer propositional contents that are objective and representational in the sense that they normatively answer to the world of objects. The main explanatory primitive on the subjective side is the concept normative attitude. However, there are no corresponding primitives available on the objective side to argue for the truth of conceptual realism. It is as if in the course of the book Brandom is driven to embrace conceptual realism because of his starting point with normative attitudes, which alone cannot secure an objective enough relation to the world to establish representational purport. In an interesting narrative twist, this result is not too different from what Brandom considers to be a central mistake of early analytic philosophy:

> Some previous varieties of logical atomism had distinguished themselves by their insistence that the only way any expression, sentential or not, could have content or contribute to the content

> of an expression of which it is a part is by standing for or representing something. Thus, not only did these views grasp the nettle of commitment to negative and conditional facts, they also were committed to "not" and "if… then…" standing for some element in a complex state of affairs. The undertakers of such commitments are admirable more for their conceptual heroism than for their good sense. (MIE, 76)

The lesson here is that we should be mindful about the possible ontological implications our theorizing on language and meaning leaves us with, for otherwise we risk putting the cart before the horse. Brandom's appeal to conceptual realism without sufficient argumentative support risks doing that, although as I have shown he has taken measures to rectify the matter later on. The final word on the matter belongs to further study, however.

University of Tampere

References

Brandom, Robert (2019), *A Spirit of Trust: A Reading of Hegel's* Phenomenology, Cambridge MA, Harvard University Press.

Brandom, Robert (2011), *Perspectives on Pragmatism: Classical, Recent, and Contemporary,* Cambridge MA, Harvard University Press.

Brandom, Robert (2008a), *Between Saying and Doing*, New York, Oxford University Press.

Brandom, Robert (2008b), "Reply to 'Are Fundamental Discursive Norms Objective?'", published in *Robert Brandom: Analytic Pragmatist,* Berlin, De Gruyter.

Brandom, Robert (2001), *Articulating Reasons: An Introduction to Inferentialism,* Cambridge MA, Harvard University Press.

Brandom, Robert (1994/MIE), *Making It Explicit: Reasoning, Representing, and Discursive Practice*, Cambridge MA, Harvard University Press.

Brennan et al. (2013), *Explaining Norms*, Geoffrey Brennan, Lina Eriksson, Robert E. Goodin, and Nicholas Southwood Brennan, Oxford, Oxford University Press.

Clausen, Andrea (2004), *How Can Conceptual Content be Social and Normative, And, at the Same Time, be Objective?* Frankfurt, Ontos. Web.

Kiryushchenko, Vitaly (2021), "I, Thou, and We: Peirce and Brandom on the Objectivity of Norms", published in *The Social Institution of Discur-*

sive Norms: Historical, Naturalistic, and Pragmatic Perspectives, edited by Leo Townsend, et al. Taylor & Francis Group, ProQuest Ebook Central.

Loeffler, Ronald (2017), *Brandom*, Polity Press. ProQuest Ebook Central.

Loeffler, Ronald (2005), "Normative Phenomenalism: On Robert Brandom's Practice-Based Explanation of Meaning", *European Journal of Philosophy*, 13:1, pp. 32–69.

Prien, Bern (2010), "Robert Brandom on Communication, Reference, and Objectivity", *International Journal of Philosophical Studies*, 18:3, pp. 433-458.

Sellars, Wilfrid (1956), *Empiricism and the Philosophy of Mind*, edited by Robert Brandom, Cambridge MA, Harvard University Press.

Wanderer, Jeremy (2008), *Robert Brandom,* Trowbridge, Cromwell Press.

The Dream Self and the Waking Self

HEIDI HAANILA

The self is the main character in one's life and an important theme in the philosophy of mind. The concept of self is multifaceted and notoriously ambiguous, and philosophers debate about the definition of self.[1] In this paper, I approach selfhood by examining dreaming and ask, how the study of dreaming can contribute to the definition of self. Or in other words, what can the dream self reveal about the waking self? Dreaming is an altered state of consciousness, which often involves extensive alterations in self-consciousness, and as such it provides an attractive way to the research of self. I start the paper by viewing the concepts of selfhood in terms of a pattern theory of self and drawing the most general distinction within the self, that is the distinction between the experiential and reflective self. Then, I consider dreaming and the methodology of using altered states of consciousness in the study of self. After which, I examine the character of the dream self and how it differs from the waking self in terms of both experiential and reflective self. The idea is that the study of dreaming can function as an instrument to distinguish different aspects of self from each other, and to bring out the connections between them and the necessary features of self. While I mainly focus on defining the most fundamental aspect of the experiential self, I also briefly consider the opportunities to study the reflective self through dreaming.

---

[1] For an overview see e.g. Gallagher (ed.) 2011 or Siderits et al. (eds.) 2011, the last mentioned involves also comparisons between eastern and western notions of self.

1. Concepts of self

Self is of the utmost importance in one's life. *Self* is the subject of experience, thinker of thoughts, and agent of action. As such 'self' or 'selfhood' is an umbrella term that comprises numerous features of self and self-consciousness. [2] In order to bring together and combine different theories of self, Shaun Gallagher has developed a *pattern theory of self* (Gallagher 2013; Gallagher & Daly 2018, see also Newen 2018). According to the pattern theory, an individual self is constituted of a complex pattern of characteristic features or certain aspects of self. Gallagher argues that the pattern theory is a useful way to organize the multidisciplinary discussion of what constitutes a self. This is because within the pattern theory, various interpretations of self can be seen as compatible or commensurable rather than being in opposition. Gallagher (2013, Gallagher & Daly 2018) presents a tentative list of significant features that contribute to the constitution of self. These aspects can be seen as variables that take different values and weightings in the dynamic constitution of self. Gallagher emphasizes that an individual self may lack a particular characteristic feature and still be considered a self. Gallagher's (2013, 3-4) list includes the following aspects:

(1) *Minimal embodied aspects*: core biological aspects, which allow the system to distinguish between itself and what is not itself.

---

[2] It can be noted at the outset that the concept of 'experiential self' used in this paper (Section 1.1.) entails the idea that self and consciousness are intertwined. This idea is denied in theories which claim that there can be experience without self. For instance, according to Pylkkö (1998), aconceptual and asubjectivist experience is fundamental and self is a construction. Or generally, the so called no-self-theories, which are advocated in many Eastern philosophies, argue that the self is illusory (see e.g. Albahari 2006; Metzinger 2009, for discussions about no-self theories, see Siderits et al. 2011). Some of the dispute between the theories highlighting the experiential self and no-self can be considered terminological; they simply mean different things with the notion 'self' (see e.g. Zahavi 2011; 2014, Ch. 4). Thus, it can be noted that an endorsement of a no-self view would not undermine the general idea of differentiating between the layers of self that is conducted in this paper, but it would entail specifying concepts for some features of the 'experiential self' without reference to 'self'. See also fn. 5.

This is an extremely basic aspect of all kinds of animal behavior, and include the aspects that define the egocentric body-centered spatial frame of reference.

(2) *Minimal experiential aspects*: to the extent that the bodily system can be conscious, it will pre-reflectively experience the self/non-self distinction in the various sensory-motor modalities available to it. Such aspects contribute to an experiential and embodied sense of ownership (the "mineness" of one's experience), and a sense of agency for one's actions (Gallagher 2000).

(3) *Affective aspects:* reflect a particular mix of affective factors that range from very basic and mostly covert or tacit bodily affects to what may be for her a typical emotional pattern or mood.

(4) *Intersubjective aspects*: humans have the innate capacity for attuning to intersubjective existence, and after language learning, this intersubjective aspect is internalized and takes the form of a dialogical process that helps to constitute the self.

(5) *Psychological/cognitive aspects*: traditional theories of the self focus on various psychological and cognitive aspects. These range from explicit self-consciousness to conceptual understanding of self as self, to personality traits of which one may not be self-conscious at all. In addition, there are strong arguments for psychological continuity and the importance of memory in the literature on personal identity (e.g. Shoemaker 2011). One can also include representational aspects here, meaning, approximately, one's ability to represent oneself as oneself.

(6) *Narrative aspects*: the basic idea is that selves are inherently narrative entities and that our self-interpretations have a narrative structure (Schechtman 2011). For some theorists, narratives are constitutive of selves.

(7) *Extended aspects*: self may include physical pieces of property, such as clothes, homes, and various things that we own. We identify ourselves with the items we own, and perhaps with the technologies we use, the institutions we work in, or the nation states that we inhabit.

(8) *Situated aspects:* include, for instance, the kind of family structure and environment where we grew up, and cultural and normative practices that define our way of living.

Different theories of self emphasize different aspects, and the pattern theory provides a framework in which the complexity of self can be endorsed. However, the pattern theory as such does not explain how the aspects are connected or what kinds of relations prevail between them. Crucially, it does not take a stand on whether some aspect or combination of aspects is necessary for self. Thus, the pattern theory does not provide answers to the quest of self, but the character of self still requires elaboration and clarification.

An advantage of the pattern theory is that it assists in distinguishing between various features of the self and in seeing how the connections between these features contribute to self. Since the list of aspects is rather long, I condense the features down to a distinction between the experiential and reflective self. This generic distinction is generally accepted and often made, although it is conceptualized differently in different theories.[3] For the purpose of this paper, the experiential self consists of embodied, experiential and affective aspects, and the reflective self consists of the psychological-cognitive and narrative aspects.[4] I elaborate these notions briefly below, and

---

[3] The distinctions has been drawn in terms of, for instance, intransitive and transitive self-consciousness (Kriegel 2004), minimal and narrative self (Gallagher 2000), nonconceptual and conceptual self-consciousness (Bermudez 2001), and pre-reflective and reflective self-consciousness (Zahavi 2005).

[4] That is, I exclude intersubjective, extended and situational aspects from the scope of this paper. These aspects are interesting for the distinction of two forms of self, since they seem to be incorporated in both experiential and reflective self and thus, might be used in investigations of the interconnections of the two forms. However, they are beyond the scope of this paper.

In addition, it can be noted that in more recent version of the pattern theory (Gallagher & Daly 2018), Gallagher has also added behavioral, reflective and normative aspects. Of these aspects, the two last mentioned can easily be included in the reflective self. However, neither of these aspects are necessary to discuss in order to present the idea of this paper and, for simplicity and brevity, the original (2013) version of the pattern theory is applied here.

then proceed to the argumentative part of the paper. In that part, I propose that the examination of the dream self is useful in revealing different layers of the self and can be used to elicit the necessary aspects of self.

1.1. The experiential self

The *experiential self* refers to the most fundamental form of selfhood, which is the basis for the cognitively more demanding and complex reflective self (Bermudez 2001; Gallagher 2000; Kriegel 2004; Zahavi 2011, 2014).[5] This concept of self emphasizes that self is always present in experience. In this elementary sense 'self' is connected to the *subjectivity* of experience. Even when one is not thinking about or focusing on herself at all, there is a subtle awareness of herself in that mental state: she is aware of herself as the owner or *subject* of the experience, and this holds true for all of her experiences. In other words, the experiential self does not refer to self as an object or content of consciousness, to a *what* of experience. Instead, it refers to the *how* of experience that is to the first-personal presence of experience. It refers to the fact that the experiences I am living through are given differently to me than to anybody else. Thus, the experiential self is an integral part of our consciousness and can be identified with the ubiquitous first-personal character of experience.

In terms of the pattern theory, the experiential self seems to include the experiential aspects by definition. In addition, many theories highlight that our basic sense of self is essentially embodied and affective (see e.g. Bermudez 2001; Colombetti & Thompson 2008; Gallagher & Zahavi 2008; Varela et al. 1991). All the experiential, embodied, and affec-

[5] The term 'experiential self' has been used by Zahavi (2011; 2014) synonymously with the terms 'pre-reflective self-consciousness' and 'for-me-ness'. Zahavi has developed a sophisticated phenomenological theory of self, and the characterization of experiential self in this paper follows Zahavi's ideas, which underline experiential self as the most fundamental form of selfhood and a constitutive feature of consciousness. However, I elaborate the notion of experiential self in terms of pattern theory which Zahavi himself does not. It also can be noted that the ideas presented in this paper are not depended on or restricted only to Zahavi's conception of self, but can be applied to others notions of self too, see fn. 3.

tive aspects of self are present in experience already without being objects of reflection. Many times self-consciousness is described as a first-person perspective (1PP in brief), and this description involves a spatiotemporal perspective that specifies a viewpoint on the environment (Metzinger 2013; Windt 2015). However, as Zahavi (2005; 2011; 2014) underlines, the essence of the notion of 1PP is that the perspective is *personal*; it is subjectively experienced.

These considerations show that although the experiential self is the most elementary form of selfhood, it involves several aspects of self that intertwine together in experiences. The richness of embodied 1PP can be noticed in a simple example of experiencing perceptions of the environment during walk. When I am walking on a sea shore, I can see the cliffs, waves and forest. All these things have a certain location in relation to my body. I can also hear the waves on the shore and the singing of birds in the forest. Further, by means of proprioception, I can sense my movements and the positions of my body that maintain its balance when walking; I need to adjust my steps to the perceived shape of the rocky shore. I can feel the excitement of being in a new place and joy when I manage to see a rare bird. Overall, the experiential self involves embodied, experiential and affective aspects, and is present in experience without any explicit thinking of self.

1.2. The reflective self

In order to do justice to human selfhood, the notion of the experiential self needs to be supplemented with the notion of the *reflective self*, which is higher in the cognitive hierarchy (Bermudez 2001; Gallagher & Zahavi 2008; Kriegel 2004; Zahavi 2005, 2014)[6]. The reflective self is capable of language use and introspection; it deliberates actions, and is shaped by its values, beliefs, commitments, goals, and decisions. This form of selfhood involves reflective self-consciousness that is

[6] The term 'reflective self' is derived from Zahavi's notion of 'reflective self-consciousness' that is used in contrast to the 'experiential self' or 'pre-reflective self-consciousness'. In addition, the notion 'reflective self' aims to take a neutral stance towards theories that highlight narrativity, although the notion embraces the narrative aspects as a significant feature of self.

the capacity to take oneself as the object of one's reasoning and to think of oneself as oneself. Reflective self-consciousness is essentially linked to our general conceptual capacities and reasoning skills. Thus, it involves at least the psychological-cognitive aspects of self. By means of reflective self-consciousness, one can focus her attention on herself, and evaluate and direct her action. Reflective self-consciousness is a necessary condition for moral self-responsibility, normative evaluation and self-critical deliberation and for that reason many theories of self find it essential (Moran 2001; Korsgaard 2009; Schechtman 2011). In addition, philosophers have been interested in the unique features of self-knowledge and self-conscious thoughts, which refer to the subject by the use of first-person pronoun 'I' and have specific epistemic and motivational features (Gertler 2011; Perry 1979; Shoemaker 1968).

Further, the reflective self has the capacity to formulate narratives and thus, involves narrative aspects of self (e.g. Gallagher 2000; Gallagher & Zahavi 2008; Schechtman 2011). This highlights the wide time-perspective of the reflective self; it is not limited to the immediate experience but extends from past to future. With these reflective and narrative capacities, one can engage in a meaningful life as a part of a community. For instance, I exercise my reflective-narrative dimensions of selfhood when I ponder about what I should do on the weekend. Should I visit an old friend in another town, or finish a work project that is significant for my future career, or take time for myself and renew my energy? Altogether, the reflective self is connected to certain ways of thinking and acting that are frequently considered characteristically human; to be a person with memories and future plans, and a deliberating moral agent.

2. Dreaming as a research tool in the study of self

A number of the recent approaches to the philosophy of mind endorse a multidisciplinary methodology and strive to be empirically informed (e.g. Gallagher 2013; Mandik 2007; Metzinger 2013; Thompson 2015; Windt 2015). In terms of self, these multidisciplinary approaches entail forming a theory of self that is conceptually coherent and empirically plau-

sible at the same time. In order to formulate such a theory, it is important to test the concepts of self against at least some (atypical) empirical examples of self-experiences, since these 'test' cases enable a more detailed evaluation of the concepts. A theory of self should be fine-grained enough to grasp self in all of its varieties: if a theory does not accomplish this, it should be developed further in order to provide an exhaustive account of the whole phenomenon. Thus, a theory of self that fails to embrace all forms of selfhood is weak: the concept of selfhood cannot be accurate enough if it cannot be applied to cases that deviate from the exemplar. Instead, a theory or conception of self that can also account for rare cases has more strength: its explanatory power is widened and the reasons to endorse it obtain support. Thus, empirical constraints are relevant for philosophers of mind. On the other hand, the conceptual analysis and theoretical knowledge from philosophy can contribute to the development of empirical theories and paradigms.

One promising methodological invention in the research on the self is to study it through different altered states of consciousness. The idea is to provide an analysis that uses an *altered state of consciousness* (ASC in brief) as a contrast condition that can elicit the features of normal self-consciousness. In other words, ASCs can be seen as a methodological tool that assists in sorting out the aspects and functions of self-consciousness. An ASC can be defined as "a temporary change in the overall pattern of subjective experience" (Farthing 1992, 205), and the strategy of examining ASCs seems highly relevant for detecting the layers of self-consciousness and the dynamics of the aspects of self.[7] The contrasts between altered and normal experience can reveal the tacit features of self that we do not normally pay attention to: only when these features change or are absent, is it possible to understand what they originally were. Thus, a profile of an ASC

[7] For a definition of an ASC see e.g. Revonsuo et al. 2009. ASCs have been useful in the examination of self; the wide range of these ASCs include: (i) meditative practices (Thompson 2015), (ii) experiences under psychedelic drugs (Carhart-Harris et al. 2012), (iii) induced illusions (Blanke & Metzinger 2009) and (iv) pathological conditions such as schizophrenia (Sass & Parnas 2003) and Depersonalisation Disorder (Ciaunica et al. 2021).

may disclose the intricacy of self-consciousness better than the normal experience.

An ASC that is interesting for the study of self is dreaming. Philosophers have argued that dreams can be used as an instrument that leads to a deeper understanding of consciousness, self-consciousness, and subjectivity (e.g. Metzinger 2013, Thompson 2015; Windt 2015). This paper follows this argumentation line and proposes that the study of dreaming can assist in dissociating different aspects or layers of self-consciousness and thus, make decisive contributions to the philosophical project of defining the concepts by which the richness of self-consciousness can be grasped.[8] The following analysis focuses on self-consciousness since it concerns the consciousness of self in dream experiences. In addition, self-consciousness provides a good general starting point for the study of self; in order to answer metaphysical questions concerning the nature of self, we need to know what the self is assumed to be, and in order to establish this, we should investigate self-experiences in self-consciousness (see e.g. Strawson 2000).

Generally, *dreaming* refers to subjective experiences during sleep.[9] Dreaming is a fully "inner" or "offline"[10] experience in

---

[8] In addition, this philosophical project to conceptually describe the layers of human self-consciousness is significant for multidisciplinary fields since it can give proper explananda for empirical research programs and assist in developing empirical theories (see e.g. Metzinger 2013; Windt 2015).

[9] In more detail, dreaming has been defined in terms of simulation (Revonsuo 2005; 2006), hallucination (Windt 2010; 2015), and imagination (Thompson 2015). The claim that dreams really are conscious experiences is also indicated in experiments with lucid dreamers (see e.g. LaBerge et al. 1981; Windt 2015; Revonsuo 2015). However, it can be noted that empirical information on dreaming and research on dreaming is still incomplete (see e.g. Windt 2015).

[10] The conceptual distinction between online and offline is used in the discussions of embodied cognition. 'Online' refers to experience that involves actual coupling with the environment. Instead, 'offline' experiences are self-generated and independent of concurrent stimulation of the senses and thus, "disconnected" from the environment. In addition to dreaming, offline sensory experiences occur during mental imagery, mind-wandering and hallucinations (Fazekas et al. 2021).

the sense that it occurs without a sensory or motor encounter with the environment but is generated by brain-activity while the body is at rest. Considering dream experiences is relevant for the study of selfhood since dream experiences involve alterations in the organization of a pattern of self and thus, different aspects of self can be more easily prominent in dreams than in waking consciousness. Further, some kind of *dream self* is present in the great majority of dreams (see e.g. Revonsuo 2005; Thompson 2015; Windt 2015). Roughly, a 'dream self' is the protagonist of the dream with whom the dreamer identifies herself. The core feature of dreaming is the immersive experience of being a self in the world, which also denotes the waking state.[11] Many times the dream self resembles the waking self, for instance, has the same kind of body and memories, although not necessarily. Despite the resemblances, typically the dream self differs from the waking self at least in its (meta)cognitive skills; the central characteristics of the dream self is a lack of the full mental capabilities of the waking self.

Because dream self and waking self differ from each other, one needs to be cautious about drawing a too straightforward relation between the dream and waking self or too simple conclusions about the complexity of selfhood. The methodological idea here is not to consider the dream self as a conclusion to philosophical questions on the nature or constitution of the self. Instead, the idea is that a careful analysis of the features of the dream self can provide premises for the arguments about the nature of self and relationships within the aspects of self (i.e. this is an application of the general methodology of neurophilosophy, see e.g. Mandik 2007).

3. The experiential self in dreams

Although the dream self can be strange, it remains as the subject of the dream experience and dreams are subjective experiences. Thus, at least the experiential self is present in dreams. In point of fact, dream research seems to be especial-

[11] But diminishes in the hypnagogic state between wake and dreaming, see e.g. Thompson 2015; Windt 2015. In addition, it is interesting that there are dreams that involve a double representation of self (e.g. Occhionero et al. 2005; Revonsuo 2005; Thompson 2015; Windt 2015).

ly relevant for the examination of the experiential self. In typical waking consciousness, minimal self-consciousness is a tacit feature that involves several aspects of self and is intertwined with the contents of experience, which makes it difficult to grasp. However, in dreams the experiential self can take less complex forms which may disclose its components more easily.

The study of dreaming is especially useful in solving a specific question about the necessary features of self. It is important to define the necessary, most minimal and fundamental forms of self since it is the starting point for a conceptually systematic account of selfhood (as e.g. Metzinger 2013; Windt 2010, 2015; Zahavi 2014 argue). Below I approach the problem of defining the minimum required for self-consciousness in terms of the pattern theory of self, and ask whether some of the aspects of self are necessary. I examine the dream self and elaborate on how the aspects of self can be omitted from it, proceeding from the cognitively higher layers of reflective self to the cognitively lower layers of experiential self. A lack of an aspect or feature of self reveals that the feature in question is unnecessary for self-consciousness. The aspect that can be found even in the cut off forms of self-consciousness has a special status in the pattern of self, since it is the most fundamental form of self-consciousness that is also the basis for other forms of selfhood.

First of all, a characteristic feature of the dream self is an unstable and disintegrated self-reflection. The dream self typically suffers from a lack of rationality and deliberation, and acts in incoherent and potentially morally dubious way. Instead of being an effective metacognitive subject of experience, the dream self has difficulties in conceptualizing and experiencing herself as a thinking, attending or deciding subject (i.e. the dream self has only weak cognitive, attentional and volitional 1PP, Windt & Metzinger 2007; Windt 2015). Further, the dream self has deficiencies of both short- and long-term memory and rather is amnestic; it does not have full access to the waking self's memories and instead can confabulate narratives. However, the dream self is not bothered by the discrepancies in its surroundings and own actions. This can be illustrated by a dream report of Evan Thompson (Thompson 2015, 136):

> I'm on the subway in Toronto. The train is above street level, and I see Paris streets below me through the window. I'm with a former girlfriend from many years ago. I'm anxious waiting for the stop, where I know I have to get off. Then the stop is past and she's out in the street. I'm more anxious and look for my suitcases. One is missing. Maybe she took it, but the train's moved on and she's gone. I wake up feeling anxious and thinking I need to find my suitcase.

The report involves a number of discontinuities, all of which the dream self fails to pay attention to. For instance, the dream self is at the same time in a subway and above street level, and in two cities. The dream self is traveling with a friend, whom the waking self has not seen in years. The dream self is waiting for the next stop, but then it is already past. The dream self remembers some important suitcases, although had not thought about them before.

Thus, it is clear that in dreams one has a sense of self but lacks the typical waking reflective self-consciousness. Many times the dream self is unable to think critically and exercise self-deliberation. In addition, the self-narrative of the dream self is often discontinuous and fragmentary. This indicates that psychological-cognitive and narrative aspects can be severely diminished in dream experience and thus, they are not necessary for self-experience. The self can be experiential without being reflective.

Further, the experiential self assumes altered forms in dreams. Thus, the dream self provides an opportunity to elaborate on the structures of experiential self, which involves the experiential, embodied and affective aspects. The dream self can involve alterations in all these aspects. The experiential aspect of the waking self many times involves the sense of agency that is "The sense that I am the one who is causing or generating an action" (Gallagher 2000, 15). This sense of agency can be missing in dreams in which the dream self remains as a mere passive observer without active participation in the dream events. This kind of dream experience shows that the sense of agency is not a necessary feature of self.[12]

---

[12] This possibility is recognized in the pattern theory (Gallagher 2013). In point of fact, it was one reason to initially draw the distinction between the two features of the minimal self (or the experiential aspect), i.e. a sense

However, maybe a more interesting feature of the dream self is that it does not lose the experiential aspect of self altogether. The experiential aspect is present as the subjectivity of consciousness or as a first-person perspective; even if the perspective is unstable and the self-experience altered, the dream self does have a perspective and undergo experience. In other words, the experiential aspect includes both the sense of agency and sense of subjectivity (that is also called 'sense of ownership', Gallagher 2000). Although a sense of agency can be lacking in the experience, the subjectivity does not disappear even in highly altered dream experiences.

With regard to the affective aspect of self, dream research indicates that the dream self cannot be considered a fully affective subject that commands a variety of emotions in the same way as the waking self (e.g. Thompson 2015; Windt & Metzinger 2007). Very often the dream self does undergo affects, and dreams can involve especially strong emotional experiences.[13] For instance, nightmares are characterized by such intense feelings that the dreaming self is woken up by them. However, the variety of affects experienced by a dream self is typically much simpler than the affects experienced by the waking self. For instance, a dream can be dominated by a single feeling, such as anxiety as in the dream report above (e.g. Thompson 2015; Windt & Metzinger 2007). In addition, some dreams can lack affectivity altogether and instead are characterized by a neutral observation of a dream scene. Since affectivity can be lacking in a dream experience, this indicates that the affective aspects are not necessary for self-consciousness.

In a similar way, the embodied aspects of self can diminish in dreams. Dreaming has been described to be phenomenally embodied only in a weak sense (Windt 2015, 339). This weak embodiment is predominantly associated with movement sensations of individual body parts. In addition, the dream

---

of agency and a sense of ownership. According to Gallagher's (2000) original idea, the sense of ownership can remain even in ASCs that lack the sense of agency.

[13] The great majority of dreams involve affects in self-rated questionaries, however, the number of affects is presented as being smaller when the affectivity in dream reports are rated by external raters. For dream affects, see Sikka 2020.

self has disturbances in multisensory integration, for instance a body part may be seen but not felt or vice versa. Moreover, the dream self only rarely has sensory experiences of pain, temperature, smell, or taste. However, the most striking example of deficiencies of embodied aspect is the dream experience in which the dream self does not have a body at all.

Metzinger (2013) and Windt (2010; 2015) have used the phenomenon of bodiless dreams as an example of a minimal phenomenal selfhood (MPS in brief). MPS refers to the simplest form of selfhood and as such the strictly necessary features of self and consciousness. According to Metzinger (2013), bodiless dreams are the best global contrast condition for isolating MPS.[14] Bodiless dreams are a rare, but well-known phenomenon in which a dreamer identifies with an extensionless point in perceptual space. Metzinger explains that in these cases the dream self has an "abstract self-representation", which does not contain any perceptual or spatially extended features of bodily content. This experience of bodiless subjectivity involves a stable sense of selfhood and an "asomatic 1PP", although the body representation is absolutely minimal. According to Metzinger (2013) and Windt (2010; 2015), bodiless dreams can reveal MPS, which they define as a "transparent self-location in a spatiotemporal frame of reference" (Metzinger 2013, 7). Since this self-location, or 1PP, only includes a point in space and a point in time, it also encompasses a highly atypical dream experience.

However, the notion of an experiential self implies that minimal selfhood should not be only defined in terms of spatio-temporal location or geometrical perspective. A robot equipped with a camera might also be said to have a geometrical perspective and locate itself in a functional sense, although it does not experience anything. Instead, the crucial feature of being a self is that 1PP is experienced subjectively; it has a subjective character that can be associated with the experiential aspects of self. Whereas, in terms of the pattern

[14] According to Metzinger (2013), in addition to dreams, there are two other experiences of bodiless subjectivity: out-of-body experiences (OBEs in brief) and meditation. However, Metzinger notices that both asomatic OBEs and "pure consciousness" experiences in meditators are rare phenomena and thus, more difficult ways to investigate MPS. See also fn. 7.

theory of the self, the geometrical perspective of the subject might be considered as an embodied aspect. Thus, the dream self can lack embodied aspects of self to the extent of lacking a representation of a body. In other words, these are not necessary for self-consciousness, since there can be self-experience without consciousness of a body at all.

Overall, these lessons from dream research make a significant contribution to the (pattern) theory of self by dissociating different layers of self-consciousness, and revealing the most fundamental aspect of self. The above examination of the dream self showed that self-experience can lack psychological-cognitive, narrative, affective and embodied aspects of self, and the sense of agency. However, the shared feature across dream experience is the presence of the experiential aspect. This indicates that the experiential aspect of self is the most fundamental level of self-consciousness: it can occur in the absence of other features of self-consciousness but not the vice versa. The experiential aspect is present in all dreams regardless of the combination of the other aspects. That is, the experiential aspect is necessary in a way that other aspects are not.

3.1. Theoretical implications of the necessity of the experiential aspect

The necessary status of the experiential aspect strengthens the idea that the experiential self is the most fundamental form of selfhood, and that it minimally involves only a subjective first-person perspective. This undermines theories of self that deny the fundamental character of the experiential aspects or claim that some other feature(s) of self is equally necessary. These theories include at least those theories that consider self as strictly narrative or reflective, claim that a representation of the body is necessary for self-consciousness, or make a too strong claim about the self's sensory-motor coupling with the environment. The shortcomings of these kinds of theories are briefly elaborated on below.

First, the manifestations of the dream self question the theories of self which claim that narratives are a constitutive necessity for being a self. For instance, according to Schechtman's (2011) Narrative Self-Constitution View (NSCV

in brief), we constitute ourselves by understanding our lives in narrative form. The narrative structure of selfhood does not require explicit narratives, but the idea is that we experience and interpret our present experience as a part of continuous narrative that gives meanings to events and experiences. The NSCV (Schechtman 2011, 405) places two constraints on self-constituting narratives: 1) the articulation constraint "involves the capacity to articulate one's narrative locally where appropriate", and the 2) reality constraint, which "demands that our narratives fit with the basic conception of reality shared by those in our community" (it probably cannot e.g. involve being able to get from Helsinki to Tokyo in one minute). Although a dream self many times participates in events that can be described with narratives and dream reports can have narrative structure, the above-mentioned two constraints are too strong. The possible narrative that a dream self would articulate would contradict the logic of the waking self's narrative and also the reality of the dream world does not meet the reality constraint in the waking world. This does not entirely refute the theories that emphasize cognitive-psychological and narrative aspects of self, but highlights the point that these aspects are not the most fundamental form of selfhood, and that the concept of a reflective self should be complemented by the concept of the experiential self (see e.g. Zahavi 2014). However, the experience of a dream self does refute the theories which claim that self-consciousness is necessarily reflective and does not recognize the significance of the experiential self (e.g. Carruthers 1996). The dream self can have vivid experiences without coherent self-reflection and does not even seem to question the lack of a continuous narrative.

Second, the theories that consider body-representation as constitutive of self-consciousness can be criticized in the light of dream experience. For instance, Blanke and Metzinger (2009) presented a theory of MPS in terms of three central defining features: 1) a globalized form of identification with the body as a whole, 2) spatiotemporal self-location, and 3) a 1PP (in the weak sense of a purely geometrical feature of perception, targeted in empirical studies investigating visuospatial perspective-taking). However, as Windt (2010, and Metzinger 2013 agrees) argues, bodiless dreams show

that this minimal form of self-experience or MPS does not require "a passive, multisensory and globalized experience of 'owning' a body" as Blanke & Metzinger (2009) present. In addition, Windt (2010; 2015) argues that the distinction between a sense of spatiotemporal self-location and a spatiotemporal 1PP is unnecessary; the subjective sense of presence involves only the sense of immersion or location in a spatiotemporal frame of reference. Thus, dream research is useful in elaborating the notion of minimal selfhood and abandoning too complex formulations.

Third, the dream self discounts theories of strong sensorimotor enactivism which claim that interaction with the environment is necessary for self-consciousness. According to sensorimotor enactivism, experiences are constituted by sensory and motor couplings with the environment.[15] In the strong version of sensorimotor enactivism, this interaction is claimed to be a necessary feature of consciousness, and this kind of general theory of consciousness can be criticized by using the dream argument (Revonsuo 2015; Loorits 2017). The dream argument points out that dream experiences are as rich and complex as the waking experience (or sufficiently similar to waking experience), and fully internally constituted. The implication of this is that necessary constitutive conditions for experiential states can be constituted only internally and thus do not require a relationship with the environment. The proponents of strong sensorimotor enactivism can answer this argument by denying that veridical and dream experiences share the same phenomenological status: It is irrelevant how dreams are constituted since they

[15] Enactivism (originating from Varela et al. 1991, for different version of enactivism, see e.g. Ward et al. 2017) is a relatively novel approach in the philosophy of mind and cognitive sciences, which proposes that cognition is a form of embodied action, and that "the human mind is embodied in our entire organism and embedded in the world, and hence is not reducible to structures inside the head" (Colombetti & Thompson 2008). Enactivism emphasizes that consciousness is central to the understanding of a cognitive system, and the concept of experiential self is significantly linked to consciousness. Thus, the enactivist theory of the nature of the conscious cognitive system can roughly be considered as a theory of self. For an enactivist view of self that also utilizes dream research, see Thompson 2015.

are not real experiences. For instance, Noë (2009, 179-180) uses this strategy and argues that: "[D]ream seeing is not really seeing at all. […] [W]e ought to think of perceiving as an activity of exploring the environment." However, this answer is rather unsuccessful since the claim that dream experiences are not real experiences is highly unintuitive and contradicts the dream research presented above. Thus, the dream argument shows that an online active interaction with the environment is not necessary for self-consciousness.[16]

4. The reflective self in dreams

As the above characterizations have shown, the dream self typically has a defective self-reflection and –narrative (Revonsuo 2005; Thompson 2015; Windt 2015; Windt & Metzinger 2007). Often the dream self does not succeed in critical thinking, makes mistakes in reasoning and acts irrationally. Even if the dream self would resemble the waking self, it suffers from difficulties in directing attention, thinking and decision making. Thus, at first sight, it seems that the study of dreaming cannot be as relevant for the examination of the reflective self as it was for understanding the experiential self and its necessary features. By contrast, the deficiencies in the reasoning of the dream self make one question as to whether it could it be considered equal with the waking reflective self at all (this seems to be the idea e.g. in Descartes' dream arguments which function as skeptical arguments, see. e.g. Windt 2015).

However, there is an interesting exceptional case of dreaming – lucid dreaming – which highlights the reflective capacities of self. In a lucid dream, a dreamer knows that she is

[16] Instead, weak versions of sensorimotor enactivism can offer more efficient strategies to answer the dream argument. The weak versions do not require an active online relationship with the environment in order to explain experiences, but propose that knowledge of the sensorimotor interaction (i.e. sensorimotor contingencies) is enough for constituting experiences (Telakivi 2020). In terms of self, it might be argued that the deficiencies in self-consciousness of the dream self actually support the enactivist idea that rich self-experiences involve interaction with the environment. However, an elaboration of the enactivist conception of self is not within the scope of this paper.

dreaming (Metzinger 2009; Noreika et al. 2010; Thompson 2015). A stronger definition of lucid dreaming (or definition of full lucidity) also highlights the following features.[17] 1) Cognitive insight and overall mental clarity that is at least as high as during normal waking states. 2) Agency is fully realized, involving the control of attention and behavior in the dream events; the dream self can do whatever she wants to – walk through walls, fly, or engage in conversations with dream figures. 3) The autobiographical memory is intact, involving full access to past waking life as well as in previous dreams. 4) The dream self's all five senses function as well as in a waking state. 5) Lucid dreaming involves more positive emotions and emotional control than non-lucid dreaming. Altogether, the experienced quality of cognition and agency is especially high in lucid dreaming; the dreamer has sharp self-reflection and perceives the environment even more intensively than when awake. Here is a famous example of a lucid dream report that does not contain all features of the strong definition of a lucid dream but presents the cognitive insight and quality of lucid dream experience (Fox, 1962, 32; quoted in Thompson 2015, 152):

> [...] Then the solution flashed upon me: though this glorious summer morning seemed as real as real could be, I was dreaming! With the realization of this fact, the quality of the dream changed in a manner very difficult to convey to one who has not had the experience. Instantly, the vividness of life increased a hundred-fold. Never had the sea and sky and trees shone with such glamorous beauty; even the commonplace houses seemed alive and mystically beautiful. Never had I felt so absolutely well, so clear-brained, so inexpressibly free! The sensation was exquisite beyond words; but it lasted only a few minutes and I awoke.

Lucid dreaming is especially interesting as a means of unfolding the layers of self, since it involves a profound change in self-experience. As Thompson (2015, 140) points out, lucid

---

[17] The list here is combined from Metzinger 2009; Noreika et al. 2010; Thompson 2015; Voss et al. 2013; Windt 2015; Windt & Metzinger 2007. For the levels of lucidity, see e.g. Noreika et al. 2010; Thompson 2015; Windt 2015.

dreaming involves two modes of self-experience. In a nonlucid dream, a dreamer identifies with the dream self and can think, for instance, that "I am flying". By contrast, in a lucid dream, the sense of self shifts when the dreamer recognizes that "I am dreaming" and the dream self is only an avatar of the dreaming self. That is, lucid dreaming involves two kinds of self-awareness; one is aware of one's self both as the dream self ("I as dreamed") and the dreaming self ("I as dreamer"). Lucid dreaming therefore involves clear and distinct introspection, the insight of the illusory character of the dream self, and the realization of different modes of self-consciousness. These insights are conjoined with an ability to control the contents of the dream and guide the dream self, and offer an interesting perspective on the self.

Thus, lucid dreaming provides an opportunity to also approach the reflective self and enable further analysis of the functions and structure of self. For instance, it would be interesting to study the transition from nonlucid dreaming to lucid dreaming more closely.[18] As the above descriptions indicate, the characteristics of the sense of self in these two types of dreaming are opposite in many ways. While nonlucid dreaming involves only a highly unstable 1PP and confused thinking, lucid dreaming is related to a stable first-person perspective and cognitive insights. In terms of the pattern theory of self, nonlucid dreaming seems to involve a rather disintegrated and partial pattern, whereas lucid dreams seem to display an integrated pattern in which the aspects are linked together. Thus, tracking the proceeding from a nonlucid to lucid dream could reveal how the layers of self unite or the aspects of self become connected. That is, it is possible that dream research can offer finding about the integration of the aspects of self, not only about their dissociation. The transition from nonlucid to lucid dreaming is also of multidisciplinary interest; experiments that trace the changes in neural activation in the transition could assist in understanding the underpinnings of waking self. On the other hand, the precise conceptualization of different features of

[18] Another useful strategy to employ lucid dreaming in the study of self is to compare self-consciousness in lucid dreaming and other ASCs, see e.g. Noreika et al 2010; Thompson 2015; Windt 2015.

self is of utmost importance in the analysis and interpretation of the experiments and thus, co-operation between philosophers and scientist is encouraged and can be mutually beneficial.

An applicable but more complex future research object could involve the well-being of self. Lucid dreaming is characterized not only by cognitive insight but also by positive emotions, and well-being-oriented studies could benefit from an examination of the pattern of self in lucid dreaming. Based on the comparison between self-consciousness in nonlucid and lucid dreaming, it seems that the less integrated nonlucid dreaming involves fewer positive feelings, or at least the feelings experienced are less controlled. In contrast, the integrated and insightful lucid dreaming involves more positive feelings (Noreika et al. 2010; Voss et al. 2013; Windt 2015). Thus, it seems that a balance and integration within the aspects of self can lead to positive emotions (although lucid dreaming can also involve experiences of dissociation, see e.g. Voss et al. 2013). A better understanding of this integration and the interconnections of the aspects of self could also be used in interventions targeting increased positive affects or well-being of the waking self. However, more studies are needed in order to take full advantage of the phenomenon of lucid dreaming in the study of self.

5. Summary

Selfhood is a multifaceted phenomenon and in need of elaboration. Dream research offers a useful tool for the study of self since the aspects of self are organized differently in dreaming than in typical waking self-consciousness. This opens a novel vantage point from which to observe the structures of self. In this paper, I examined the dream self in terms of the general conceptual distinction between experiential and reflective self, both of which involve several more detailed aspects of self. The main focus of the paper was on the experiential self and its manifestations in dreams. It transpired that a dream experience can basically lack all aspects of self except the experiential aspect, and this was used as an argument for the necessary status of the experiential aspect in the pattern theory of self. That is to say, the empirical evi-

dence from dream research strengthens the concept of the subjective first-person perspective as the minimal or fundamental feature of selfhood. This significance of the experiential aspect undermines theories of self which claim some contingent features of self to be necessary. These features involve coherent self-reflection and -narrative, representation of a body, and online interaction with the environment. Concerning the reflective self, the quality of the rationality and reflection of a dream self varies. Typically, the dream self is characterized by deficiencies in thinking. However, lucid dreaming is an interesting exceptional case of self-reflection, characterized by a stable and integrated first-person perspective and specific cognitive clarity and control. Because of the high quality of self-reflection and integration of aspects of self, it could be useful to target lucid dreaming in more detail in future investigations of self. Overall, dreaming provides an interesting instrument with which to study the self. Although the dream self is not exactly the same as the waking self, it provides a means to learn more about the dimensions of self.[19]

University of Turku

References

Albahari, M. (2006), *Analytical Buddhism: The Two-tiered Illusion of Self*, Palgrave Macmillan, London.

Barrett, D. and McNamara, P. (eds.) (2007), *The new science of dreaming: Vol. 3. Cultural and theoretical perspectives*, Praeger Perspectives, Westport.

Bermudez, J. L. (2001), "Nonconceptual Self-Consciousness And Cognitive Science", Synthese 129 (1), pp. 129-149.

Blanke, O. and Metzinger, T. (2009), "Full-body illusions and minimal phenomenal selfhood", *Trends in Cognitive Sciences* 13 (1), pp. 7-13.

Carhart-Harris, R. L., Erritzoe, D.,Williams, T., Stone, J. M., Reed, L. J., Colasanti, A., et al. (2012), "Neural correlates of the psychedelic state

[19] I am grateful to Valtteri Arstila, Antti Revonsuo, Nils Sandman, and two anonymous referees for their helpful comments. In addition, I thank Otto A. Malm Foundation and Finnish Cultural Foundation for financial support to this study.

as determined by fMRI studies with psilocybin", *Proc. Natl. Acad. Sci. U.S.A.* 109, pp. 2138–2143.

Carruthers, P. (1996), *Language, Thoughts and Consciousness. An Essay in Philosophical Psychology*, Cambridge University Press, Cambridge.

Ciaunica, A., Charlton, J. and Farmer, H. (2021), "When the Window Cracks: Transparency and the Fractured Self in Depersonalisation", *Phenomenology and the Cognitive Sciences* 20 (1), pp. 1-19.

Colombetti, G. and Thompson, E. (2008), "The Feeling Body: Toward an Enactive Approach to Emotion", in W. Overton, U. Muller and J. Newman (2008), pp. 45–68.

Farthing, G. W. (1992), *The psychology of consciousness*. Prentice-Hall, Inc.

Fazekas, P., Nanay, B. and Pearson, J. (2021), "Offline perception: an introduction", *Philosophical Transactions of the Royal Society B*, 376 (1817).

Feinberg, T.E. and Keenan, J.P. (eds.) (2005), *The Lost Self: Pathologies of the Brain and Identity*, Oxford University Press, New York.

Gallagher, S. (2000), "Philosophical conceptions of the self: Implications for cognitive science", *Trends in Cognitive Sciences* 4(1), pp. 14–21.

Gallagher, S. (2013), "A pattern theory of self", *Frontiers in Human Neuroscience* 7, pp. 1-7.

Gallagher, S. (ed.) (2011), *The Oxford Handbook of the Self*, Oxford University Press.

Gallagher, S. and Daly, A. (2018), "Dynamical Relations in the Self-Pattern", *Frontiers in Psychology* 9, pp. 1-13.

Gallagher, S. and Zahavi, D. (2008), *The Phenomenological Mind: An Introduction to Philosophy of Mind and Cognitive Science*, Routledge, New York.

Gertler, B. (2011), *Self-Knowledge*, Routledge, London.

Kriegel, U. (2004), "Consciousness and Self-consciousness", *The Monist* 87, pp. 185-209.

Korsgaard, C. M. (2009), *Self-Constitution: Agency, Identity, and Integrity*, Oxford University Press.

La Berge, S. P., Nagel, L. E, Dement, W. C., and Zarcone, V. P. (1981), "Lucid dreaming verified by volitional communication during REM sleep", *Perceptual and Motor Skills* 52(3), pp. 727–732.

Loorits, K. (2017), "Dreaming about Perceiving: A Challenge for Sensorimotor Enactivism", *Journal of Consciousness Studies* 24 (7–8), pp. 106–129.

Mandik, P. (2007), "The neurophilosophy of Consciousness', in M. Velmans and S. Schneider (2007), pp. 458-471.

Metzinger, T. (2009), *The Ego Tunnel: The Science of the Mind and the Myth of the Self*, Basic Books, NewYork.

Metzinger, T. (2013), "Why are dreams interesting for philosophers? The example of minimal phenomenal selfhood, plus an agenda for future research", *Frontiers in Psychology* 4, pp. 1-17.

Moran, R. (2001), *Authority and Estrangement: An Essay on Self-Knowledge*. Princeton University Press.

Newen, A. (2018), "The Embodied Self, the Pattern Theory of Self, and the Predictive Mind", *Frontiers in Psychology* 9, pp. 1-14.

Noë, A. (2009), *Out of Our Heads*, Hill & Wang, New York.

Noreika, V., Windt, J. M., Lenggenhager, B. and Karim, A. A. (2010), "New perspectives for the study of lucid dreaming: From brain stimulation to philosophical theories of self-consciousness", *International Journal of Dream Research* 3 (1), pp. 36-45.

Occhionero, M., Cicogna, P., Natale, V., Esposito, M. J. and Bosinelli, M. (2005), "Representation of self in SWS and REM dreams", *Sleep and Hypnosis* 7(2), pp. 77–83.

Overton W., Muller U. and Newman J. (eds.) (2008), *Developmental Perspectives on Embodiment and Consciousness*, Lawrence Erlbaum, New York.

Perry, J. (1979), "The problem of the essential indexical", *Nous* 13, pp. 3–21.

Pylkkö, P. (1998), *The Aconceptual Mind*, John Benjamins, Amsterdam.

Revonsuo, A. (2005), "Dream Self", in T.E. Feinberg & J.P. Keenan (2005), pp. 206-219.

Revonsuo, A. (2006*), Inner Presence: Consciousness as a Biological Phenomenon*, MIT Press.

Revonsuo, A. (2015), "Hard to See the Problem?", *Journal of Consciousness Studies* 22 (3–4), pp. 52–67.

Revonsuo, A., Kallio, S. and Sikka, P. (2009), "What is an altered state of consciousness?", *Philosophical Psychology* 22, pp. 187–204.

Sass, L.A., and Parnas, J. (2003), "Schizophrenia, consciousness, and the self", *Schizophrenia Bulletin* 29 (3), pp. 427–444.

Schechtman, M. (2011), "The narrative self," in S. Gallagher (2011), pp. 394–416.

Shoemaker, S. (1968), "Self-Reference and Self-Awareness", *The Journal of Philosophy* 65, pp. 555–67.

Shoemaker, S. (2011), "On what we are," in S. Gallagher (2011), pp.352–371.

Siderits, M., Thompson, E. and Zahavi, D. (eds.) (2011), *Self, No Self? Perspectives from Analytical, Phenomenological, and Indian Traditions*, Oxford University Press, Oxford.

Sikka, P. (2020), *Dream affect: conceptual and methodological issues in the study of emotions and moods experienced in dreams,* University of Turku, Turku.

Strawson, G. (2000), "The Phenomenology and Ontology of the Self", in D. Zahavi (2000), pp. 39-54.

Telakivi, P. (2020), *Extending the Extended Mind: From Cognition to Consciousness*, University of Helsinki, Helsinki.

Thompson, E. (2015), *Waking, dreaming, being: self and consciousness in neuroscience, meditation, and philosophy*, Columbia University Press, New York.

Varela, F., Thompson, E. and Rosch, E. (1991), *The Embodied Mind: Cognitive Science and Human Experience*, The MIT Press, Cambridge, MA.

Velmans, M. and Schneider, S. (eds.) (2007), *The Blackwell Companion to Consciousness*, Blackwell Publishing.

Voss, U., Schermelleh-Engel, K., Windt, J., Frenzel, C. and Hobson, A. (2013), "Measuring consciousness in dreams: The lucidity and consciousness in dreams scale", *Consciousness and Cognition* 22 (1), pp. 8-21.

Ward, D., Silverman, D., and Villalobos, M. (2017), "Introduction: The Varieties of Enactivism", *Topoi* 36, pp. 365–375.

Windt, J. M. (2010), "The immersive spatiotemporal hallucination model of dreaming", *Phenomenology and the Cognitive Sciences* 9 (2), pp. 295-316.

Windt, J. M. (2015), *Dreaming: A Conceptual Framework for Philosophy of Mind and Empirical Research*, MIT Press, Cambridge, MA.

Windt, J. M. and Metzinger, T. (2007), "The philosophy of dreaming and self-consciousness: What happens to the experiential subject during the dream state?" in D. Barrett & P. McNamara (2007), pp. 193-248.

Zahavi, D. (2005), *Subjectivity and Selfhood: Investigating the First-Person Perspective,* MIT Press, Cambridge, MA.

Zahavi, D. (2011), "The Experiential Self: Objections and Clarifications', in M. Siderits, E. Thompson and D. Zahavi (2011), pp. 56-78.

Zahavi, D. (2014), *Self and Other: Exploring Subjectivity, Empathy, and Shame*, Oxford University Press.

Zahavi, D. (ed.) (2000), *Exploring the Self*, John Benjamins.

Without a Voice of One's Own: *Aphonia* as an Obstacle to Political Freedom

JOONAS S. MARTIKAINEN

Introduction

What does it mean to have a public "voice of one's own", either as an individual or as a group? What does it mean to lose that voice? We live in a time of a sharpening social divide between those with opportunities to participate in political life, and those who feel left behind by politics altogether, remaining passive. It seems that alongside increasing social and economic inequality, there is a growing divide between those who actively participate in the political life of the society around them, and those who perceive their ability to influence politics as non-existent.

In this article I make use of Maurice Merleau-Ponty's existential phenomenology to conceptualize and present a disclosing critique of a phenomenon which I have decided to call *aphonia*, or the loss of one's own public voice. This conception of *aphonia* offers a correction to theories of democracy which tacitly assume a citizen who is unproblematically willing and able to voice their opinion on matters concerning them. I investigate how public speech can be understood as an expressive modality of the lived body, which can be lost due to experiences of suffering economic and social marginalization. I claim that this makes it hard for already dominated groups to make their frustrations and concerns known in public. Entire social groups therefore can, through no fault of their own, become silenced.

I first discuss Maurice Merleau-Ponty's existential phenomenology as a tool for disclosing critique. I then present some examples from sociological literature, which describe

how economically marginalized and politically apathetic persons and groups can perceive their political marginalization as frustrating and constraining, while feeling unable to speak out in public about their marginalization. These examples present a phenomenon which appears widespread in Western democracies while remaining largely unconceptualized by mainstream political philosophy. I conceptualize such *aphonia* as a result of the internalization of experiences of social marginalization into what Merleau-Ponty calls the "operative intentionality" of the lived body. Merleau-Ponty's phenomenology helps me to show how this process also has a perceptual effect. As one slowly and unknowingly acquires the habit of remaining silent in public and withdrawing from public participation due to, for example, feelings of frustration, powerlessness, and shame, one may quite literally have the words with which to protest one's condition taken out of one's mouth. *Aphonia*, then, does not mean losing the human capacity of speech, but losing the ability to see oneself as a credible and able public speaker who is allowed to have a voice of one's own. I finish with a call for grass-roots initiatives for engendering political participation among those suffering from *aphonia*.

Phenomenology as a Method for Disclosing Critique

My chosen method situates this article alongside a recent wave of interest in "political phenomenology": a philosophical approach which draws from twentieth-century phenomenology. This approach is exemplified by the recent broad collection of articles edited by Thomas Bedorf and Steffen Hermann (2020a). Political phenomenology does not present a unified movement with either a well-defined methodological toolkit or a shared normative stance. It is instead an attempt to rethink fundamental themes of contemporary political philosophy through engagement with the tradition of phenomenology. Political phenomenology uses phenomenological description and diagnoses of particular experiences, as well as diagnoses of contemporary political phenomena, to investigate the commonplace abstractions of mainstream political philosophy. In this article I use existential phenomenology as a tool for what Nikolas Kompridis (2005) calls

"disclosing critique"; a way of revealing possibilities that have been left unnoticed in current social arrangements, and also a way to give voice to experiences of suffering that have heretofore been left unexamined.

Contemporary political philosophy and theory has approached political marginalization and apathy in various ways. Traditionally theories of participatory democracy have described political marginalization as a lack of social goods, for example, the opportunities and resources for effective political participation. Such a lack is a matter to be addressed by a more just distribution of resources to ensure equality of opportunities and the formulation of more inclusive democratic procedures. James Bohman (1997) has objected to this view, noting that what is at issue is not only the poverty of opportunities and resources, but also the lack of recognition as an equal and the inability to acquire the cognitive and communicative capabilities required for effective participation in public, or what he calls "equality of effective freedom". Another strand of political theory describes the phenomenon as political exclusion, or the exclusion from democratic participation in public deliberative processes on decisions that concern one's own interests (Allen 2005; Benhabib 2004; Young 2000).

What has been left unnoticed by all of these approaches is something that theories of participatory democracy tend take for granted: the motivation of citizens to become politically engaged even when they are formally included within democratic processes; and the possibilities, resources, and communicative and cognitive capabilities to participate in public deliberation. After all, who would not want to participate in the making of decisions that affect one's own interests when presented with the opportunity to do so? When political philosophy tacitly assumes a motivated and capable subject, withdrawing one's democratic participation becomes understood as the result of a knowingly made choice. When someone does not participate in political processes by, for example, making their concerns known in public deliberation, or even by casting their vote when given the chance, it is easy to describe them as having knowingly and wilfully delegated one's share of political power over to others.

Approaching the matter as a form of lacking resources or capabilities, or as a form of being left outside the political community, or as the result of a willingly made choice to not participate, does not describe an important facet of political passivity. The picture of political agency behind much of contemporary democratic theory does not consider the effect of social marginalization on one's ability to perceive oneself as a capable and credible political agent. Possessing political agency not only means having the communicative capabilities required for effective participation, but also involves a subjective, affective component: *feeling* included, feeling like an able and credible political agent who is allowed to participate in the public life of one's community. Losing this feeling is a form of marginalization that can be approached as being separate from poverty as the lack of resources and opportunities, or the lack of cognitive and communicative capabilities required for effective political participation.

What, then, is exactly meant by *aphonia*, and how is existential phenomenology to be used to approach the issue? My usage of the term follows Nikolas Kompridis (2008, 301–3), who uses it to describe the loss of a sense of having a "voice of one's own" in public. Kompridis, however, does not develop the term much further in the article, leaving me room to perform a disclosing critique that, through a phenomenological approach, aims to give voice to a heretofore unacknowledged injustice.

Something of the nature of *aphonia*, losing the ability to express oneself, can be apprehended from a clinical example from Maurice Merleau-Ponty. In his *Phenomenology of Perception*, Merleau-Ponty (2012, 164 ff.) describes how a teenager, due to a tense family situation, suddenly starts suffering from muteness or *aphonia*. The inability to speak is clearly connected to the emotionally loaded conflict between family members, and as the situation is resolved, the *aphonia* disappears as well.

What catches Merleau-Ponty's attention is the way that the *aphonia* takes over the person in a fashion that is independent of their conscious will. The person suffering from *aphonia* does not, even on a deeper unconscious level, "choose" to remain silent. *Aphonia* is not a case of a voluntary limitation of one's freedom, since "To have lost one's voice is not to

keep quiet: one only keeps quiet when one can speak," (Merleau-Ponty 2012, 164). What has taken place is an impoverishment of a form of intentional experience that is "prior to both knowledge and ignorance, and prior to voluntary assertion and negation." (Merleau-Ponty 2012, 165) What has taken place is the disappearance of an expressive modality of one's being, a disappearance of the possibility to speak from experience:

> The will presupposes a field of possible among which I choose: here is Pierre, I can choose to speak to him or not. If, however, I lose my power of speech, then Pierre no longer exists for me as a desired or rejected interlocutor. The entire field of possibilities collapses, and I even cut myself off from the mode of communication and signification that is silence. (Merleau-Ponty 2012, 165)

I claim that such a "collapse of the field of possibilities" can, to a less radical degree, also happen to one's ability to speak out in public. Kompridis (2008, 300–2) describes the way suffering social hardships can also lead to political *aphonia* as suffering not just from the lack of words to put one's suffering in, but a "voice of one's own", the very ability to even attempt to articulate one's suffering politically in the first place. What is missing in such situations is something often tacitly taken for granted by philosophy of democracy: the ability to speak and express oneself authentically in public, and with it, the desire and motivation to engage with politics. I propose that in cases of *aphonia*, what has disappeared is a sense of being able to perceive speaking and acting politically as a meaningful possibility in one's intentional experience. I will next present some empirical examples of ways that internalization of negative social experiences can have an effect on the ability of an individual or a group to openly express themselves in public.

Examples from Sociological Literature

The usual approach of theories of justice, based on the scientific appraisal of social phenomena and comparing them to a rationally derived set of "objectively" valid philosophical norms, is badly suited for identifying the kinds of injustices that only signal themselves as *absences* of something. I be-

lieve that a phenomenon like *aphonia* is best approached by beginning from an investigation of particular examples that present different facets of an experienced injustice. Through a phenomenological diagnosis of such examples, a picture emerges: in this case an inability to speak in a voice of one's own in public. This is qualitatively different from an objectively measurable lack of participational resources and opportunities, or cognitive and communicative capabilities required for effective deliberative participation (cf. Bohman 1997).

I begin from a description by Finnish sociologist Eeva Luhtakallio and journalist Maria Mustranta of the frustration felt by the residents of a disadvantaged neighbourhood in Helsinki:

> During her years of fieldwork Eeva became more and more bothered by the observation that among the residents of the area the primary feeling associated with belonging to a society was frustration. Getting to know the residents made quickly clear that people were not – of course not – stupid or inactive, far from it. But many seemed to lack an understanding of what could be done about frustrating things, and the faith in the capacity of one's own actions to change things.[1]

Luhtakallio and Mustranta then proceed to relate a description of a social milieu whose inhabitants, despite their awareness of their situation, and definitely despite not being "stupid and inactive", remain unmotivated to become politically engaged and to challenge their political exclusion. This is due to lacking both the understanding of how to change things, and the faith, or confidence, in their ability to enact that change. This loss of confidence in oneself is described by Luhtakallio and Mustranta as the internalized result of a stream of negative social experiences, often in the hands of

---

[1] "Vuosien kenttätöiden aikana Eevaa alkoi yhä enemmän vaivata havainto siitä, että alueen asukkaiden päällimmäisin yhteiskuntaan kuulumiseen liittyvä tunne tuntui olevan turhautuminen. Asukkaisiin tutustuminen teki nopeasti selväksi sen, että ihmiset eivät – tietenkään – olleet tyhmiä tai toimettomia, kaukana siitä. Mutta monilta tuntui puuttuvan käsitys siitä, mitä turhauttaville asioille voisi tehdä, ja usko siihen, että omalla tekemisellä voi olla vaikutusta." (Luhtakallio and Mustranta 2017, 14)
All translations from Finnish are by the author.

ostensibly well-meaning actors and social institutions who are supposed to be helping them. Especially illustrative is their description of how the residents described their experience of public education, the social institution that was supposed to be helping them towards social advancement:

> School memories are not the same for everyone. Here they appeared to become a seamless part of that humiliating inheritance of being branded stupid and incapable that many carried with them as their main experience of the whole of society.[2]

Luhtakallio and Mustranta (2017, 55) describe how, for many of the residents, their entire experience of society is a stream of experiences of humiliation and being made to feel inferior by public actors and institutions, often beginning already from school. Such experiences have left many of the residents intensely suspicious of anything "official" and mistrustful of the same political and social institutions that are supposed to be helping them. Feelings of frustration, hostility, and mistrust appear to colour the perception of the residents when it comes to society as a whole. (Luhtakallio and Mustranta 2017, 56–7) This sense of being left outside society is even noted to have a perceptual effect:

> While the networked activist is browsing through the contact information of ten different council members on their cell phone to push their agenda forward, there is elsewhere a group that does not protest or participate in associations, nor set up trendy street festivals. Their city looks completely different – it is not a playground of imagination where everyone can bring their own contribution, nor are the decision makers reachable by phone or a Facebook message, but could just as well reside in another reality. They see their possibilities to influence society, or even to belong to it, as non-existent.[3]

---

[2] "Koulumuistot eivät kuitenkaan ole kaikille samanlaisia. Täällä ne tuntuivat liittyvän saumattomaksi osaksi sitä nöyryyttävää tyhmäksi ja kyvyttömäksi leimaamisen perintöä, jota moni kantoi mukanaan päällimmäisenä kokemuksenaan koko yhteiskunnasta." (Luhtakallio and Mustranta 2017, 55)

[3] "Mutta samaan aikaan kun verkostoitunut aktivisti selaa kymmenen kunnanvaltuutetun yhteystietoja kännykästään viedäkseen asiaansa

This perceptual effect, the inability to see society as something one can influence, or even belong to, is something which appears to be intimately connected to the ability to express oneself in public, an idea I will return to below.

It is important to note that democratic politics thrive on a certain amount of distrust. Mark E. Warren notes that it is by distrusting those in power that we also come to democratically hold them to account (Warren 1999a, 320). However, democracy cannot function without a kind of "generalized trust" which reflects "the capacities of individuals and groups to act for common ends as well as to represent their interests to the state", without which corrupt governments can take hold of society (Warren 1999b, 12). This kind of generalized trust can also be understood as giving to experience a certain "background sense" of security that allows a citizen to make use of their abilities to reach outside oneself in engagement with their social world. Such action is facilitated when one is reasonably secure in the knowledge that they won't be received with ridicule, indifference, or hostility. The loss of such a trust can be experienced as debilitating, as described by Luhtakallio and Mustranta:

> When you discuss politics, participation, and influencing with the local residents, the conversations convey a sense of disappointment and distrust. Society should be the guarantor of help in face of life's ordeals, but this promise has been repeatedly broken. No-one has noticed their distress, or it has not been responded to. The comments also echo with the bitterness brought about by false promises:

---

eteenpäin, on toisaalla joukko, joka ei osoita mieltään tai osallistu juuri yhdistystoimintaan sen enempää kuin järjestä trendikkäitä katufestareitakaan. Heidän kaupunkinsa on aivan eri näköinen – se ei ole mielikuvituksen temmellyskenttä, johon jokainen voi tuoda oman panoksensa, eikä sitä koskeva päätöksenteko ole puhelinnumeron tai Facebook-viestin päässä vaan pikemminkin aivan toisessa todellisuudessa. He näkevät mahdollisuutensa vaikuttaa yhteiskuntaan, jopa ylipäätään kuulumisensa siihen, olemattomina." (Luhtakallio and Mustranta 2017, 118)

> *"Everyone can make something out of themselves." "If given the chance." "But they always pull the rug from under your feet."*[4]

A loss of such a trust from one's perception is, in many ways, analogous to the loss of what Anthony Giddens (1992, 37–8) describes as "ontological security". It is the basis for a stable sense of positive self-identity and social continuity and a sense of belonging to a social fabric. It illuminates one's fundamental background characteristic of *aphonia*: the damaging of the background affective disposition of trust in one's peers and social institutions which forms the bedrock of a sense of having political agency. This loss of trust in others of the social world is mirrored in loss of trust in one's own ability to effectively function inside that world in concert with similarly situated others. This also has a deleterious effect on being able to perceive oneself as a credible and able agent, a person with a voice of their own. Such a loss of a sense of security changes one's perception of the social world and the possibilities within it.

Luhtakallio and Mustranta's description of the lives of residents of the unnamed Helsinki neighbourhood testify to a similar kind of reality as Simon Charlesworth's 2000 phenomenological study on his old hometown of Rotherham in South Yorkshire and its working-class inhabitants, a world from which he himself hails. He uses the tools of existential phenomenology to describe a social milieu suffering from extreme economic hardship, characterized by some of the worst levels of poverty in the Western world. Loss of industry since the 1980s has led to the loss of a credible future horizon for social advancement of its working-class residents. To compound their hopelessness, they are also largely failed by the inability of public institutions, such as education and job programs, to provide meaningful ways forward. What espe-

---

[4] "Kun alueen asukkaiden kanssa puhuu politiikasta, osallistumisesta ja vaikuttamisesta, keskusteluista välittyy pettymys ja epäluottamus. Yhteiskunnan pitäisi olla takuu avusta silloin, kun elämä koettelee, mutta tämä lupaus on toistuvasti rikottu. Kukaan ei ole huomannut hätää, tai siihen ei ole vastattu. Kommenteissa kaikuu myös falskien lupausten herättämä katkeruus:
"Kaikista on johonkin." "Jos annetaan mahdollisuus." "Mutta kun aina vedetään matto jalkojen alta." (Luhtakallio and Mustranta 2017, 26–7)

cially interests Charlesworth is the destructive effect that living in such conditions has on the ability of the residents to authentically express their discomfort and political domination. Working-class Rotherham is described as a world that actively curtails and frustrates the "generative competences for language use and expressive behaviours" of its residents, a phenomenon not adequately captured by the statistical tools often employed by sociologists (Charlesworth 2000, 3).

Charlesworth describes how his research work was made harder by the fact that even people with whom he was intimate did not feel comfortable being formally interviewed. Charlesworth expresses his frustration at the way how people who he knew "to be articulate, thoughtful, insightful and powerfully evocative in their speech, exhibited tendencies of shy restraint as soon as one formalizes the situation, even simply through the introduction of a tape recorder." (Charlesworth 2000, 137) People who were just moments earlier presenting insightful analyses of their situation became silent with the insertion of the recorder into the equation. According to Charlesworth, there is something in the lived experience of the subjects of his study which leads them to suffering from a form of damage done to the very capability to authentically express oneself in public, as if one were afraid of the expressive medium itself (Charlesworth 2000, 283). This is combined with a sense of not being a part of the processes of the society around them. It is as if they are left only with the role of spectators, not agents in their own right:

> The world has become occurrent to them, something they experience 'from the outside', that is, from a position of non-involvement. Possibilities no longer solicit them. They experience a radical discontinuity, an unsettledness emanating from the grounds of the body's projection into the future which creates a sense of the loss of meaning of their lives and yet which makes the meaninglessness of the world in which they live more explicit. (Charlesworth 2000, 79)

Such descriptions reveal another important aspect of the phenomenon. It is not that Charlesworth's interviewees did not possess the linguistic skills to describe their situation, nor have they suddenly chosen to remain silent at the introduction of the tape recorder. Instead, they suddenly find them-

selves unable to speak in a situation perceived as "official", resembling the residents described by Luhtakallio and Mustranta. They do not feel "socially instituted to have opinions", and as the official language doesn't feel like theirs to use, they often fall back to the "ultimate euphemization of silence", a feeling so strong it takes hold of them against their own will. (Charlesworth 2000, 135–7) Charlesworth is worried that the entire social world he is speaking of is being enveloped by silence, as those living in it lose the ability to give voice to their situation (Charlesworth 2000, 3).

This silence that is endured, in a sense, even against one's will, encapsulates what I mean by not having a political voice of one's own. Such *aphonia* must be approached as something negative, something which reveals itself only as an absence: public silence not as choosing to stay silent, but rather as the absence in the field of experience of the very possibility of speaking out. It has its roots not in lacking (narrowly defined) cognitive and communicative capabilities, but in the way human beings inhabit their environments and interact with them on a pre-cognitive level of embodied awareness that is primordial to conscious awareness. It is also at this pre-cognitive level of bodily existence that our expressive capacities, use of language included, take root as certain types of habits, which are intimately connected to our bodily existence and the way our bodies perceive their environments as perceptual fields already flush with a sense of meaning and possibility or their absence. I believe that the best method for approaching this pre-cognitive level of intentional experience is Merleau-Ponty's phenomenology of the perceiving and expressive body.

Speech as an Expressive Modality of the Body

I believe that the above examples provided by both Luhtakallio and Mustranta and Charlesworth testify to different aspects of a phenomenon of losing one's own political voice, an injustice which contemporary theories of democracy have not yet adequately conceptualized. This is due to an insufficient analysis of the experiential conditions of being able to speak in public in the first place. While critical conceptions of exclusion and injustice go a long way towards helping us under-

stand the problem, they are unable to consider the subjective side of the equation: the feeling of being unable to appear and speak competently in public.

What is at issue is a deeply seated experience of oneself as not capable of speaking in public situations, the feeling of lacking the "proper" words, even if in private one was, like Charlesworth's interlocutors, an intelligent and eloquent analyst of one's situation. To understand this phenomenon, I use Maurice Merleau-Ponty's phenomenology to approach public speech as one modality of a more general embodied expressivity. Merleau-Ponty's concepts of operative intentionality, body-subject and the lived body allow us to understand how human beings continuously relate to their social environment on a pre-cognitive, pre-discursive embodied level of perceptual intentionality.

This approach is indebted to the work of Edmund Husserl, and his conception of intentionality, the way that conscious experience is always consciousness *about* something, upon which Merleau-Ponty builds. With "operative intentionality," Merleau-Ponty describes a form of intentionality that is primordial to conscious reflective intentionality; the knowledge that, for example, I am aware that I am currently experiencing something (Merleau-Ponty 2012, lxxxii). Operative intentionality is a pre-cognitive, embodied form of intentionality that offers conscious experience the perceptual field which it encounters as objectivity. This field is already experienced as meaningful, that is, a field of not only visual phenomena, but also an affective field. This means that we perceive the world as already presenting a field which pulls us towards certain possibilities while repelling us from others. As such, the world already presents us with solicitations for action.

Central to Merleau-Ponty's project in the *Phenomenology of Perception* (2012) is the conception of the human being as a body-subject: the intertwining of a subjective, reflexive consciousness and the objective being of the body as a physical object which is encountered as such by other human beings. This intertwining is mediated by the "lived body", something which, in a sense, "comes before" our conscious sense of self as a perceiving being. Our experience is always rooted in an "anonymous" level of bodily experience that exhibits an agency of its own that is somewhat alien to us, as it is the re-

pository of unconscious habits and meanings, acquired during one's lifetime. It is the lived end-product of an extensive social conditioning through the process of living a singular, particular life in a particular place and time. The resulting "habitual body" includes not only mostly unconscious habits of comporting one's body and expressing oneself, but also a certain style of perceptual organisation, and even creation, of space around oneself in a perceptual field that the body-subject encounters as objectivity. As Monika Langer describes it, "The 'habitual body' already projects a habitual setting around itself, thereby giving a general structure to the subject's situation." (Langer 1989, 31.) In this way the "sedimentation", or incorporation of past experiences, meanings and knowledge into one's lived body, results in a personal style of being, acquired over time, combining cognition, perception and the motor intentionality of the body itself (Merleau-Ponty 2012, 113). The body-subject projects meaningful space around itself through interrogating its surroundings in an organic manner, giving it a sense of meaningfulness. This allows the world to present in perceptual experience a field of meaningful possibilities for action that the body-subject can act towards.

The lived body, then, is primordial to one's reflective intentionality and awareness or consciousness, presenting a somewhat autonomous and anonymous level of being. It is a product of sedimentation of meanings and social practices into somewhat stable habits of acting and seeing, forming the ground of all our agentic capacities. The spontaneity inherent to human freedom must be understood as the movement in which a body-subject improvises on meanings already sedimented within itself according to a developed personal style to answer solicitations present in its perceptual field. This expressive movement of the body outside itself mixes the cognitive, the perceptual and the motor intentionality of the lived body, and also includes our capacity for language – the primary medium of politics.

The lived body spontaneously relates to its social environment, improvising on the shared meanings it finds within it and thereby answers the solicitations presented by the field with which it is engaged. Expressivity includes ways of inhabiting a place, of bodily comportment and action through

which we appear to others both as human beings and as physical objects that are perceived and valued by others. This is a relationship that can become damaged or even destroyed when the experiential conditions for acquiring these expressive capacities are curtailed or lacking in the situation one inhabits in a given society.

The ability to express oneself, to project oneself freely into the world, can be understood as a product of sedimentation of positive social experiences. In successful social interactions, a kind of a positive feedback loop takes place. As one is given positive feedback for one's overtures towards the world, a positive sense of "being able" begins to take root in experience. However, a similar feedback loop can also feed on negative experiences, curtailing one's ability to freely express oneself. The bleak social realities in which many marginalized groups find themselves often actively frustrate the acquisition of expressive capacities and the ability to relate to ones' social environment. These social realities, therefore, can even constrain the cognitive capabilities of those inhabiting them to make sense of their experiences and to express them in public.

We are always part and parcel of the historical situation we find ourselves always already thrown into and constituted by; this does not foreclose our freedom to act out of our own will, but instead gives us the field of meaningful possibilities we can act towards. Our ontological intertwining with our situation means that political agency is not merely a matter of exercising one's autonomous will, but instead an unstable process that is shot through with the ambiguity inherent to our embodied being. It is due to our lived body projecting meaning around itself in this process of intertwining that our situations appear to us as soliciting us to act upon them. It is in the pre-reflective perceptual relationship of the lived body to its social environment that the body-subject encounters a world and communicates with it. It is also there that we can start to understand the root of *aphonia* as an obstacle to exercising one's political freedoms as a citizen of a democratic community.

Aphonia as an Obstacle to Freedom

We can now turn towards *aphonia* as a political phenomenon. *Aphonia* is not an objective state of affairs that could be judged to be a violation of a pre-political, philosophically reached norm given by an "ideal" theory of justice, such as that of John Rawls (1999). Examples of such injustices would be material poverty or exclusion from equal democratic participation. Nor is it a purely internal matter of subjective attitude. The experienced incapability to speak out in public and to make a difference is the incorporated result of inhabiting a certain place in society, a certain historical situation. As seen in the above examples provided by Luhtakallio & Mustranta and Charlesworth, *aphonia* is associated with inhabiting a certain kind of social world which offers little possibilities for learning how political engagement "works" and provides little hope that one's actions could actually change something for the better. What mainstream philosophical approaches lack is precisely the focus on how such conditions become *incorporated* through sedimentation and can form an obstacle to feeling like one is a capable and credible witness of one's own suffering.

It is now possible to see how *aphonia* is a phenomenon intimately connected to the expressive capacities of the lived body, and therefore also connected to the affective background of perception. A person "without a voice of their own" is not necessarily objectively lacking in resources or the communicative capabilities required for effective democratic participation, even if the lack of such capabilities contributes to their *aphonia*. Instead, they might be unable to perceive in their environment a field that would welcome their participation, or give a sense of opportunities for speaking out. Quite the contrary, the world might appear as indifferent to their concerns or even actively hostile to their presence.

Aphonia might be a result of multiple empirical causes: these include being a member of a social group whose members are not recognized as social equals; objectively lacking the linguistic capabilities which would allow them to present themselves as credible political speakers, or even due to simply feeling like "persons like them" are "not political". I leave aside the empirical matter of exact causation, as what this

article attempts to describe is the political quality of *aphonia*, the way it constitutes an injustice. *Aphonia* is an injustice because it constitutes an obstacle to exercising one's share of political freedom. According to Merleau-Ponty, freedom is meaningless unless considered against its context, the situation in which the body-subject finds itself. Freedom is experienced as meaningful when there is a situation one can perceive as calling for action, that presents some possibilities or "cycles of behaviour" one can take up which would be left unrealized without acting upon them (Merleau-Ponty 2012, 462). Freedom can be understood as a successful interaction between a body-subject and a field, a kind of fit Merleau-Ponty describes as a "gearing into" a situation which "calls forth privileged modes of resolution and that it, by itself, lacks the power to procure any of them." (Merleau-Ponty 2012, 467) Freedom, then, always means "taking up" the solicitations presented by an existing situation through becoming engaged and involved with it, and perhaps even attempting to change it.

Nick Crossley (1996, 151) provides an interesting approach to freedom by describing how citizenship can be understood as an intersubjectively constituted role that one must inhabit in a meaningful sense to experience political participation as a meaningful possibility:

> In order to perform their role, citizens must have a shared sense of that role, a sense of citizenship. And they must have the know-how required to perform that role competently. 'Citizenship' must be meaningful to them as a group. It must be a constitutive feature of their shared interworld and an identity which each assumes therein. It must be embedded in the texture of taken-for-granted assumptions which comprise the meaning horizon of our everyday life; that is, in the (intersubjectively constituted) lifeworld (Roche 1987; Schutz 1964).

Crossley describes a way to approach citizenship as a role that one inhabits to a degree that it becomes invisible: citizenship becomes a part of the horizon of experience against which things and phenomena of the social world can appear as meaningful objects of political engagement. The social world comes to present a meaningful field one can become engaged in. This could be compared to the way Hannah Ar-

endt describes the public world as something which only exists in interactions between men, a "web" that action and speech knit over the material world of artefacts (Arendt 1998, 182–3). This way of relating to the social world as a political matter can be approached as a facet of the expressivity of the lived body, something that a person can also lose due to little fault of their own.

As an example, Miranda Fricker describes how a negative social stereotype about a social group, for example, "the members of this group are not politically active", can become internalized by the members of said group to a degree that they withdraw their political participation – as if through a learned habit:

> …we can imagine an informally disenfranchised group, whose tendency not to vote arises from the fact that their collectively imagined social identity is such that they are not the sort of people who go in for political thinking and discussion. 'People like us aren't political'; and so they do not vote. (Fricker 2007, 16)

Fricker notes that the converse can also be the case: members of a group whose shared identity includes an image of the self as an active and capable political agent are more likely, for instance, to go out and vote (ibid.). Fricker's example of an internalized negative self-image describes something of the phenomenon at hand: that of an internalized sense of oneself as incapable to act, to speak out in public. This sense can become an obstacle to freedom, as one loses a sense of oneself as a political agent, and of political engagement as a field which presents meaningful possibilities for action.

Production of citizenship as a meaningful role is connected to a welcoming and egalitarian civic culture. It is not simply a discursive affair, but a product of embodied *mimesis*, of learning "body-to-body" how to appear in public as a citizen among equals. Charlesworth describes how the subjects of his study have, as a class, fallen outside such civic culture:

> Such conditions of scarcity amidst affluence, of severe vulnerability amidst images of security, of dislocation without movement, have led to the creation of a class in which many have come to appear 'odd', abject, because they have been unable to

> participate in spaces in which they could learn, mimetically, body-to-body, the manners and styles of deportment of the accomplished adult, attuned to the respectable world of a civilized realm in which there exists, practically and dispositionally, a civic culture oriented to public civility. (Charlesworth 2000, 159.)

One learns such civic culture through the practices of being a citizen and in time these become incorporated into the horizon of operative intentional experience. This is sedimentation at work: one incorporates a civil form of being and perceiving into one's living body, allowing the world to appear as a field of possibilities for speaking and acting. In speech, the body-subject re-activates those sedimented meanings and practices. When one lacks possibilities for learning how to be in civil society, the very public realm itself begins to appear as a distant place one does not inhabit, a separate world from one's own. To compound the problem, those who do not know "how to be" in public often stand out as abject, stigmatized beings, running the danger of becoming objects of shunning or ridicule by others.

The cultural specificity and social power relations present in conversational norms which guide public discussion are an important facet of political exclusion (Young 2000, 55 ff.). However, we should also pay attention to how economic poverty and social marginalization can become sedimented into the habitual body as the limiting of the body's very capacity to express itself more or less confidently in public. What to a middle-class educated eye appears as self-evident, the fact of human beings equally possessing speech, is revealed by a phenomenological analysis to be a socially constituted and contingent state of affairs.

Pierre Bourdieu notes that politically dominated groups can silently reject the "official" world and language of politics while remaining dispossessed of the means of presenting one's own alternative to them (Bourdieu 1991, 51–2). Effective political agency is contingent on being able to perceive politics as a field of meaningful possibilities towards which one may act. If one becomes unable to perceive the world around them as containing meaningful possibilities, or possibilities to engage in social practices which could create such possibilities, then one is left without a horizon in which effecting change is perceived as possible.

Aphonia, then, is something more than a lack of objectively measurable resources or capabilities. As Merleau-Ponty writes:

> If there were no cycles of behaviour, no open situations that call for a certain completion and that can act as a foundation, either for a decision that confirms them or for one that transforms them, then freedom would never take place. […] If freedom is to have *a field to work with*, if it must be able to assert itself as freedom, then something must separate freedom from its ends, freedom must have *a field*; that is, it must have some privileged possibilities or realities that must tend to be preserved in being. (Merleau-Ponty 2012, 462)

If one inhabits a situation lacking in such "possibilities or realities", then it is this very lack that may become sedimented into one's lived body. This curtails the ability to perceive possibilities where someone better situated might be able to see them and act upon them. One might become subjectively justified in thinking that speaking has become meaningless even when there is no objective confirmation to justify such a feeling. After all, why speak when there is nobody who would listen? This means it would be meaningless to ask a person suffering from *aphonia* for confirmation to the contrary. What matters is precisely that they are unable to perceive any possibility for effecting meaningful change while having lost confidence in their own ability to voice their frustration and suffering in public.

Combating *Aphonia* through Therapeutic Encouragement and Engagement

As described in the examples above, the tendency among marginalized persons to withdraw from civil public life to suffer in silence is often combined with a suspicion towards, and even fear of, anything "official". In such a situation, even well-meaning measures towards political inclusion through, for example, participation in democratic processes which are directed from above, are apt to backfire. When one already perceives "official" situations as off-putting due to, for example, the shame one feels when forced to appear and speak in public in the presence of "one's betters", even well-meaning

demands for inclusive democratic participation can have the unwanted result of driving away the very people that one is trying to draw into political engagement. I do not claim to present ready solutions to ameliorate such situations. However, I will end this essay by presenting the story of a possible approach which begins from grass-roots engagement with politically marginalized persons.

Luhtakallio and Mustranta (2017) tell their story in the context of a community project which attempts to draw the politically marginalized residents of a disadvantaged Helsinki neighbourhood into political engagement. After many failures in getting through to the residents of the neighbourhood, usually stemming from their own preconceived notions of what political engagement "should" look like, Luhtakallio and Mustranta finally themselves realize that what is needed is a change of perspective and approach. Instead of dragging the residents into protests and badgering them to participate in direct political action, something which none of them had ever seen any meaning in, Luhtakallio and Mustranta began attending to the affective needs of the residents. They tried to make the residents feel welcome and appreciated through simple and relatively inexpensive things like offering meals and drinks, making sure that single mothers have the possibility to attend the meetings through providing childcare, and even paying for a taxi trip to the theatre, a rare luxury for the residents.

An important facet of the issue was that many of the residents had never felt wanted and appreciated in public situations, something which an educated middle-class outsider has a hard time understanding. After this sense of being unwelcome and ignored was attended to, things started moving forward. Ultimately the project leaders and the residents collaborated in a community theatre project in which an entire political play about marginalized immigrant residents of the neighbourhood was written by the residents themselves, with assistance from theatre professionals. Most of the residents had never conceived of themselves as capable of such a feat, yet with resources and encouragement, they enthusiastically completed the project and held their play at the local community centre.

After the play, the project disbanded, with seemingly little results: some participants continued their new hobby as amateur thespians, but for most, life assumed its usual course. However, this is only a failure on the standards of seasoned political activists, with their own preconceived notions of what constitutes a success. For the participants, the project was an important experience of finally having their say in a world that is largely indifferent, even hostile, to their concerns, an experience which might bear fruit later in unexpected ways. Perhaps one of their children may decide to join a political party or start a social media drive for some cause. Perhaps they teach someone else that engagement is possible and does matter. There are no certain outcomes, only the opening of new possibilities.

I have decided to describe Luhtakallio and Mustranta's grass roots approach as "therapeutic", because, instead of worrying about systemic and society-encompassing objectively verifiable forms of political injustice, it attends to the subjective experiences and emotions of marginalized persons and attempts to acknowledge and address them first. The results achieved by such an approach are, perhaps, somewhat beneath the grand systemic ambitions of critical theories of democracy. This does not mean that they would not still have an important effect among persons and groups suffering from *aphonia.* Helping individual persons and groups to find their own voice as citizens after an entire lifetime of being made to feel inferior and incapable might also be an efficient way to combat *aphonia* in a way that is felt as effective by those who have heretofore been unable to authentically express themselves on their own terms.

Concluding Remarks

We can now see how the conception of *aphonia* I have presented describes a situation which is qualitatively different from not having the opportunity to speak, whether due to being excluded from a political community or due to lacking the necessary material resources and cognitive and communicative capabilities to do so. *Aphonia* describes the disappearance of the possibility of speaking out in public from the field of possibilities present in experience. It is a phenomenon

that, as I have shown above, can be disclosed by a phenomenological approach to critique which begins from an understanding of speech as rooted in the broader expressive capacicapacities of the lived body. Merleau-Ponty's conception of operative intentionality as a dialectic between the lived body and its environment is uniquely suited to understanding how language and speech form but one expressive modality of the lived body, and as such, are always dependent on the continuous process of the intertwining of the lived body with its environment and the meanings found therein.

The expressivity of the lived body is a dimension of experience which is often not adequately recognized by democratic theory. Instead of thinking public speech in terms of capabilities, as skills or abilities which allow one to "function" effectively in society (Bohman 1997, 325), we should be thinking about speech, and the more general category of expressivity, from a phenomenological perspective which presents expressivity as a body-subject's way of responding to the solicitations present in its perceived field of possibilities.

The kinds of structurally enforced silences, curtailings of expressive capacities and experiences of *aphonia* suffered by marginalized groups the world over are such burning injustices because those suffering from them are denied the possibility of appearing as capable political agents amongst their peers. Instead, the lack of confidence to express themselves in public prevents the political voicing of their concerns and furthers their political marginalization. *Aphonia* is thus not another type of objective inequality that could be approached as a question of unjust distribution of social goods, be they resources, recognition, or adequate capabilities. It is a question of losing a practical sense of certain possibilities that, in a functioning democracy, should in principle be present in the field of experience of every citizen.

In an egalitarian society, every citizen should have the equal possibility of learning to feel the role of citizenship as something which belongs to them, conferring them a voice of their own. Since *aphonia* can be presented as the result of a sedimentation of experienced political domination into relatively stable habits of perception, comportment, and use of language, it also describes how the least powerful can end up being the least capable of expressing themselves in public. As

one is stripped of their sources of positive self-identity as an equal member of a political community, one can lose one's own voice, the capability to express oneself authentically on one's own terms, to make one's frustrations and sufferings known in public.

I have in this article presented a disclosing critique of the phenomenon of *aphonia* as the loss of a political "voice of one's own" and shown how it can be approached through Maurice Merleau-Ponty's existential phenomenology as the sedimentation of negative social experiences into the lived body of an individual due to them inhabiting a social situation involving economic and social marginalization. Such experiences can become internalized and incorporated into the operative intentionality of a body-subject to a degree it becomes almost impossible for them to perceive themselves as capable and credible citizens, and the world around them as a field of possibilities for political engagement which welcomes them as equals. This can damage the expressive capacities of the lived body to a degree that it becomes impossible to express one's frustrations and sufferings authentically in public: to have a voice of one's own.

University of Helsinki

References

Allen, D. (2005), "Invisible Citizens: Political Exclusion and Domination in Arendt and Ellison", in *Political Exclusion and Domination*, S. Macedo and M. S. Williams (eds.), New York, New York University Press, pp. 29–76.

Arendt, H. (1998), *The Human Condition,* 2nd ed., Chicago, University of Chicago Press, original edition, 1958.

Benhabib, S. (2004), *The Rights of Others. Aliens, Residents and Citizens,* Cambridge, Cambridge University Press.

Bohman, J. (1997), "Deliberative Democracy and Effective Social Freedom: Capabilities, Resources, and Opportunities", in *Deliberative Democracy. Essays on Reason and Politics*, J. Bohman and W. Rehg (eds.), Cambridge, MA, MIT Press, pp. 321–348.

Bourdieu, P. (1991), *Language and Symbolic Power,* translated by G. Raymond and M. Adamson, Cambridge, Polity Press.

Charlesworth, S. J. (2000), *A Phenomenology of Working-Class Experience,* Cambridge, Cambridge University Press.

Crossley, N. (1996), *Intersubjectivity: The Fabric of Social Becoming,* London, SAGE Publishing.

Fricker, M. (2007), *Epistemic Injustice. Power & the Ethics of Knowing,* Oxford, Oxford University Press.

Giddens, A. (1992), *Modernity and Self-Identity. Self and Society in the Late Modern Age,* Cambridge, Polity.

Kompridis, N. (2005), "Disclosing Possibility: The Past and Future of Critical Theory", *International Journal of Philosophical Studies* 13 (3): pp. 325–351.

Kompridis, N. (2008), "Struggling over the Meaning of Recognition", in *Adding Insult to Injury. Nancy Fraser Debates Her Critics*, K. Olson (ed.), London, Verso, pp. 295–309.

Langer, M. M. (1989), *Merleau-Ponty's Phenomenology of Perception. A Guide and Commentary,* London, Macmillan.

Luhtakallio, E., and M. Mustranta (2017), *Demokratia suomalaisessa lähiössä,* Helsinki, Into Kustannus.

Merleau-Ponty, M. (2012), *Phenomenology of Perception,* translated by D. A. Landes, London, Routledge. Original edition, 1945.

Rawls, J. (1999), *A Theory of Justice. Revised Edition,* Cambridge, MA, Harvard University Press. Original edition, 1971.

Warren, M. E. (1999a), "Democratic theory and trust", in *Democracy and Trust*, M. E. Warren (ed.), Cambridge, Cambridge University Press, pp. 310-345.

Warren, M. E. (1999b), "Introduction." In *Democracy and Trust*, M. E. Warren (ed.), Cambridge, Cambridge University Press, pp. 1-21.

Young, I. M. (2000), *Inclusion and Democracy,* Oxford, Oxford University Press.

The Stratigraphic Fallacy or the Anthropocene as an Epistemic Question

AGOSTINO CERA

Introduction

This paper provides an *epistemic investigation of the Anthropocene idea*.[1] It highlights the intrinsic epistemic ambiguity/peculiarity of this idea, which emerges halfway between ideology and geology and subverts the traditional distinction between the "two cultures" (natural and human sciences), going beyond the difference between *physis* (nature) and *techne* (culture) itself.

To prove that the Anthropocene is a "threshold concept" capable of shaking the very foundation of the sciences and forms of knowledge that study it (in particular geology), I will deal with Carlos Gray Santana's critical discussion of this aspirant geological epoch. More precisely, I will focus on the *Stratigraphic Fallacy* at the base of his "Stratigraphic Miso-Anthropocene". In conclusion, I will put forward an *ad hoc* definition of the Anthropocene inspired by Timothy Morton's work: the Anthropocene is an *epistemic hyperobject with a (geo-) historical barycenter*.

1. Birth of an Epoch

As is well-known, Dutch chemist Paul Jozef Crutzen – 1995 Nobel Prize winner for his 1970s work on anthropogenic damage to the ozone layer – used the word "*Anthropocene*" in the year 2000 to represent the geological break between our current age and the *Holocene*: the second epoch (following the

---

[1] Translated from Italian by Jack Spittle. The author would like to thank the reviewers of this paper, whose comments and suggestions have been extremely helpful.

Pleistocene) of the Quaternary or Neozoic period. Inspired by an intuition of Charles Lyell, the term "*Holocene*" (*Holocéne*) was coined in 1850 by French paleontologist and entomologist Paul Gervais and formalized – apparently – at the *Second International Geological Congress* (IGC) in Bologna in 1881. The Holocene began 11,650 years ago with the end of the last glacial period. It is characterized by a significant increase in average temperature and sea level, both of which subsequently stabilized approximately 8,000 years ago.[2]

Some months after Crutzen's unofficial announcement during an *International Geosphere-Biosphere Programme* (IGBP) conference in Cuernavaca, Mexico, the official one came: a very brief article – a kind of *anthropocenic manifesto* – entitled *The "Anthropocene"*. Published in the *Newsletter of the IGBP,* it was signed by Crutzen and Eugene Filmore Stoermer,[3] an American biologist who had already used the word "Anthropocene" informally in the 1980s. The article featured a list of the parameters that objectively show an escalation in the anthropic variable (human agency) over the last three centuries: increase in human population, urbanization, exploitation of fossil fuels, the so-called "sixth mass extinction", climate change and the concentration of greenhouse gases. According to Crutzen, the exponential increase of CO_2 in the atmosphere – of obvious anthropogenic origin – is conclusive proof of this new geological epoch.[4]

The anthropocenic idea can be summed up as follows: "human activities have become so pervasive and profound that they rival the great forces of Nature and are pushing the Earth into planetary *terra incognita*. The Earth is rapidly moving into a less biologically diverse, less forested, much warmer, and probably wetter and stormier state." (Steffen and Crutzen *et al.* 2011a, 614.)

---

[2] See Zalasiewicz *et al*. 2008, 4–5.

[3] See Crutzen and Stoermer 2000. The *International Geosphere-Biosphere Programm*e (IGBP) is an international program created in 1987 and based in Sweden. With its establishment, "Earth system science gained the institutional capacity it would need to build a robust interdisciplinary community of scientists dedicated to advancing Earth system science" (Ellis 2018, chap. 2.7).

[4] For a summary of these parameters see Steffen and Crutzen *et al.* 2011b, 851–852.

Beyond helping us imagine the concrete future of our planet, this vivid image of a *terra incognita* (unknown land) is also valid from a conceptual point of view. Despite the enthusiasm of the so-called "Anthropocenologists" – described by Christophe Bonneuil and Jean-Baptiste Fressoz as the "phalanx of renowned scholars who made the bold gesture of naming our epoch" (Bonneuil and Fressoz 2016, chap. 3.0) – it quickly became clear that the Anthropocene was not so much an epoch as a "discourse", that is "in the strong sense of organizing the perception of a world picture (past, present, and future) through a set of ideas and prescriptions" (Crist 2016, 24).

It could even be considered an *ideology*, a *Weltanschauung* or a "paradigm dressed as epoch" (Baskin 2015).[5] As Bonneuil and Fressoz correctly affirm, l'*Evénement Anthropocène* as "geohistorical event" establishes a new *grand récit* in which the human being ensures its full power within "a hegemonic system for representing the world as a totality to be governed." (Bonneuil and Fressoz 2016, chap. 3.0).

This is why they suggest calling the new epoch Capitalocene or Oliganthropocene, rather than Anthropocene.

One thing is sure: the Anthropocene idea emerges as an *epistemically unstable dispositif* due to its intrinsic tendency to exceed its original epistemic boundaries of the natural sciences. What I am going to argue may sound quite similar to environmental humanities scholar Timothy Clark's definition of the Anthropocene as a "Threshold Concept" (Clark 2015). Where we differ is that, in my view, its instability/ambiguity does not concern a single topic (i.e., ecological crisis), but the epistemic dimension as such. More precisely, I think the

---

[5] When considering the Anthropocene, my Gramscian use of the term "ideology" is well summarized by these lines from Manfred Steger: "Ideology can be defined as a system of widely shared ideas, patterned beliefs, guiding norms and values, and lofty ideals accepted as 'fact' or 'truth' by significant groups in society. Codified by social elites, ideologies offer individuals a more or less coherent picture of the world not only as it is but also as it should be. In doing so, they organize the tremendous complexity of human experience into fairly simple and understandable images that, in turn, provide people with a normative orientation in time and space and in means and ends." (Steger 2009, 6).

Anthropocene emerges as a kind of hyperobject (following Timothy Morton's definition). Or better, an *epistemic hyperobject*, as it puts the traditional distinction between "the two cultures" – following Charles Snow's definition – in radical and perhaps irreversible crisis. As a result of this epistemic ambiguity/peculiarity, while the scientific community is still hesitant to support the legitimacy of the aspirant geological epoch, the word "Anthropocene" has become a trend topic within the cultural debate (the opposite epistemic side). Beyond all the books, papers, conferences, workshops, websites, movies… on the subject, there are now even monothematic journals (*The Anthropocene Review* and *Anthropocene*), *Atlas* (Gemenne and Rankovic 2019) and an *Encyclopedia of the Anthropocene* (DellaSala and Goldstein 2017).

According to Ben Dibley, cultural studies scholar and author of *Seven Theses on the Anthropocene,* the Anthropocene must be considered an "ambivalent formulation", that is an epoch and discourse at the same time (thesis I). On the one hand, it presents a new epoch and a new geological agent "which would make any distinction between nature and society untenable", giving birth, as a result, to "hybrid naturecultures" (Dibley 2012, 144). On the other it "retains nostalgia for that very distinction", specifically a "*nostalgia for the human*" (Dibley 2012, 142 – thesis IV). The idea that human agency has become a "geological/geophysical force" or "geological power" – i.e., the main force within the *Earth System* – clearly transcends such a distinction. When characterizing the Anthropocene chronologically, for instance, one does not use natural time – as has always been the case for geological epochs – but historical time. In other words, we are faced with *a historical time used as a geological chronology*. As Dibley states: "*The Anthropocene is the crease of time,* […] the advent of the human as a geological agent demands ways of thinking these temporalities [i.e., the deep time of geology and a rather shorter history of capital] together." (Dibley 2012, 140 – thesis II.)

A much more critical take on this matter comes from the sociologist and ecological activist Eileen Crist, whose approach is an example of *Bad Anthropocene,* a position opposite to the *Good* or *Great* Anthropocene expressed by the

Ecomodernists/Ecopragmatists.[6] She thinks the Anthropocene performs a space-time colonization, as "in the Anthropocene discourse we witness […] history's conquest not only of geographical space but now of geological time as well." (Crist 2016, 17.)

One thing these different interpretations seem to agree on is that we are dealing with an indissoluble *natural-cultural temporality,*[7] and there are many proposals that seek to epistemically normalize this anomaly. For instance, the post-colonial historian Dipesh Chakrabarty speaks of "deep history" or "negative universal history" (see Chakrabarty 2009) while the philosopher Clive Hamilton speaks of "geohistory". In Hamilton's view, "We are shifting from humanist history as the story written by free humans to geohistory, a story penned by geohistorians, attempting to tell the story of powerful beings soon to be overwhelmed by more powerful forces." (Hamilton 2017, chap. 4.4.)

Another peculiarity of the Anthropocene is that it represents the first *predictive* geological epoch. Not only does it geologically rename the recent past and current epoch, it seeks to characterize or even mortgage future ages (centuries and even millennia). Some scientists (geologists) are embarrassed by this situation, and use it to justify their skepticism of the anthropocenic hypothesis, in particular their aversion to what the geologist and paleontologist Jan Zalasiewicz has been attempting with the *Anthropocene Working Group* (AWG) since 2009. We will deal with this skepticism directly later, when

---

[6] "*Good Anthropocene*" refers to the most optimistic interpretation of the Anthropocene according to which any environmental challenge related to the new epoch (starting with climate change) can be overcome by means of a technological solution. The leading figures of this technolatric optimism are the so-called *Ecomodernists* or *Ecopragmatists*, a group of scholars linked to *The Breakthrough Institute*, a think tank located in San Diego (https://thebreakthrough.org/). Their agenda is outlined in the *Ecomodernist Manifesto* (see Asafou-Adjaye *et al.* 2015).

[7] Since 1977, the standard method for defining and dating geological epochs is that of "golden spikes" (scientific name: *Global Boundaries Stratotype Section and Point* – GSSP). These are "distinctive biostratigraphic signs" used as markers to identify the stratigraphic limits necessary for dividing geologic time (see Ellis 2018, chap. 3.4).

discussing Santana's *Stratigraphic Miso-Anthropocene*.[8] With respect to this predictive tendency, Crutzen's *anthropocenic manifesto* had already argued that "without major catastrophes like an enormous volcanic eruption, an unexpected epidemic, a large-scale nuclear war, an asteroid impact, a new ice age, or continued plundering of Earth's resources by partially still primitive technology […] mankind will remain a major geological force for many millennia, maybe millions of years, to come." (Crutzen and Stoermer 2000, 18.)

It follows that the Anthropocene is *the first geological epoch to be theorized ex ante*.

2. Beyond the Two Cultures: Epistemology of the Anthropocene

A Hermeneutic Taxonomy of the Anthropocene

Bad Anthropocene = Eileen Crist, Timothy Clark

Beautiful or *Great* Anthropocene = the Ecomodernists, Christian Schwägerl, Bruno Latour

Critical Anthropocene = Jeremy Baskin

Eco-anthropocene (or *Ecological* Anthropocene) = Andreas Malm, Alf Hornborg, Ian Angus, Timothy Clark

Geological Anthropocene (or *Standard* Anthropocene) = Paul Crutzen

Good Anthropocene = Erle Ellis, Andrew Revkin

Hard Anthropocene = Clive Hamilton

Historical Anthropocene = Dipesh Chakrabarty

Moral Anthropocene = Lisa Sideris, Eileen Crist

---

[8] In 2009 Zalasiewicz founded the *Anthropocene Working Group* (AWG), which he still presides over. Established by the Subcommission on Quaternary Stratigraphy (SQS) of the International Commission on Stratigraphy (ICS) – itself a constituent of the International Union of Geological Sciences (IUGS) – the AWG is a leader in the attempt to geologically legitimize this aspirant new epoch. Zalasiewicz recently edited a book that is a *summa* of scientific knowledge on the Anthropocene (see Zalasiewicz and Waters 2019).

Political Anthropocene = Chistophe Bonneuil and Jean-Baptiste Fressoz, Jason W. Moore

Romantic Anthropocene (or *Soft* Anthropocene) = Jedediah Purdy

Stratigraphic Philo-Anthropocene = Jan Zalasiewicz

Stratigraphic Miso-Anthropocene = Carlos Gray Santana

As is well-known, many different starting dates have been proposed for the Anthropocene. What this diagram shows is that, beyond a chronological taxonomy of the various interpretations, it is possible to build a *hermeneutic taxonomy* of them. In addition to "when did the Anthropocene begin?" we can ask "what does the Anthropocene mean?" While certainly a useful tool at first glance (another means for deciphering the new paradigm), this additional taxonomy actually raises a radical objection, because the very existence of the anthropocenic hypothesis puts the condition of possibility of such a classification up for discussion. This condition of possibility equates to the classical distinction between the two cultures: hard sciences and humanities. It is this distinction which allows us to establish whether the Anthropocene is a geological or a historical phenomenon/epoch and thus a topic for the hard sciences or the humanities.

As mentioned above, the epistemically crucial point here lies precisely in the fact that the Anthropocene idea is proving more and more to be a hybrid: an object (hyperobject) whose very existence requires the setting aside of such a distinction. Even in its most basic and neutral interpretation (i.e. as a potential geological epoch) the Anthropocene's *conditio per quam* is the overcoming not only of the difference between two forms of knowledge (the two cultures), but of the difference between the categories of nature (*physis*) and culture (*techne*) themselves. It moves towards an osmosis of them (or an *Aufhebung*, if you prefer). This overcoming/osmosis should be considered a kind of transcendental of the anthropocenic hypothesis: the Anthropocene's basic feature or what I call the *anthropocenic Urphänomen* according to the

Goethean meaning of the word.[9] In this very particular case, "beyond the two cultures" means "beyond Nature and Culture",[10] that is the epistemic side immediately involves the categorial/ontological side. This is the epistemic instability I am trying to highlight, an instability which establishes the primacy of the hermeneutical dimension as a result. To sum up, given that the Anthropocene can legitimately be considered both a geological and/or historical phenomenon, its ontological status depends on its hermeneutical status. The answer to "what is the Anthropocene?" depends on how we classify it, on how we choose to interpret it.

Scholars support different perspectives on this instability. Three are the most important.

1. Underestimation or even negation of the instability. In this case scholars negate the newness of the Anthropocene as such, to the point of contesting the utility of an *ad hoc* definition. Among scientists, this position is put forth by palaeoclimatologist William F. Ruddiman's *Early Anthropogenic Hypothesis* (or *Palaeo-anthropocene)* (Ruddiman 2003) and reiterated in a recent article by Carlos Gray Santana that I will address shortly. It is worth noting that within geological science there is another approach that, despite adopting a similar geological/stratigraphic focus as Santana's *Stratigraphic Miso-Anthropocene*, actually claims the official recognition of the new geological epoch rather than rejecting it.[11] This *Stratigraphic Philo-Anthropocene* is led by the aforementioned Jan Zalasiewicz: the most active supporter of the Anthropocene's geological legitimacy for the better part of ten years now.

2. Recognition of the Anthropocene's peculiarity while seeking to offer a normalizing interpretation of it. These

---

[9] Goethe calls some phenomena "*Urphänomen*" (archetypal phenomena) because "nothing higher manifests itself in the world; such phenomena, on the other hand, make it possible for us to descend, just as we ascended, by going step by step from the archetypal phenomena to the most mundane occurrence in our daily experience." (Goethe 1983, 195.)

[10] Obviously, I refer here to Philippe Descola's well known 2005 book (see Descola 2013). An inspiring discussion of Descola's approach to the Anthropocene can be found in Hamilton 2017, chap. 3.4.

[11] With the term "Stratigraphic Miso-Anthropocene" I use the prefix "miso-", from the Greek "miseo" (to hate).

scholars pertain more or less to the *Good, Great* or even *Beautiful Anthropocene* and include the geographer Erle C. Ellis, the journalist Andrew Revkin and, generally speaking, the Ecomodernists. In their eyes, the Anthropocene as both geological and historical aspirant epoch does represent a newness, but one no different than any other historical or geological discontinuity. In other words, we are not facing a genuine "*epistemological rupture*" (Bachelard), a "*paradigm shift*" (Kuhn) or even a radical rupture. The Anthropocene may be new, but it is not an authentic "uniqueness" (Hamilton and Grinevald). It is important to emphasize that this normalizing approach is wound up with and functional to an optimistic vision of anthropocenic challenges, the ecological one first and foremost. Whether the objective is to return to "business as usual" by tweaking holocenic parameters (such as environmental scientist Johan Rockström does with the so-called *planetary boundaries*[12]) or to welcome true epochal change (according to the Ecomodernists), these "normalizers" firmly believe that the human being (i.e., as *homo faber*, or better, *homo oeconomicus-technologicus*), thanks to its agency (i.e., to the constant growth of the anthropic/technological factor within the natural context), will be capable of fulfilling the role that destiny has assigned to it: Steward or Manager of the Earth System. At the heart of this unwavering optimism, this techno-*philia* that turns out to be techno-*latry,* lies the imperative "*Love your Monsters*": the moral of Mary Shelley's *Frankenstein* (the modern Prometheus) according to Bruno Latour's re-reading. Latour argues, "our sin is not that we created technologies but that we failed to love and care for

[12] Rockström, the executive director of the *Stockholm Resilience Centre,* lists nine parameters (i.e. "climate change; ocean acidification; stratospheric ozone; biogeochemical nitrogen (N) cycle and phosphorus (P) cycle; global freshwater use; land system change; the rate at which biological diversity is lost; chemical pollution; atmospheric aerosol loading") which establish the actual distance between the current epoch and the Holocene. They "define, as it were, the boundaries of the 'planetary playing field' for humanity if we want to be sure of avoiding major human-induced environmental change on a global scale" (Rockström *et al*. 2009). On the topic see also Rockström and Gaffney 2021.

them. It is as if we decided that we are unable to follow through with the education of our children." (Latour 2011.)[13]

The key feature of this ideology is trust in geoengineering and its radical solutions, though with "techno-latry" it would probably be better to talk of faith.[14] In the eyes of these scholars, geoengineering represents the ideal tool for achieving planetary management, that is for "decoupling human development from environmental impacts" and thus allowing "for a good, or even great, Anthropocene" (Asafou-Adjaye *et al.* 2015, 7).

3. The Anthropocene is something peculiar, but for strictly epistemic reasons. Adherents to this perspective, which include the aforementioned Dipesh Chakrabarty and Clive Hamilton, are more worried than others about the problems currently posed by the Anthropocene, problems that will only grow as time passes. Chakrabarty holds that the Anthropocene forces us to rethink certain traditional categories of historical thought, namely to go beyond the distinction between human history and natural history. According to Chakrabarty, we need a "new universal history of humans that flashes up in the moment of the danger that is climate change". But it cannot be a Hegelian universal, i.e. a universal which subsumes particularities. Instead, our current situation "calls for a global approach to politics without the myth of a global identity […] We may provisionally call it a 'negative universal history'." (Chakrabarty 2009, 220.)

The Anthropocene is therefore an event capable of producing a veritable crisis in the foundation of history as knowledge. That is why I've called his approach *Historical Anthropocene*. Hamilton, on the other hand, author of one of the most philosophically solid investigations on the subject, highlights the Anthropocene's "*uniqueness*", namely its status as a radical "*rupture*" with "no precursors".[15] In his view, this rupture became possible only after the birth of a new epistemic paradigm: *Earth System Science* (ESS), and anyone who deals with the anthropocenic question from outside this par-

[13] On Latour and the Anthropocene see Latour 2017, in particular 111–145.
[14] On geoengineering in the Anthropocene see Baskin 2019.
[15] On Anthropocene's uniqueness see Hamilton and Grinevald 2015; Hamilton 2016.

adigm literally does not know what he/she is talking about. Because of its radicalness, I call Hamilton's proposal *Hard Anthropocene*, which at times leans towards an *Esoteric Anthropocene*: an almost "initiatic" approach where only a select few have access to true knowledge of the phenomenon. They are the Earth System scientists: the priests of the anthropocenic cult.

3. The Stratigraphic Fallacy

Now I would like to clarify my own position about the Anthropocene's epistemic peculiarity. I will do so by discussing what I consider to be an epistemically emblematic interpretation of it. I refer to the geological interpretation presented in Carlos Gray Santana's recent article *Waiting for the Anthropocene*, 2019 winner of the prestigious "Karl Popper Prize" established by *The British Journal for the Philosophy of Science*. Santana opts for what I have called *Stratigraphic Miso-Anthropocene*. By this term I mean an approach that rejects the anthropocenic hypothesis on the basis of purely stratigraphic evidences. Adopting a philosophy of science perspective, Santana argues that "formal recognition of the Anthropocene should be indefinitely deferred" (Santana 2019, 1073), namely it "should remain informal" (Santana 2019, 1075).

In so doing, he rejects the AWG's 2016 recommendation to formally recognize the Anthropocene. In my view, Santana's perspective can be considered the natural *pendant* of Zalasiewicz's one. Both firmly believe that the Anthropocene's legitimacy as an aspirant new epoch depends entirely on its *scientific* evidence, which in turn equates to its *geological*, or better *stratigraphic*, evidence. Yet they arrive at opposite conclusions: Zalasiewicz's Stratigraphic *Philo*-Anthropocene claims the stratigraphic legitimacy of the new epoch, while Santana's Stratigraphic *Miso*-Anthropocene supports its illegitimacy.

Santana's thesis is based on two arguments. The first is a "synchronic perspective" which sees "the possible existence of the Anthropocene as a question about our current relationship to our planet" (Santana 2019, 1076), namely a potential justification on the basis of its political utility. He argues that "we have little evidence that ratifying the Anthropocene will

have the political effect its proponents suggest" (Santana 2019, 1077).

The second argument, a "future geologist's perspective", is a predictive justification of the Anthropocene. On this point he says that "we should hold off on formal approval of the Anthropocene because extant geological changes don't reach the thresholds necessary to define a new epoch" (Santana 2019, 1077), that is one different from the Holocene.

I will briefly discuss these two arguments in order to emphasize what I consider to be the aporia of Santana's approach: his *Stratigraphic Fallacy*. Santana constructs his argument with competence and clarity by way of the following thought experiment. Given that the proposed epoch is currently in progress (or, better, only just beginning), we cannot use the normal stratigraphic methodology. We must thus "consider the question 'from the viewpoint of a geologist viewing sequences [of rock] thousands or millions of years in the future'. Would that future geologist see justification for driving a golden spike into a rock layer that was formed around the twentieth century?" (Santana 2019, 1076.)

Despite his skepticism – which stems from the fact that this experiment forces geology, a *historical science*, to become a *science of prediction* – Santana accepts the challenge. After examining the most commonly proposed markers of the Anthropocene (climate change, human and non-human fossil record, direct anthropogenic deposits, chemical markers, hydrology), he presents three reasons for which "we are not yet justified in claiming that current human geological activity will rise to epochal significance from the future geologist's perspective" (Santana 2019, 1078).

These are: 1) "many of our geological impacts can be mitigated by future human behavior. To the future geologist, this may make them relatively insignificant, brief anomalies"; 2) "Some anthropogenic activities are best conceived as continuations of processes that originated in the Holocene […] Thus, those markers go back into the Holocene"; 3) "Many clear examples of human impact will be seen by the future geologist as local catastrophes rather than geologic events of global reach and long-term impact." (Santana 2019, 1078.)

The first of these reasons is the most important, and from it the stratigraphic fallacy clearly emerges. In my view, Santana

shows a basic misunderstanding of the Anthropocene's ontological characterization, one that Crutzen never fell for. As early as his anthropocenic manifesto in 2000, in fact, Crutzen saw the Anthropocene as a situation in which "humans and our societies have become a global geophysical force", in which "human activities have become so pervasive and profound that they rival the great forces of Nature and are pushing the Earth into planetary *terra incognita*." (Steffen and Crutzen *et al.* 2007, 614.)

This means that the Anthropocene's uniqueness (i.e. its peculiarity as an aspirant geological epoch) doesn't depend on the fact that the human being irreversibly alters the homeostasis of the Earth System with its technological agency, but that it produces an irreversible stratigraphic discontinuity. In other words, even from an epistemic-scientific perspective the crucial element has nothing to do with the emergence of a clearly recognizable "anthropogenic golden spike", so to speak, but with the *order of magnitude* that the anthropic variable reaches for the first time. For the first time human agency has become a *technological omni-power*.[16] And that is exactly the point. As technological omni-power, our agency becomes a decisive condition of possibility for the destiny of the Earth, even from a geological perspective. The human being may or may not produce such a stratigraphic discontinuity (the

[16] There is not enough space here to explain this term, which is part of my much longer work aimed at establishing a *Philosophy of Technology in the Nominative Case* (TECNOM). On this topic see Cera 2017 and 2020. So let it suffice that by "technological omni-power" I mean the absolutization of our technological agency, namely the conviction that we are now able to make anything we want to make, which is a kind of omnipotence. In my view the main outcome of this absolutization is the definitive metamorphosis of potentiality (the Greek *dynamis*) into power (the German *Macht*). More clearly, it equates to the idea that the only possible way of acting/doing is the technological one. On this basis, I consider technological omni-power to be the entelechy of modern anthropocentrism: the culmination of the world's disenchantment (according to Max Weber), the definitive proof that technology represents the current "subject of history" (according to Günther Anders). As Peter Haff says, insofar as technology has become *de facto* "a geological phenomenon", we are living in a "technosphere" (see Haff 2014). A similar approach can be found in Yuk Hui's idea of "cosmotechnics" (see Hui and Lemmens 2021).

aforementioned anthropogenic golden spike), but both situations would be the result of its action, including non-action: its eventual choice, that is, to limit itself, to not use all of its omni-power, to avoid a full challenging of the ecosystem. From both human history and natural history, the human being has for the first time become a "geologic agent" (Wilkinson),[17] "geological/geophysical force" (Crutzen and Steffen) or "geological power" (Hamilton). It is able to "rival the great forces of Nature", namely it is in the order of magnitude of those forces. And this is true regardless of whether we give rise to a Good or Bad Anthropocene.

Geology – or at least Santana's idea of geology – does not seem capable of realizing this. It presumes to simply continue on as usual, limiting itself to the verification of what has happened by looking *exclusively* at the effects, not the causes or intentions. Yet even from a scientific point of view the level of effects proves to be an insufficient dimension within the current, brand-new epochal framework. Faced with the Anthropocene hyperobject, science can no longer ignore the intentional dimension. Given that technological development makes us increasingly "capable of doing what we want", of realizing our goals and *desiderata* (or, better, given that our technological omni-power forces us into this position), it is precisely on the level of intentions that the human being's quantum leap in its relationship with nature (world, environment, ecosystem, Earth System…) should be measured and evaluated. Incidentally, the idea of Earth System,[18] with its new paradigm of Earth System Science (ESS), represents a good example of science's effort to acknowledge this radically new situation and *scientifically* respond to it.

It is worth repeating that the truth of the matter, even from a strictly geological standpoint, is not that the possible eco-

---

[17] See Wilkinson 2005.

[18] According to Crutzen and Steffen, the Earth System is "the suite of interacting physical, chemical and biological global-scale cycles and energy fluxes that provide the life-support system for life at the surface of the planet […] the Earth System includes humans, our societies, and our activities; thus, humans are not an outside force perturbing an otherwise natural system but rather an integral and interacting part of the Earth System itself." (Steffen and Crutzen *et al.* 2007, 615.)

logical damages or geological-stratigraphic alterations are anthropogenic, but that the eventual solutions will almost certainly be anthropogenic too. The destiny of nature is now literally in our hands. From here on out, human agency as technological omni-power will increasingly be both the problem and the possible solution, the illness and the cure. Here the "pharmacological" status of contemporary technology – according to a Stieglerian interpretation of it – emerges, further confirmation that it equates to an *integral epochal phenomenon*. Like it or not, the human being will increasingly play the role of *arbiter* (i.e., steward, manager) of the destiny of the natural environment to which it belongs.

The most emblematic example of this is once again that of geoengineering solutions as corrections to the ecological imbalances. These proposals show that any future restoration of a "natural" or "normal" evolution of the Earth System will come from a technological-artificial intervention, an increase in anthropic meddling within the natural environment. Geoengineering – "the most Promethean form of environmental governance" Ellis (2018, chap. 8.5) – appears as the perfect tool for achieving planetary management, the form taken by technology when it presumes to play the part of nature as well. Potential problems, it is held, can only be solved by an increase in human agency in regards to nature. The general principle can be summarized as such: against technology (i.e. its excesses), more technology is needed. As a result, geoengineering is not only "techno-fix" (Ellis 2018, chap. 8.5). It is *Techno-care*.[19] The driving force of these approaches is no longer a simple "*love* your monsters" (according to Latour), but "*worship* your monsters", a fetishizing and idolatry of them. It is the metamorphosis of technology into an *idolum*, namely something which requires faith and promises salvation. Geongineering solutions are *de facto* declarations of this kind of faith. *Techno-care's basic assumption is techno-latry.*

The same argument can be used in the case of non-action. Even if the human being decides not to make a geologically decisive mark on the course of the Earth System by limiting

[19] An additional emblematic example in this sense can be found in so-called *rewilding* or conservation biology, which puts a *technological re-naturalization of nature* into practice.

its own agency/challenging, this "neutral" result would for the first time depend entirely on its own choice. This "non effect" would still be a product of its action, its assumption of responsibility. Traditional geology, on the other hand, doesn't seem capable of taking the effective value of this possible abstention into account, namely the real consequences of the human being's eventual choice not to exercise its technological omni-power.

In some distant future, it could very well be that geology, looking only at stratigraphic evidence, would not claim the advent of an Anthropocene because we had gotten things under control in the meantime, reversing the destructive tendency of our impact on the Earth System. This is the ecomodernists' hope: to manage the planet by means of "knowledge and technology, applied with wisdom" (Asafou-Adjaye *et al.* 2015, 6). Yet from a conceptual, epistemic point of view, such a situation would not be, as Santana argues, an objection against the Anthropocene's existence. It would on the contrary be a crucial confirmation of it, because – it bears repeating – that (re-)normalized, (re-)naturalized stratigraphic situation would depend exclusively on our decision and thus our responsibility. Whether the soil – i.e., the flesh of the Earth System's body – continues to tell the same (old) story of the Holocene or a radically new one will be largely an anthropogenic outcome. This shows that stratigraphic evidences need a historical-anthropological etiology to be correctly understood and interpreted, a hermeneutic. Here lies the paradigmatic novelty of the situation that Santana – who insists precisely upon geology's character as "historical science" – proves unable to recognize. If Earth's destiny will be increasingly in our hands, even from a geological point of view, and if geology is incapable of recognizing this using its traditional methods, then the question becomes: does this say something about the Anthropocene or about geology? Isn't this totally unprecedented scenario an indication that geology should put its own foundation as a science up for discussion? Confining itself to the negation of the anthropocenic hypothesis, hiding in its own epistemic coat-of-arms while claiming a priori the efficacy of its traditional methods and approaches – isn't this too simple a response to a question of such magnitude?

I am not a geologist, and nor am I a philosopher of science, but I think I can state that to confine oneself to stratigraphic evidences when confronting such a complex phenomenon is no longer sufficient. *The epistemic ambiguity of the anthropocenic idea gives birth to an epistemic crisis in geology*, and ignoring the signs of this potential crisis is not the best way to address the challenge. Precisely in order to continue being a historical science, to avoid becoming an "*antiquarian* science" (in the Nietzschean sense of the word),[20] geology should probably learn to read and recount, scientifically, a different story.[21]

3.1 The future geologist or "The Man without a Past"

The second element of Santana's stratigraphic fallacy is the anthropological *Naivität* which clearly emerges from his portrait of "the future geologist": the protagonist of his mental experiment.

My objection is the following. Is Santana's future geologist a credible example of a human being? I mean, is it plausible to imagine a future scientist confining him or herself to an exclusive – not to mention stubborn – observation of strati-

---

[20] Here I refer to Nietzsche's *Second Untimely Meditation*, that is its distinction between "monumental", "antiquarian" and "critical" approaches to history. In particular, one "who likes to persist in the familiar and the revered of old, tends the past as an antiquarian historian" (Nietzsche 1997, 72).

[21] A reviewer of this paper deemed my idea of an epistemic crisis in geology caused by the Anthropocene to be excessive, and suggested I consider a very recent article by Jan Zalasiewicz (see Zalasiewicz *et al.* 2021) as a proof that geology is able to take points of view from other disciplines into account. I must thank the reviewer, for the article is indeed of great value, especially the last section where it deals with *Potential Acceptance and Utility of the Chronostratigraphic (Geological) Anthropocene Beyond Geology*. However, I have to remark that my point does not refer to a "simple" lack of interdisciplinarity, (i.e. to geology's capability to listen to other disciplines), but rather to geology's capability to open itself up (i.e. *qua* geology) to the new idea of history emerging from the Anthropocene (Hamilton's geohistory, for instance). Zalansiewicz's paper undoubtedly represents an encouraging first step in the right direction, but it cannot be considered the solution for the epistemic crisis I am trying to point out.

graphic evidence? Is it plausible to imagine that he/she would not know nor want to know anything about the non-stratigraphic causes of those stratigraphic evidences? About the fact that, from a certain point of history on, the geological and stratigraphic evolution of our planet is determined at least in part if not predominately by human choices, decisions and actions (abstentions included)? To paraphrase Feuerbach, we should remember that "even as a *geologist*, I am a man in togetherness with men" (Feuerbach 2012, 244). A geologist is an intrinsically historical entity, a *zoon historikon* (historical animal) to paraphrase Aristotle. This being said, it is difficult to imagine Santana's future scientist as an unhistorical being, an animal which, "fettered to the moment" (Nietzsche 2007, 60),[22] does not know nor wants to know anything about the general context in which a given geological-stratigraphic evidence has developed.

If geology is a historical science, then the geologist can be nothing but a *historical scientist*. Yet Santana's portrait of the future geologist closely resembles the protagonist of Aki Kaurismaki's *The Man without a Past*. As a result, Santana seems to be a victim of that *anthropological misunderstanding* which tries to reduce the human being to a *rational agent*, to convince us that we are – or worse, should become – rational agents: mere certifiers of (presumed) evidences, worshippers of facts in the form of *data*. I would even call it an *anthropological perversion*. The rational agent is only capable of acting and thinking in a *rational* manner, with "rational" understood (and lessened) as a synonym of "rationalized" that is utilitarian, computational, executive... By definition, the rational agent is a "*calculator*" in all meanings of the word. It is "a machine *in potentia*". This parody of the human phenomenon cannot be considered an authentic "agent", or better, it is an agent only insofar as agency refers to the execution of programs and procedures and not to "action".[23] The rational

[22] According to Nietzsche "the animal lives unhistorically".

[23] I refer here to Hannah Arendt's interpretation of "action" as the defining feature of a *vita activa*. One could say that, according to Arendt, the difference between human being and *animal laborans* depends on the idea of "action". That is, the human being is different from the *animal laborans* as long as it is able to authentically act. (See Arendt 1958, 175–247.)

agent is nothing but a performer, an *executor*. More precisely, the ideal executor: someone whose only interest is the ascertainment/certification of facts and execution of programs; someone whose thirst for knowledge is quenched by the mere acknowledgment of reality as it is. Historical sensibility would sound foreign to such a human type, as would its most direct consequence: the hermeneutic attitude, the uniquely human capability (and even necessity) to interpret and comprehend phenomena. That is to say, not only to acknowledge or certify them, but to give them meaning and sense.

This amputated form of human being (i.e. as rational agent or "man without a past") is the only way a future geologist could limit himself/herself to certify that, stratigraphically, nothing changed from the Holocene on. The only way he/she could make such an assertion ignoring that, from both human and natural history, humankind (*anthropos*) had found itself for the first time in the position to produce such an outcome, to assume such a responsibility. If the future geologist were to take these facts into account, he/she would be forced to claim that the geological-stratigraphic phenomenon at hand is definitely an anthropogenic – and therefore potentially anthropocenic – phenomenon.

Only two catastrophic, post-apocalyptic scenarios could validate Santana's historical and hermeneutical reluctance as expressed by his future geologist's blindness toward phenomena. Let's imagine a future geologist who knows nothing about his/her ancestors and their history because a catastrophic event had annihilated all traces of them, cutting the string of history. Such an event could either be of natural or anthropogenic origin: a nuclear or ecological holocaust. But an anthropogenic catastrophe would almost certainly leave some stratigraphic evidence, thus proving a geological discontinuity and confirming the Anthropocene *de facto* (though this particular future geologist would not be capable of recognizing it as such). As a consequence, only a catastrophic event of natural origin, one so strong as to wipe out all trace of humanity, could represent a sufficient precondition for the validation of Santana's thought experiment and its rejection of the anthropocenic hypothesis. Interestingly, this eventuality was already envisaged by Crutzen and Stoermer in their anthropocenic manifesto: "*Without major catastrophes* like an

enormous volcanic eruption, an unexpected epidemic, a large-scale nuclear war, an asteroid impact, a new ice age, or continued plundering of Earth's resources by partially still primitive technology […] *mankind will remain a major geological force for many millennia, maybe millions of years, to come.*" (Crutzen 2000, 18 – my italics.)

In conclusion, what unintentionally but clearly emerges from the stratigraphic fallacy of Santana's argument is that, *sic stantibus rebus*, the Anthropocene is already a reality/evidence *de facto*, even from a geological and stratigraphic point of view. Paradoxically, the real problem lies in geology's capacity to take this very particular reality/evidence into account, in its ability to adapt to a historical and epistemic paradigm in rapid and deep evolution.

Conclusion: An Epistemic Hyperobject?

At the basis of Santana's *Stratigraphic Miso-Anthropocene* – his rejection of the anthropocenic hypothesis – I have found a *Stratigraphic Fallacy* which demonstrates the difficulty geology has as a "historical science" in adequately exploring the anthropocenic phenomenon, and specifically in fully recognizing its newness.

This difficulty is further proof that the Anthropocene, with all its epistemic peculiarity, needs an *ad hoc* definition. This integral epochal phenomenon not only places itself beyond the "two cultures", but beyond Nature and Culture themselves: as a result, it needs to be approached in a brand-new way. That is why, to conclude this paper, I would like to suggest a possible epistemic definition of the Anthropocene. I will do so by using Timothy Morton's idea of hyperobject.

As is well known, Morton speaks of hyperobjects in order to rethink the ecological crisis and establish an incisive ontology of the present, which in his case is inspired by the so-called Oriented Object Ontology (OOO). In his view we are currently surrounded by phenomena that transcend the human dimension, such as the ecological crisis and the Anthropocene itself. These phenomena require an *anthropo-decentered perspective* to be comprehended. In other words, Morton agrees with Hamilton's aforementioned argument in favor of geohistory: "we are no longer able to think history as

exclusively human, for the very reason that we are in the Anthropocene." (Morton 2013, 5.) Morton defines hyperobjects as "objects that are so massively distributed in time and space as to transcend spatiotemporal specificity" (Morton 2010, 130–135).

The basic features of hyperobjects are: *viscosity* (they "'stick' to beings that are involved with them"); *nonlocality* ("any 'local manifestation' of a hyperobject is not directly the hyperobject"); *temporal undulation* ("they involve profoundly different temporalities than the human-scale ones we are used to"); *phasing* (hyperobjects occupy a high-dimensional phase space that makes them invisible to humans for stretches of time); and *interobjectivity* (they can be "detected in a space that consists in interrelationships between aesthetic properties of objects") (Morton 2013, 1).

As an epistemic entity – that is as an object of knowledge and discourse, and not only as an aspirant geological epoch – the Anthropocene seems to possess all these features. It is certainly "viscous", as any field of knowledge in the vicinity is inevitably attracted, with each proposing its own interpretation of the Anthropocene. It is certainly "nonlocal", as it is clearly everywhere and nowhere. It is certainly *temporally undulated* and *phased*, as it continuously alters its space-time framework according to the different thematic and disciplinary fields. Finally, as an *epistemic* hyperobject, the Anthropocene is not only inter-*objectual* but inter- and trans-*disciplinary*, as its epistemic peculiarity/ambiguity places it *ipso facto* beyond the two cultures. My hope is that this "beyond" becomes an authentic "with", namely not only inter- or trans-disciplinarity but *co-disciplinarity*: a real encounter, not the relationship of power between the two cultures which is currently at stake, with the minority status of humanities (and philosophy *in primis*) in regards to the hard sciences.

Like every hyperobject, the Anthropocene is "necessarily uncanny", meaning it produces "a feeling of strange familiarity and familiar strangeness." (Morton 2013, 55). Yet as an *epistemic* hyperobject it reveals its uniqueness, establishing itself as an *integral epochal phenomenon*: one of those discontinuities/singularities that "make history". This epistemic hyperobject presents a constitutive *historical barycenter* with the

adjective "historical" referring to a geohistory which brings together natural history and human history.

Given my assumption that overcoming the distinction between nature (*physis*) and culture (techne) represents the *anthropocenic Urphänomen*, namely the very condition of possibility for the Anthropocene as new epoch (not only geological), we define the Anthropocene as an *epistemic hyperobject with a (geo-)historical barycenter*.

University of Ferrara

References

Asafou-Adjaye, J. and L. Blomqvist *et al.* (2015), *An Ecomodernist Manifesto*. April 2015, http://www.ecomodernism.org/manifesto-english.

Arendt, H. (1958), *The Human Condition*, University of Chicago Press, Chicago.

Baskin, J. (2015), "Paradigm Dressed as Epoch: The Ideology of the Anthropocene", *Environmental Values* 24 (1), pp. 9-29.

Baskin, J. (2019), *Geoengineering, the Anthropocene and the End of Nature*, Palgrave Macmillan, New York.

Bonneuil, Ch. and J.-B. Fressoz (2016), *The Shock of the Anthropocene: The Earth, History and Us*, transl. by D. Fernbach, Verso, London and New York (Kindle edition).

Cera, A. (2017), "The Technocene or Technology as (Neo)environment", *Techné: Research in Philosophy and Technology* 21(2/3), pp. 243-28.

Cera, A. (2020), "*Lineamenti di una Filosofia della Tecnica al Nominativo (TECNOM)",* in M. Pavanini (a cura di), *Tecnica. Figure e strutture dell'artificio*, Kaiak Edizioni, Tricase, pp. 93-130.

Chakrabarty, D. (2009), "The Climate of History: Four Theses", *Critical Inquiry* 35 (2), pp. 197–222.

Clark, T. (2015), *Ecocriticism on the Edge: The Anthropocene as a Threshold Concept*, Bloomsbury, London, New Delhi, New York, Sidney.

Crist, E. (2016). "On the Poverty of our Nomenclature", in J. W. Moore (ed.) *Anthropocene or Capitalocene? Nature, History, and the Crisis of Capitalism*, PM Press, Oakland, pp. 14–33.

Crutzen, P. J., and E. F. Stoermer (2000), "The 'Anthropocene'", *Global Change Newsletter* 41, pp. 17–8.

DellaSala, D. A. and M. I. Goldstein (eds.) (2017), *Encyclopedia of the Anthropocene,* Elsevier, Oxford.

Descola, Ph. (2013), *Beyond Nature and Culture*, transl. by J. Lloyd, The University of Chicago Press, Chicago.

Dibley, B. (2012), "The Shape of Things to Come: Seven Theses on the Anthropocene and Attachment", *Australian Humanities Review* 52: pp. 139–153.

Ellis, E. C. (2018), *Anthropocene: A Very Short Introduction*, Oxford University Press, Oxford (Kindle edition).

Feuerbach, L. (2012), "Principles of the Philosophy of the Future," in *The Fiery Brook: Selected Writings*, transl. by Z. Hanfi, Verso, London and New York, pp. 175–245.

Gemenne F. and A. Rankovic (eds.) (2019), *Atlas de l'Anthropocène,* Presses de Sciences Po, Paris.

Goethe, J. W. von (1983), "Theory of Color", in *Scientific Studies (Goethe Collected Works vol. 12)*, transl. by D. E. Miller, Suhrkamp, New York, pp. 157–298.

Haff, P. K. (2014), "Technology as a Geological Phenomenon: Implications for Human Well-being", *Geological Society Special Publications* 395 (1), pp. 301–309.

Hamilton, C. and J. Grinevald (2015), "Was the Anthropocene anticipated?", The *Anthropocene Review* 2 (1), pp. 59–72.

Hamilton, C. (2016), "The Anthropocene as Rupture", *The Anthropocene Review* 3 (2), pp. 93–106.

Hamilton, C. (2017), *Defiant Earth: The Fate of Humans in the Anthropocene*, Allen & Unwin, Sidney, Melbourne, Auckland, London (Kindle edition).

Hui, Y. and P. Lemmens (eds.) (2021), *Cosmotechnics: For a Renewed Concept of Technology in the Anthropocene*, Routledge, London and New York.

Latour, B. (2011), "Love Your Monsters: Why we must care for our technologies as we do our children", in M. Shellenberger and T. Nordhaus (eds.), *Love your Monsters: Postenvironmentalism and the Anthropocene*, The Breakthrough Institute, San Diego (Kindle edition).

Latour, B. (2017), *Facing Gaia Eight Lectures on the New Climatic Regime*, transl. by C. Porter, Polity Press Cambridge (UK) and Medford (MA).

Morton, T. (2010), *The Ecological Thought*, Harvard University Press, Cambridge (Mass.) and London.

Morton, T. (2013), *Hyperobjects: Philosophy and Ecology after the End of the World*, University of Minnesota Press, Minneapolis.

Nietzsche, F. (1997), "On the Uses and Disadvantages of History for Life", in *Untimely Meditations*, ed. by D. Breazeale, transl. by R. J. Hollingdale, Cambridge University Press, Cambridge and New York, pp. 57–123.

Rockström, J. and W. Steffen *et al*. (2009), "Planetary Boundaries: Exploring the Safe Operating Space for Humanity", *Ecology and Society* 14(2), p. 32.

Rockström, J. and O. Gaffney (2021), *Breaking Boundaries: The Science of Our Planet*, Penguin Random House, London.

Ruddiman, W. F. (2003), "The Anthropogenic Greenhouse Era Began Thousands of Years Ago", *Climatic change* 61 (3), pp. 261–293.

Santana, C. G. (2019), "Waiting for the Anthropocene", *The British Journal for the Philosophy of Science* 70 (4), pp. 1073–1096.

Snow, Ch. P. (2012), *The Two Cultures*, Cambridge University Press, Cambridge (UK).

Steffen, W. and P. Crutzen *et al.* (2007), "The Anthropocene: Are Humans Now Overwhelming the Great Forces of Nature?", *Ambio* 36 (8), pp. 614–621.

Steger, M. B. (2009), *Globalisms: The Great Ideological Struggle of the Twenty-first Century* (third edition), Rowman & Littlefield, New York and Toronto.

Wilkinson, B. H. (2005), "*Humans as geologic agents: A deep-time perspective*", *Geology* 33 (3), pp. 161–164.

Zalasiewicz, J. and M. Williams *et al*. (2008), "Are We now Living in the Anthropocene?", *GSA Today* 18 (2), pp. 4–8.

Zalasiewicz, J. and C. N. Waters *et al*. (eds.) (2019), *The Anthropocene as a Geological Time Unit: A Guide to the Scientific Evidence and Current Debate*, Cambridge University Press, Cambridge (UK).

Zalasiewicz J. and C. N. Waters et al. (2021), "The Anthropocene: Comparing Its Meaning in Geology (Chronostratigraphy) with Conceptual Approaches Arising in Other Disciplines", *Earth's Future* 3 (9).

The Non-Identity Problem and Its Harm-Based Solutions[1]

SIMO KYLLÖNEN

1. Introduction

We know that many of our actions have a great effect on how well- or badly-off some future people are. For instance, by deciding to continue using fossil fuels as we do today, we may pollute the climate and make the life of very many future people miserable. Yet, according to Derek Parfit's infamous Non-Identity Problem (NIP)[2], our decision does not harm those badly-off future people of the polluted climate, if that same decision is also a necessary condition of the very existence of those future people.

The NIP is a problem that results from a set of common-sense assumptions which most people find intuitively plausible but which yield unacceptable conclusions in identity-affecting future-related cases, such as a choice related to future climate pollution. According to these assumptions, we hold, for instance, that an action (or inaction) harms a person only if it makes that person *worse off* in some respect than the person would have been had the action not been performed. Yet choices like the one of increasing pollution do not make the future people who will suffer the polluted climate worse

[1] I want to thank the two anonymous reviewers for their careful and constructive reviews. The article also benefited from input from participants of the FiPhi 2020 Conference and the Philosophy Research Seminar of Tampere University. This work was supported by the Strategic Research Council at the Academy of Finland (grant numbers 312671/326662).

[2] The problem seems to have been discovered by several authors in the late 1970s (e.g., Schwartz 1978) but it was Parfit's treatment of the problem in *Reasons and Persons* that popularised the name Non-identity problem. See also Boonin (2014).

off, because had we acted otherwise, that is, stopped pollution, these future people would never have even existed.[3] This is so, because the decision to use fossil fuels over renewables will eventually have an impact on whom people meet during their lives, with whom they will have children, when they will have children, how many children if any, etc. In the long run, we can quite safely assume that these small differences in individual parental choices will result in a whole set of different people existing than would have been the case had we selected the swift change to renewable energy and prevented climate pollution.

The implausible conclusion of the NIP – that we do not harm future people by polluting the climate – rests on the fact that, although our pollution is likely to make future people very much badly off, it does not make them worse off.[4] In this paper, I investigate solutions to the NIP which do not deny this, but which argue for an alternative understanding of harm. I suggest that the so-called harm-based accounts are unaffected by the NIP, because according to them the identi-

---

[3] Claiming that we have made the future people of polluted climate worse off by our actions would imply that we were able to compare the state of these people to their non-existence, and then conclude that these people would have been better off if they had not existed at all. Given that the future world of polluted climate is not so terrible that it would make the life of the people of that world dubiously worth living, such a claim would be highly implausible.

[4] There are also other assumptions related to the NIP: many people may also find it plausible that one can wrong a person only if one harms the person. In that case, the implausible conclusion of the NIP would be that we do not wrong future people by polluting the climate. Boonin's (2014) treatment includes these assumptions about the relation between harm and wronging, and he critically evaluates solutions to the NIP that deny that there is a necessary relation between them. These solutions aim to show that even if future people are not harmed in non-identity cases, they are wronged in some other way, e.g., by violating their rights. In this paper I focus purely on the conclusion at the level of harm, which I assume to be already implausible enough. That an act harms gives us pro tanto compelling reason against the act. If it can be established that, contra the NIP, harm occurs in non-identity cases that supports also our judgment that polluting the climate is (morally) wrong if harm is also wronging. Whether harming someone is the only feature that determines whether we also wrong the harmed one is left open here.

fication of harm does not require us to be able to compare the state of an individual to her better-off state in a situation that would have obtained in the absence of the harmful action.

However, harm-based accounts have been criticised as being both over- and under-inclusive. On the one hand, harm-based accounts judge some acts as objectionably harmful even if they are clearly not so. On the other hand, their scope seems to be too narrow, since they cannot account for some intuitively plausible cases of harm.

Together these objections weaken the solution that harm-based accounts claim to offer against the NIP. According to Boonin (2014), a successful solution needs not only to offer a way to avoid the implausible conclusion of the NIP by rejecting some of its seemingly plausible assumptions – such as the harm-based accounts' rejection of the comparative *worse off* notion of harm. In the successful solution, the proposed rejection of an assumption that generates the NIP must be sufficiently (i) *independent* from our aim to avoid the implausible conclusion, (ii) *robust* to be able to offer a solution to any weakened version of the rejected assumption that would still generate the NIP's implausible conclusion, and (iii) *modest,* that is, not yielding conclusions that are even more implausible than the conclusion of the NIP. Boonin's general objection against harm-based accounts as a solution to the NIP is based on their inability to fulfil these requirements. They do not offer a sufficiently independent account of harm that would be able to block the appeal to any weakened version of the comparative notion (still able to generate the NIP). And in the cases that they seem to do so, they yield conclusions that are even more implausible than the one resulting from the NIP.[5]

---

[5] Such would be the case if the defender of the harm-based account accepts that harm makes also the act morally wrong and then argues that the case of *Surgery* discussed below shows that the patient is harmed even though she is made better off overall in the comparative sense. This would yield the notion that the surgeon is also wronging the patient, which Boonin takes to be even more implausible than the one that results from the NIP. But here a defender of the harm-based solutions might respond that even if the fact that an act harms someone provides a reason against it, the reason is only pro tanto and can be overridden by some other reason, and it is the balance of the overall "conclusive" reasons that makes the acts right or wrong (see e.g., Scanlon 2007).

In this paper, I defend a specific harm-based account against these objections and argue that a suitably qualified disjunctive understanding of harming, called the *Additional reasons* view, offers a plausible and sufficiently robust way to account for harm both in ordinary as well as in non-identity cases. This defense is organised as follows. First, I present the Non-Identity Problem and how it has been suggested that harm-based accounts of harming offer a solution to the NIP. In the third section, different interpretations of the non-comparative version of the harm-based approach are discussed and evaluated against various same individual and non-identity cases. The section ends by suggesting a disjunctive understanding of the non-comparative account, and the fourth section evaluates different ways to interpret that understanding. In the following section, I defend the *Additional reasons* interpretation of the disjunctive against Boonin's objections. Finally, I evaluate whether the claimed incompatibility of the *Additional reasons* view with Derek Parfit's No-Difference View is a serious objection against harm-based accounts. I conclude that if someone finds the incompatibility implausible, it is less so than the conclusion of the NIP.

2. The Non-Identity Problem and different notions of harming

The NIP is based on an understanding of harm that we often find intuitively plausible. According to this understanding, what could be called the *counterfactually comparative* notion of harming:

> HARMING I: An action (or inaction), A, harms an individual, S, if and only if, had A not occurred, S would have been better off in some respect.[6]

The implausible conclusion of the NIP may become clearer if we shift the choice from the indirect collective level of climate policy to the decisions of individual mothers concerning their

[6] I adopt the useful formulations in the paper mostly from Gardner (2015, 2017, 2019) but focus only on actions or (inactions) while her more general formulations also allow that events can harm and actions (inactions) harm by causing events that harm.

children to be. Consider the following example provided by Parfit (1986):

> *Two women.* While Carla is pregnant, she learns that, unless she takes some treatment, there is a risk that her child may have a certain severe and irreversible disability. She decides not to take this treatment. As a result, her child, Carl, is seriously disabled.
>
> While Paula is trying to become pregnant, she learns that, if she conceives a child now, there is a risk that they may have the same severe and irreversible disability. If she waits two months before conceiving the child, there would be no such risk. She decides not to wait. As a result, her child, Paul, is seriously disabled.

It seems clear to us that Carla harms Carl by not taking the treatment and HARMING I explains why this is so: Carla makes Carl worse off than he would have been had she decided to take the treatment. But most people would like to say the same about Paula's choice: as a result of her decision, Paul likewise suffers a similar severe disability. But this intuition cannot be explained by referring to HARMING I due to the NIP. Had Paula waited two months, Paul would not have existed, and thus he is not made worse off in accordance with HARMING I.

Thus, in order to explain our intuitions that Carla and Paula harm their children equally, and that they equally had a harm-based reason against their decision, we need an alternative to HARMING I. Parfit's original solution to the NIP was to appeal to a principle Q that allows us to compare the state of Paul to the counterfactual state of *another* future child of Paula's that would have existed had she waited two months (Parfit 1984, 360). As a notion of harming, Parfit's principle could be formulated as follows:

> HARMING II: An action (or inaction), A, that brings an individual S into existence, harms S, if and only if A causes S to be *worse off* than *another individual* who would have existed had A not occurred.

Parfit's principle Q has faced a lot of critical attention (e.g., Hanser 1990; Woodward 1986, 1987; Woollard 2012). It has been noted, for instance, how this principle would imply that

we could harm a person whom we have brought into existence whenever we could have brought into existence another person who could have been better off than the person whose existence we actually brought about. Accordingly, parents would be harming their perfectly healthy and well-off child, had it been possible to them to have a child who would have had an even higher level of wellbeing. On the other hand, consider the following example (Parfit 1986, 861; see also Woodward 1986, 816):

> *Parents*. Petra and Paul are considering having a child but they know that there is a risk that any child of theirs would have a severe and irreversible disability. Though they know this, they decide to have a child. As a result their child, Peter, is born with the disability.

According to HARMING II, there is nothing objectionable with Petra's and Paul's choice as long as the disability, though serious, is not such that it makes Peter's life dubiously worth living (cf. Parfit 1983).

Yet, both of these conclusions implied by HARMING II seem to go against what many people would think in these situations (cf. Woodward 1986). Moreover, the comparative notion of HARMING II does not seem to capture the features that might explain our reactions.[7] One such feature seems to be connected to the undesirability and badness of the state of the children that results from parents' decision. In considering whether parents have acted in a harmful way when they decided to have the child they actually have, the crucial thing appears to be how badly off they could have foreseen their forthcoming child was likely to be, and not the possibility of

---

[7] Parfit has later admitted that in such cases there might be reasons for objecting to the mother's choice other than those provided by the principle Q (Parfit 1984, 358; 1986, 860). However, he also claims that principle Q is still needed to capture the difference we make when comparing Petra's choice to the choice of a mother who could have a child without a disability if she only waited some months (cf. Paula in the *Two Women* example) (Parfit 1987, 816). In his last, posthumously published, paper Parfit (2017) propose a new way to ground principle Q by a "wider theory" which he believed to be found in the principle that combines effects at the collective level to all and at the individual level to each.

having another better-off child. If the child is likely to be seriously badly off, as in the case of *Parents*, this clearly offers a reason – sometimes even the conclusive one – not to have the child (cf. Woodward 1986, 816).

The harm-based accounts of harming point directly to this feature of harm as a resulting bad state. According to these accounts:

> HARMING III: An action (or inaction), A, harms an individual, S, if and only if A causes a state of affairs that is a harm for S.

HARMING III is a general form of harm-based accounts. The central contrast between the counterfactually comparative HARMING I and HARMING III is that, according to the latter, actions can harm even if the harmed person would not have been made worse off than she would have been had the action not occurred. The harm-based accounts are thus unaffected by the NIP, because according to them the identification of harm does not require us to compare the state of the harmed person to her better-off state in a situation that would have obtained in the absence of the harmful action. Instead, HARMING III requires that we be able to identify the effects of the action that constitute a harmful state for the affected person in a relevant sense, regardless of whether that person would have been better off in the absence of the action.

In this sense, harm-based accounts are "effect-relative" rather than "action-relative", as Molly Gardner (2015, 2017, 2019) describes. Actions or events are harmful by virtue of their effects on the individuals – in terms of the resulting states of affairs for them – rather than by virtue of the difference that the actions make. Being "effect-relative" in this way allows HARMING III to provide us with reasons to say that parents who have decided to have a child who is well-off, but who is not necessarily the best-off child they could have had, have not harmed their child. No bad state of affairs for their well-off child results from their decision although they could possibly have had a child who was even better off. Petra and Paul, in contrast, harm Peter (at least in one sense), because as a result of their decision a bad state of affairs, a severe and irreversible disability, obtains that is a harm for Peter.

Yet various harm-based accounts differ in how they define a harm.[8] The purely *non-comparative* accounts define harm as a *bad* state of affairs for the affected individual:

> *Non-comparative* HARMING III: An action (or inaction), A, harms an individual, S, if and only if A causes a state of affairs that is a harm for S.
>
> HARM (*non-comparative*): A state of affairs, T, is a harm for an individual, S, if and only if T is a *bad* state of affairs for S.

Accounts within the non-comparative camp then differ from each other in how they define a bad state of affairs. Elizabeth Harman (2004, 2009) offers a list of bad states.

---

[8] In addition to the non-comparative accounts discussed in the paper, Gardner (2015, 2017, 2019) has suggested a so-called *existential* account that offers a harm-based alternative to the non-comparative understanding. According to this account, an action (or inaction) harms someone if it causes a state of affairs that detracts her from the well-being she would have had if she existed and the state of affairs did not obtain. Gardner's existential account would yield results like the *Additional reasons* view defended in this paper (including the rejection of the No-Difference View), but without any reference to comparative notion. Moreover, the account needs no predefined understanding of what counts as a *bad* state of affairs for an individual. A potential weakness of the existential account is that it needs to reject several metaphysical claims about causation and counterfactuals that many find highly plausible (see Gardner 2019). Those claims hold, for instance, that counterfactual dependency is a sufficient condition for causation. But if counterfactual dependency is accepted as a sufficient condition, then, according to the existential account, Paula would not only harm Paul in *Two Women* by causing him a severe disability; she would also cause all the harms suffered by Paul over his entire life. Furthermore, in order to explain why, e.g., in *Physician,* the patient is not harmed, the existential account needs to accept that some "backtracking" counterfactuals are true: if the patient existed and state of affairs in which she has dim vision did not obtain, then the physician would not have operated and the patient would have been worse off in some respect. In the *Additional reasons* view, potentially only the first claim about the counterfactual dependency needs to be weakened: without doing so, parents' act of conceiving their child would seem to count a sufficient counterfactual cause of their children's death, which would not have happened without the act of conception (see Boonin 2014, 253–254). The limitations of this paper do not allow me any deeper discussion about these claims, neither of which evaluate the existential account properly.

Those include "pain, early death, bodily damage, and deformation" (2004) as well as "mental and physical discomfort, disease, or disability" (2009). On Seana Shiffrin's view, "harm involves conditions that generate a significant chasm or conflict between one's will and one's experience, one's life more broadly understood, or one's circumstances" (Shiffrin 1999, 123). Lucas Meyer (2003, 2008) and Eduardo Rivera-Lopez (2009) defend a threshold-view, according to which there is a morally relevant threshold, and an action (or inaction) harms someone only if the agent thereby causes the individual to be in a sub-threshold state.

In what follows, I will evaluate the purely non-comparative accounts of the harm-based approach and the objections raised against them. I conclude the section by suggesting a disjunctive interpretation that is able to respond to the objections.

3. Non-comparative harming and the need for a disjunctive notion

There are several objections that have been raised against the *non-comparative* accounts of HARMING III. The first type of objection claims that the non-comparative accounts are *too inclusive*, that is, the accounts judge acts as objectionably harmful that seem clearly not to be so. Consider, for instance, the following case presented by Harman (2004, 91):

> *Surgery.* A surgeon cuts a hole in my abdomen in order to remove my swollen appendix. Cutting open my abdomen causes me pain (as I recover); but if the operation had not been performed, I would have suffered worse pain and died very soon.

According to the objection, a defender of the non-comparative notion of HARMING III is forced to admit that whatever the surgeon would do will count as harmful even if she has improved my condition as much as she can. This clearly seems to go against the view that many find intuitively appealing here: the surgeon does not harm me, at least in the sense that the harm would count as a compelling rea-

son against the surgery.[9] One way for a non-comparative account to respond to such cases is to accept the outcome that the surgeon is indeed harming me – by causing the bodily damage resulting from the incision in my abdomen and the post-operative pain as I recover – but then regard the harms that she causes as *permissible*. Shiffrin (1999) and Harman (2009), for instance, accept that harms that are caused in order to *prevent* even greater harm are permissible. Central to this justification is that it is limited only to cases in which the act is performed as a necessary means to prevent a greater *harm,* but not in which the act confers "purely" greater benefits on the one harmed. Thus, in *Surgery* the surgeon causes a permissible harm to her patient, while performing an equally serious and painful surgery would not qualify as a permissible harm "even if necessary to endow [her patient] with valuable, physical benefits, such as supernormal memory, a useful store of encyclopedic knowledge, twenty IQ points worth of extra intellectual ability, or the ability to consume immoderate amounts of alcohol or fat without side effects" (Shiffrin 1999, 127).

This solution would still retain the non-comparative accounts' strength against non-identity cases, since the justification based on permissible harm would not be available to Paula (or to Petra and Paul in *Parents*), since Paul (and Peter) would not have suffered any greater harm had their parents acted differently, since in that case they would not even have existed. Moreover, allowing an appeal to the permissible harm only in cases of harm prevention works against those views that would try to justify Paula's (or Petra's and Paul's) decision with the great benefits that it brings to Paul (or Peter), namely all the benefits they would enjoy during their life, which is still worth of living despite the severe disability.[10]

---

[9] If harming is necessarily connected to wronging, the judgment that the surgeon would wrong me would appear even more implausible (see more Boonin 2014, 85).

[10] For an extensive discussion of the permissible harm and whether the defender of Non-comparative HARMING III can appeal to it in her solution to the NIP, see Boonin (2014, 74–92).

The non-comparative account might still remain over-inclusive, however. Consider the following case, adopted from Gardner (2015):

> *Physician*. A physician faces a situation in which she can improve the patient's blindness, but only to a level that still leaves the patient's vision dim.

If the resulting state of affairs, dim vision, is a bad state for the patient, then, according to the non-comparative HARMING III, the physician's operation should count as harmful, even if it improves the patient's condition. This would be the case in Harman's non-comparative account and in threshold views which hold that harmful states are those "that are worse in some way than the normal healthy state for a member of one's species" (Harman 2009, 139).[11] Perhaps again a defender of a non-comparative account could bite the bullet and count the operation as a permissible harm as it prevents the patient's future blindness. That appears to stretch the concept of harming too far, however.

Instead, to avoid the outcome, the non-comparative account can further qualify or limit the scope of what is regarded as a harm. Following Shiffrin (1999), a harm could be qualified as a state that is in conflict with the individual's will. Or, since this solution would potentially lead to the conclusion that non-human animals cannot be harmed unless they have a will, harmful states could rather be understood as states that are *seriously below* the normal healthy level, allowing dim vision not to be counted as a harm.

However, these qualifications make the non-comparative accounts vulnerable to the objection of *under-inclusiveness*, since they would then exclude too many acts that most people consider harmful but the qualified non-comparative accounts would not. Consider the following cases:

> *Clumsy physician*. A physician operates on a patient's slightly short-sighted eyes and causes the patient to have dim vision.

> *Robbery 1.* Adam breaks into the garage of Wayne's mansion and steals Wayne's new convertible while Wayne is at his penthouse in the city. (Meyer 2003, 153)

---

[11] For a further argument against Harman's view, see Gardner (2015).

If dim vision is not counted as a bad state of affairs for the patient, then the clumsy physician's operation does not harm the patient. Similarly, Adam's theft is not likely to cause Wayne to fall seriously below any plausible level of a normal healthy state, and thus the qualified non-comparative account alone does not provide us with a reason to object to Adam's act as a harm to Wayne.

A straightforward way to avoid these complications and counterintuitive results would be to supplement the non-comparative notion of HARMING III with the counterfactual comparative notion of HARMING I. According to such a "disjunctive" notion of harming:[12]

> HARMING IV: An action (or inaction), A, harms an individual, S, if and only if either
>
> - had A not occurred, S would have been better off in some respect (*comparative*); or
> - A causes a state of affairs that is a harm for S (*non-comparative*).

The advantage of the disjunctive notion of harming is that it allows the defender of the harm-based account to directly appeal to the comparative notion in the relevant cases above. By appealing to the comparative notion, she can explain why the physician harms the patient in *Clumsy Physician* and Adam harms Wayne in *Robbery 1*. In order to explain why the overall harm-based reasons are not against the operations in *Surgery* and in *Physician,* she can further explicate the definition of non-comparative harm in the following way:

> HARM IV (*non-comparative*): A state of affairs, T, is a harm for an individual, S, if and only if
>
> - T is a *bad* state of affairs for S[13]; and

---

[12] For the disjunctive notion, see also Meyer (2003, 2008).

[13] I leave it open here how exactly the bad state of affairs should be defined. Generally, I contend that a relevant bad state of affairs for an individual S is in some respect worse than the state of affairs in which S ought to be. This understanding of the bad state of affairs allows us to include the bad states mentioned by Harman (pain, early death, bodily damage, deformation, mental and physical discomfort, disease, or disability) as

- T is *not only an improvement* on the *bad* state of affairs in which S were before A occurred.
- The *more* T is an *improvement* on the *bad* state of affairs in which S were, had A not occurred, the *less* HARM IV is a reason against A.

This definition of the non-comparative harm allows the defender of HARMING IV to judge, first, that *Surgery* causes harm because it causes a bad state of affairs for me: the bodily damage resulting from the incision in my abdomen and the post-operative pain as I recover. But then she can conclude that these non-comparative harms provide no overall harm-based reasons against the surgery because I am not made worse off (but in fact am made better off overall) in a comparative sense and the resulting bad states (the hole and the post-recovery pain) are in fact a great improvement on the bad state of affairs I would have experienced had I not been operated on (worse pain and death). The non-comparative notion thus provides only a very weak harm-based argument against the surgery. Moreover, in *Physician*, HARM IV would allow the defender of HARMING IV to plausibly hold that the operation causes *no harm* at all to the patient, because she is not made worse off and the resulting bad state (dim vision) is in fact an improvement on her earlier bad state (blindness).

These explications seem to be plausible even beyond these cases. It seems odd to count a bad state, even a serious one, a harm if it only improves the earlier bad state of the person. But certainly, a bad state that is *not* only an improvement on an earlier, even worse state can count as a harm. Consider that we would assume that I am in extreme pains before the *Surgery* and thus the hole and the post-operational pain would be an improvement on my earlier bad state. Even in that case, the defender of HARM IV could claim that at least the hole in abdomen constitutes a separate bad state that is not *only* an improvement on my extreme pains before the surgery. Section 5 provides further discussion on this matter.

---

relevant and serious bad states. Also, dim vision in *Clumsy Physician* could be understood as a bad state for the patient, although much less serious than blindness.

Finally, the disjunctive HARMING IV and HARM IV offer a harm-based account that we could rely on in non-identity cases, in which the comparative notion does not apply at all. In *Two Women* and in *Parents,* the non-comparative HARM IV would judge Paul's and Peter's serious disability as bad states of affairs for both of them and not an improvement on any bad state of affairs they would have experienced had their parents not conceived them.[14]

4. How should we understand the disjunctive harming?

The disjunctive notion raises difficult questions of interpretation of its own, however. One particular area of questions concerns the relation of the two notions of harming – *non-comparative* and *comparative* – in the disjunction. One may ask, for instance, whether either of the notions has priority over the other, that is, does either one always apply primarily? Does the priority of either one also mean that the primary notion always provides a stronger objection against an action than the other? Or, if both notions of the disjunction can apply simultaneously to some act, does this mean that this provides stronger reasons to object to this act?

A possible way to answer the first question would be to prioritise the non-comparative notion: the non-comparative notion should be appealed to whenever it applies, and the comparative notion only when conditions for the non-comparative notion are not satisfied (e.g., in situations like *Robbery 1*). This would be in line with the central idea of the non-comparative accounts: to harm someone is to cause a morally bad state for that person, and this provides always a strong reason against the act (e.g. Harman 2009).

A concern with this solution is that by prioritising the non-comparative notion, even in cases when conditions for the comparative notion are satisfied, we would lose much of our normative capacity to explain the *variations* in the strength of the reasons against harming (see Gardner 2017). Consider the following modification of the *Clumsy physician:*

---

[14] I contend here that non-existence is not a *bad* state of affairs for anyone. Death, becoming non-existent, can be a bad state for the person who has been existent up to the point of her death.

> *Clumsy physician 2.* A physician operates on the eyes of two patients. The first one is slightly short-sighted and the second one already had badly dim vision. After the operation both patients are blind.

Since the operation causes a bad state of affairs (blindness) to both patients, they are both equally seriously harmed by the non-comparative notion. But the non-comparative notion would not recognise that the operation makes the first patient much worse off than the second. Yet such considerations about the degree of harm, understood as a difference in the harmed person's wellbeing, play a central role in our moral and legal understanding of harm, restitution and compensation.

Thus, reflecting the central place that the comparative notion has in our moral and legal thinking about harming (Parfit 1984, Woollard 2012), one option could be that, instead of giving priority to the non-comparative notion of harming, we should prioritise the comparative one. According to this understanding of the disjunctive HARMING IV, we would appeal to the comparative notion whenever it applies, that is, in cases such as *Clumsy Physician* and *Robbery 1,* and limit the appeal to the non-comparative notion only to non-identity and other exceptional cases, such as *preemption* (discussed below), in which victims are not made worse off relative to their state had the act not occurred at all, but our intuitive judgements hold that the victims have been harmed.

The significance of the comparative notion has been emphasised by Fiona Woollard (2012), who argues that though we can have reasons based on both comparative and non-comparative notions of harming, the reasons provided by the comparative one are stronger. Woollard bases her argument on our intuitions in certain *preemption* cases where an act that harms someone non-comparatively but not in the comparative sense is necessary to benefit a third person. Consider the following example she gives (Woollard 2012, 685; see also Parfit 1984, 71):

> *Saving Sarah.* Barney is about to shoot and kill Wayne. Adam has no way of preventing this. Sarah is about to die. Adam can save her but doing so would have the side-effect that he kills Wayne. Adam saves Sarah's life and kills Wayne.

According to Parfit, who considers a similar case, the fact that Adam's killing of Wayne does not make Wayne worse-off and greatly benefits Sarah, justifies Adam's behaviour as what he morally ought to do (Parfit 1984, 71). Woollard accepts this, though she admits that our intuitions in this case might change if the benefits to Sarah would be much smaller. If, for instance, by killing Wayne Adam could only save Sarah "from a scraped knee or a painful bruise", Woollard suggests that many of us would not permit the killing of Wayne even if he is not made worse off by that act. However, she argues that we are inclined to make a difference between the comparative and non-comparative harming in these cases. We tend to permit harming someone in a non-comparative way as a side-effect of benefitting some other person, whereas harming someone in a comparative sense would not be similarly permitted. This suggests, according to Woollard, that our reasons based on non-comparative notion of harming are weaker than the reasons provided by the comparative notion.

If Woollard is right, then the comparative notion would not only be prioritised whenever it applies but also that the reasons based on the comparative notion would be stronger. This understanding of the disjunctive HARMING IV would have the advantage of explaining the variations in the strength of the reasons we have in cases like *Clumsy physician 2* and *Saving Sarah*. Still, granting priority to the comparative notion in this way seems to have some counterintuitive consequences, which those who doubt the centrality of the comparative notion have pointed out. Consider, for instance, James Woodward's (1986) example of *Viktor* Frankl. In Woodward's example, Viktor Frankl's imprisonment in a Nazi concentration camp leads to a major enrichment of Frankl's later life, thus benefitting him overall. Therefore, according to the comparative notion of harming, Frankl is not harmed, which, being counterintuitive, indicates that the need to appeal to a non-comparative notion is larger and not limited to exceptional cases such as those named above (i.e. non-identity and pre-emption).

But Woodward also makes a further note about the cases like this: even if the necessary condition for the comparative notion were satisfied in this example, the harm-causing features that the comparative notion points to might not be the

right one. Let us, for instance, accept what Boonin (2014, 79) calls the "Short-term" version of the comparative notion.[15] According to this version, an act harms when it makes someone *at least at some point in time* worse off than she would have been at that point had the act not occurred (even if the act makes her better off overall than she would have been had the act not occurred). Thus, the Short-term version can rightly account for *Viktor's* treatment in the camp as a harm. Still, the defender of the harm-based account can hold that the Short-term version of the comparative notion would not pick out the right features and the seriousness of the harm. It is not only that Frankl is made worse off than he would have been had he not been in the camp, it is the awful bad state of affairs caused by the imprisonment and the horrible treatment in the concentration camp (e.g. Harman 2004). To pick out the seriousness of the kind of harm inflicted on Frankl as an awful bad state, we would also need the non-comparative notion of harming to apply.

This suggests, *pace* Woollard's argument, that the strength of the reasons against harming is not straightforwardly related to whether an action causes comparative harm. Gardner (2017) also offers a further reason against Woollard's argument. According to Gardner, the reasons against killing Wayne in *Saving Sarah* are not weaker because the action harms Wayne "only" in a non-comparative way. Instead, the reasons are weaker due to the action's additional feature of causing *redundant harm*, that is, harm that the victim would have been suffered anyway, had the action not been performed.[16] For Gardner, the reasons against redundant harm-

---

[15] Boonin argues that because the "Short-term" version of the comparative HARMING I is able to explain harming in examples like *Viktor,* such cases do not work against the comparative notion and thus fail to provide a satisfactory ground for the non-comparative accounts to solve the NIP by rejecting the comparative notion.

[16] To support her argument, Gardner (2017) offers a pair of cases, in the first of which the action qualifies *both* as non-comparative harming and redundant harming and in the second of which the action harms only non-comparatively:

> "*Inducing Paralysis*: It is Leland's 5th birthday, but unfortunately, a piece of debris, A, is flying through the air towards Leland's midsection. If Bernard does nothing, A will hit Leland and para-

ing are generally weaker than the reasons against non-redundantly harming and thus *Saving Sarah* offers no general argument for the claim that the non-comparative notion provides weaker reasons against harming than the comparative notion.

Often the fact that an action makes someone not only worse off but also pushes them into a serious enough bad state seems to add to the reasons that we have against the action. Consider the following:

> *Robbery 2*. Adam can either steal from wealthy Wayne or from poor Barney. Both acts would make either Wayne or Barney worse off by the same amount but stealing from Barney would cause him to be unable to afford enough food. Wealthy Wayne can easily afford food even after the theft.

All other things being equal, it seems to be in accordance with our intuitions that Adam has more reasons against stealing from Barney than from Wayne. Yet giving priority to the comparative notion and allowing the non-comparative notion to apply only when the comparative one does not and could not pick up this kind of difference in the strength of the objections against these two thefts.

> lyze him from the waist down. Bernard's only other option is to push Leland into the path of another piece of debris, B, which would cause Leland a qualitatively identical injury; however, pushing Leland into the path of B would prevent Sarah from having a broken arm. Bernard pushes Leland into the path of B.
>
> *Selecting for Paralysis*. Enid is choosing which embryo to implant into her uterus. Embryo B will become an individual named Sheldon. B has a mutation that will cause Sheldon to become instantly and permanently paralyzed from the waist down on his 5th birthday. However, implanting embryo B will prevent Sarah from having a broken arm. Enid implants B."

For Gardner the pair cases show that causing redundant harm makes Bernard's non-comparative harming morally less bad than Enid's non-comparative harming. Thus the case shows that redundant harming is a separate and additional feature of an action that also explains our intuitions in *Saving Sarah*.

Such cases seem to support the understanding of the disjunctive notion in a way that allows both notions to apply whenever the conditions for them are satisfied. Furthermore, in order to explain the difference in the strengths of objections we hold between stealing from either Barney or Wayne, we could allow that the conditions for both comparative and non-comparative notions of harming being satisfied would add to the strength of the objection we have against that act. According to this, what could be called the *Additional reasons* interpretation of the disjunctive notion, our intuition that stealing from Barney is more objectionable than from Wayne would be explained by the fact that only the comparative notion applies and provides us with reason against stealing in the case of stealing from Wayne, while in the case of stealing from Barney the conditions for both notions of harming are satisfied and they both give us reasons (that are additional to each other) to object to stealing from Barney.

In *Saving Sarah,* the *Additional reasons* understanding of the disjunctive HARMING IV is able to recognise the difference in the strength of reasons against killing Wayne without committing to Woollard's view that the comparative notion of harming always gives stronger reasons than the non-comparative one. According to the *Additional reasons* understanding, in *Saving Sarah* only the non-comparative notion applies, which makes the objection against Adam's action weaker than it would have been if Wayne had not been killed anyway. Moreover, that the killing causes redundant harm could be added as a weakening feature to the *Additional reasons* view: the weakness of the reason against killing in *Saving Sarah* in relation to other non-comparative serious harming would be explained by the redundant harming. The weaker objection is reflected in our intuition to accept Wayne's preemptive killing as a side-effect of saving Sarah. Yet, our inclination *not* to accept such killing as a side-effect of a less significant benefit reflects the reasons that the non-comparative notion gives us. Causing a seriously bad state of affairs for Wayne (even though that makes him not worse off) can only be justified by strong reasons, for instance, to avoid some other person, Sarah, from suffering an equally serious state. Were the justifying reasons lesser, e.g., that the act would "only" prevent the other person from being worse off but not

from a seriously bad state of affairs, they would not justify the non-comparative harm to Wayne.

The *Additional reasons* reading of the disjunctive notion seems to have some plausibility in the cases above. It allows us to retain the comparative notion that we often apply when we justify the reasons (and the strength of these reasons) we have against acts that make others worse off than they would have been without those acts. But as the cases of *Viktor* Frankl and *Robbery 2* show, we often also need to appeal to the non-comparative notion to fully account for the right kind of seriousness of the harm-based reasons, and the *Additional reasons* understanding allows us to do this. But is it able to provide a plausible account of harming that would offer a solution to the NIP? In the following section, I consider the *Additional reasons* view against the objections that David Boonin (2014) has raised against the harm-based views.

5. *Additional reasons* understanding and the NIP

Boonin (2014) offers a rigorous argument against the harm-based solutions to the NIP. His central claim is that harm-based solutions do not fulfil the *independence* and *robustness* requirements, since in order to do so they would need to provide an independent and plausibly robust reason to reject the comparative HARMING I – and any weakened version of it, such as the "Short-term" version discussed above – in each of the cases above, and thus block the move to the NIP's implausible conclusion that results from the comparative account. In Boonin's view, they are unable to do this. In cases such as *Surgery* and *Physician,* the comparative HARMING I provides the most intuitively plausible explanation for why these cases involve no harming in the first place: the patient is not made worse off than she would have been had the operation not taken place. In cases like *Viktor,* where the original HARMING I seems to have problems in explaining the intuitively obvious harm inflicted on the victim, there is a version of the comparative account (like the "Short-term" version) that can both explain why the act is harmful and still generate the NIP.

Since the disjunctive HARMING IV and its *Additional reasons* understanding allows us to appeal to the comparative

notion in the cases above, it does not face the same challenge of full rejection of the comparative HARMING I and any of its weakened versions. Instead, the *Additional reasons* view needs to show that, even if the comparative notion applies, the relevant bad state in those cases provides an *additional* and *independent* harm-based reason to which we can then appeal in the non-identity cases. I have argued above that even if Boonin were right, and the "Short-term" version of the comparative HARMING I would apply in cases like *Viktor,* there would still be an additional reason based on the non-comparative harm as a bad state for Frankl. It is only by appealing to those reasons that we can account for the overall reasons we have against Frankl's treatment in the Nazi's concentration camp. In *Robbery 2* I also contend that we find it intuitively plausible that Barney's bad state adds an independent reason that makes a difference between those two thefts. But let us test these intuitions against another example:

> *3 Surgeries.* A surgeon operates on 3 patients. In each operation, the surgeon cuts a hole in the patient's abdomen in order to remove their swollen appendix. Due to painkillers, the appendix caused only moderate pain to the patients before the surgery, but if the operation had not been performed, all the patients would have suffered extreme pain and died very soon after.
>
> In *Operation 1* cutting open the patient's abdomen causes her bodily damage resulting from the incision but *no* post-operative pain (as the patient recovers).
>
> In *Operation 2* the patient suffers the bodily damage resulting from the incision in her abdomen and moderate post-operative headache as the patient recovers.
>
> In *Operation 3* the patient suffers the bodily damage resulting from the incision in her abdomen and a *deformation* in the patient's nerve system that causes a reoccurring painful post-operative headache for the rest of her life.

Each of these operations cause the patients to be in a harmful bad state of experiencing the bodily damage resulting from the incision in their abdomen. But, according to Boonin (2014), the harm caused by the hole in patients' abdomen is unable to establish a solution to the NIP for two reasons.

First, the short-term harm resulting from the hole can be accounted for by the "Short-term" version of the comparative HARMING I that also leads to the NIP. Second, the harm resulting from the hole is structurally very different from the non-identity cases in which the act causes the victim to be in the bad state for the long term and not only in the short term. Only the more long-term post-operative pain makes the surgeries analogous to the non-identity cases.

So, to offer a solution to the NIP, the *Additional reasons* view needs to show that there is an additional and independent reason based on the post-operational pain of the patients even if they are not made worse off. *3 Surgeries* aims to do that. It is clear that there is a difference in the strength of the harm-based reasons between the operations even if they all save the patient's life and thus make them overall better off. *Operation 1* is the least harmful since there is no post-operative pain at all. *Operation 2* and *Operation 3* involve post-operative pain but the pain caused by *Operation 3* is much more severe (reoccurring painful headache caused by a permanent deformation) and more long term (rest of the patient's life) than the post-operational pain in *Operation 2* (moderate headache during patient's recovery). Still, neither post-operational pain makes the patient worse off by any version of the comparative HARMING I and thus the comparative account alone is unable to account for the difference in the harm-based reasons we have in *3 Surgeries*. In each of the *Operations,* the comparative account recognises only the harm that results from the hole in the patient's abdomen, while the *Additional reasons* view and the HARM IV definition of non-comparative harm is able to account for the differences in the post-operational bad states resulting from the operations. According to HARM IV, *Operations 2* and *3* cause bad post-operative states that do not improve on the patient's earlier bad state before the operation (moderate pain).[17] The result-

[17] If we assume that the patients are in extreme pain before the surgery, then the *Additional reasons* view could hold that while the post-operational pain even in *Operation 3* may be an improvement on the bad state the patient was in before the surgery, the deformation of the nerve system would also be an independent bad state that is a harm, and not only an improvement on the earlier bad state of the patient (like dim vision is for a totally blind person in *Physician*).

ing post-operational bad states in *Operations 2* and *3* are therefore harms even though their strength as reasons is weakened by that fact that they are *improvements* on the *bad* state of affairs in which the patients would be, had they not been operated on (worse pain and death). But since the bad post-operational state resulting from *Operation 3* is more serious and less of an improvement than the resulting bad state in *Operation 2*, the reasons against the surgery are still stronger in *Operation 3* even if they may plausibly be overridden by the fact that the surgery makes the patient better off overall in the comparative sense, and the resulting bad state, while serious, is still a great improvement on the bad state of affairs the patient would have suffered had she not been operated on (extreme pain and death).

Therefore, the *Additional reasons* view has the strength to be able to account for the harm-based reasons we have in cases like *3 Surgeries*. Such examples also add the plausibility that the non-comparative notion provides an independent reason against harming: a reason that we can appeal to in non-identity cases. Thus, the *Additional reasons* view seems to offer a harm-based account that would be able to block the move to the NIP.

However, the view appears to go against Parfit's No-Difference View. In the following section I will evaluate how implausible this commitment is and whether it makes the *Additional reasons* view vulnerable to the objection that it would not fulfil Boonin's modesty requirement. According to the modesty requirement, the solution to the NIP should not yield conclusions that are even more implausible than the conclusion of the NIP.

6. Harm-based solutions and the No-difference view

The No-Difference View holds that an action being identity-affecting should not affect the strength of the reasons we have against the act. In other words, the reasons to object to an action that makes a person badly off but not worse off in an overall comparative sense should be as strong as the reasons

to object to an action that makes a person badly off in the same way but also worse off overall.[18]

Thus, according to the No-difference view, there should be equally strong objections against Carla's and Paula's decision in *Two women*. They both harm their child and the fact that Carl is made worse off, while Paul is not, should not make a difference. However, the *Additional Reasons* view would make a difference. According to the *Additional Reasons* understanding, we would have both comparative and non-comparative based reasons against Carla's decision while only non-comparative based reasons against Paula's decision.

In this section I consider how serious objection this incompatibility with the No-Difference View is for the *Additional reasons* view. Consider first the following case, which is a slightly modified version of *Two women*:

> *Two Women 2.* Carla has a normal one-year old boy, Carl. She learns that, unless she provides Carl with some treatment, there is a risk that Carl may develop a certain severe and irreversible disability. She decides not to take this treatment. As a result, Carl is handicapped with the disability.
>
> While Paula is trying to become pregnant, she learns that, if she conceives a child now, there is a risk that they may develop the same severe disability. If she waits two months before conceiving the child, there would be no such risk. She decides not to wait. As a result, her child, Paul, is born with the disability.[19]

Again, as in the original example of *Two Women,* Carla is harming her child both in a comparative as well as in a non-comparative way. However, asking people to compare her decision to that of Paula now might easily obtain answers that are not in line with the No-Difference View. Carla's decision, which is harmful to her healthy child, seems much worse than Paula's. This asymmetry that many people would

---

[18] There is also another stronger reading of the No-difference view, which requires that we should have the *same* reasons for objecting to both actions (Parfit 1986; Woodward 1987). The non-comparative notions of harming (HARMING III) would be compatible with this reading, while disjunctive notions that allow that we may have (different) reasons based on comparative or non-comparative notions would be incompatible.

[19] The modification follows Woodward (1987, 812).

now allow between their judgements about the decisions of these two mothers suggests that our intuitive reactions in these cases are more ambiguous and controversial than those who hold them as providing indubitable support for the No-Difference View would like to think them to be.

Justin Weinberg (2013) offers further reasons in support for the asymmetry in our judgements between same individual and non-identity cases. When introducing the No-difference view in *Reasons and Persons*, Parfit presents the example of *The Medical Programmes,* which has the same structure as *Two Women,* only at the collective level. The first programme will offer a million pregnant mothers, "Carlas", a treatment that will prevent their children from developing a severe disability once they are born. The second programme will prevent the same number of children from being born with that same disability by warning a million mothers, "Paulas", to postpone conception for at least two months. Parfit then asks, which programme we should cancel if we have money only for one programme.

When facing *The Medical Programmes* so described, we might well be tempted to judge the outcomes of both programmes as "morally equivalent", as Parfit wishes (Parfit 1984, 369). Still, this judgement can easily be a result of representing the situation as a choice between two outcomes rather than acts or choices that have an identifiable agent and a victim. Given this description of the examples, it invites us to make the judgement on the level of the alternative outcomes, which include large numbers of people. However, as Weinberg points out, it is an oft-mentioned fact about our moral psychology that we tend to pay less attention to "bad effects brought to our attention when such effects happen to very large numbers of people and pay more attention when such effects happen to small numbers of people we can identify" (Weinberg 2013, 29). If this is true, then our intuitions related to such a choice would be more reliable in the *Two Women* or *Two women 2* cases that are structurally similar to *The Medical Programmes* but represent the choice in a way that attracts our moral attention and consideration more strongly.

Therefore, evaluating the outcomes of *The Medical Programmes* as morally equivalent need not imply that we would also judge that both of the acts we are comparing are equally

objectionable, or that those who make the choices in these cases should have equally strong reasons for their choice. In other words, even if our intuitions, when faced with examples like *The Medical Programmes,* were in line with Parfit's No-Difference View, this would not automatically mean that we would also judge that Carla and Paula have equally strong reasons against their choices. Our intuitions might only reflect how we evaluate the situation when comparing the moral value of the outcomes including large numbers of people.

If these observations are correct, they would weaken the claim that being against the No-Difference View makes harm-based accounts seriously implausible.[20] We should also note that allowing a difference in the strength against same individual and non-identity harming would not lead to dramatic changes in our practical judgements about what we ought to do in many non-identity cases. In the *Additional Reasons* view, it is plausible to hold that non-comparative HARM IV always provides us with strong reason against the act that causes the defined bad state of affairs for the victims (and is not an improvement in the state they would have been in had the act not occurred). While the fact that an act would also make someone worse off as required by the comparative sense would add to the reasons against the harmful act, the addition in the overall strength of the reasons might be much less than what the non-comparative reason already provides.

Thus, for instance, in addressing the problem of anthropogenic climate change, the reasons based on any plausible harm-based account would require very much stronger actions from us than we are performing at the moment (see e.g., Kyllönen 2018; Cripps 2013). The fact that the comparative notion does not apply does not make these requirements significantly weaker, according to the *Additional reasons* view. Therefore, I contend that the incompatibility with the No-Difference View does not make the *Additional reasons* view

---

[20] Gardner (2017) offers further examples against the No-Difference View and argues that sometimes it actually "may be a *good thing* for an account of harming to allow for the possibility that the reason against non-comparative harming is weaker than the reason against comparative harming." (80; emphasis in the original)

vulnerable to the objection that it would not fulfil Boonin's modesty requirement. That our reasons against harming in non-identity cases are to some degree weaker is, after all, a much less implausible result than accepting the NIP, which denies that we have *any harm-reasons* against polluting the climate that are related to the future people who will suffer the most severe consequences of the pollution.

7. Concluding remarks

In this paper I have investigated the arguments in favour of a harm-based solution to the Non-Identity Problem. I have focused my attention on a particular disjunctive understanding of harming, the so-called *Additional reasons* view, that allows us to appeal to both comparative and non-comparative notions of harming. I have argued that it offers a plausible explanation of the intuitions we have in several cases of harming and thus it has independent plausibility even outside the NIP. Since the *Additional reasons* view aims only to add non-comparative harm-based reasons to the comparative ones, I have also suggested that the view finds it easier to satisfy Boonin's robustness requirement than harm-based accounts that aim at full rejection of the comparative account. Finally, I have argued that the fact that the *Additional reasons* view is incompatible with Parfit's No-Difference View does not cause the view to have seriously implausible conclusions. On the contrary: a seriously bad state that is not an improvement for the victim always provides a strong reason against the act that causes the bad state and the fact that it would also make the victim worse off in a comparative sense only adds to this strong reason.

University of Helsinki

References

Boonin, D. (2014), *The Non-Identity Problem and the Ethics of Future People*, Oxford, Oxford University Press.

Cripps, E. (2013), *Climate Change and the Moral Agent. Individual Duties in an Interdependent World*, Oxford, Oxford University Press.

Hanser, M. (1990), "Harming future people", *Philosophy & Public Affairs* 19, pp. 47–70.

Harman, E. (2004), "Can we harm and benefit in creating", *Philosophical Perspectives* 18, pp. 89–113.

Harman, E. (2009), "Harming as causing harm", in Roberts, M. and Wasserman, D. (eds.), *Harming Future Persons: Ethics, Genetics and the Nonidentity Problem,* Springer.

Heyd, D. (1992), *Genethics. Moral Issues in the Creation of People*, Berkeley, University of California Press.

Gardner, M. (2015), "A Harm-Based Solution to the Non-Identity Problem", *Ergo* 2, pp. 427–44.

Gardner, M. (2017), "On the Strength of the Reason Against Harming", *Journal of Moral Philosophy* 14, pp. 73–87.

Gardner, M. (2019), "David Boonin on the Non-Identity Argument: Rejecting the Second Premise", *LEAP* 7, pp. 29–47.

Kyllönen, S. (2018), "Climate change, no-harm principle, and moral responsibility of individual emitters", *Journal of Applied Philosophy* 35(4), pp. 737–758. doi: 10.1111/japp.12253

Meyer, L. (2003), "Past and Future. The Case for Threshold Notion of Harm" in Meyer, L., Paulson, S.L. and Pogge, T. (eds.) *Rights, Culture and the Law. Themes from the Legal and Political Philosophy of Joseph Raz,* Oxford, Oxford University Press.

Meyer, L. (2008), "Intergenerational Justice" in *Stanford Encyclopedia of Philosophy,* available at <http://plato.stanford.edu/entries/justice-intergenerational/>

Meyer, L., Roser, D. (2009), "Enough for the Future" in Gosseries, A., Meyer, L. (eds.), *Intergenerational Justice*, Oxford, Oxford University Press.

Parfit, D. (1983), "Energy Policy and the Further Future: The Identity Problem" in MacLean D. and Brown P. (eds.) *Energy and the Future,* Totowa, N.J., Rowman & Allanheld.

Parfit, D. (1984), *Reasons and Persons*, Oxford, Clarendon Press.

Parfit, D. (1986), "Comments", *Ethics* Vol. 96, No. 4, pp. 832-872.

Parfit, D. (2017), "Future People, the Non-Identity Problem, and Person-Affecting Principles", *Philosophy & Public Affairs* 45(2), pp. 118–157.

Rivera-Lopez, E. (2009), "Individual procreative responsibility and the non-identity problem", *Pacific Philosophical Quarterly* 90, pp. 336–363.

Scanlon, T.M. (2007), "Wrongness and Reasons: A Re-Examination", *Oxford Studies in Metaethics* Vol. 2, pp. 5–20.

Shiffrin, S. (1999), "Wrongful life, procreative responsibility, and the significance of harm", *Legal Theory* 5(2), pp 117 – 148.

Schwartz, T., (1978), "Obligations to Posterity", in Sikora, R. and Barry, B. (eds.) *Obligations to Future Generations*, Philadelphia, Temple University Press.

Weinberg, J. (2013), "Non-Identity Matters, Sometimes", *Utilitas* 26(1), pp. 1–11.

Woodward, J. (1986), "The Non-Identity Problem", *Ethics* Vol. 96, No. 4, pp. 804-831.

Woodward, J. (1987), "Reply to Parfit" *Ethics*, Vol. 97, No. 4, pp. 800-816.

Woollard, F. (2012), "Have We Solved the Non-Identity Problem", *Ethical Theory and Moral Practice* 15, pp. 677–690.

No Safe Haven for Truth Pluralists

TEEMU TAURIANEN

1. Introduction

Truth pluralism has become a much-discussed position in contemporary truth-theoretic debates (Pedersen & Wright 2013; Wyatt 2013; Wyatt & Lynch 2016; Wyatt, Pedersen & Kellen 2018; Edwards 2018a, 2018b).[1] The general thesis of truth pluralism is that there are many ways for truthbearers to be true.[2] According to the standard explanation, sentences get to be true in different ways based on their domain membership. For example, sentences addressing ethical matters, or composed of ethical concepts, belong to the domain of ethics, which is governed by an adequate truth-grounding property such as *coherence*. Other sentences are about extensional states of affairs, thus belonging to the domain of physics, which is governed by an appropriate truth-grounding property such as *correspondence*. By accommodating both, coherence and correspondence criteria, truth pluralists aim to offer a definition of truth that scales over the full range of natural truth-apt discourses, thus offering a viable alternative to traditional monist and deflationary theories (Pedersen, Wyatt & Kellen 2018, 4).[3]

---

[1] The term "truth pluralism" was introduced by Crispin Wright (1992) in *Truth and Objectivity*. One of its original goals was to arrive at a definition of truth that would allow both realist and anti-realist intuitions to be satisfied.

[2] For practical reasons, I commit to treating sentence tokens as truthbearers.

[3] Note that one can be a pluralist in the context of a single truth property such as correspondence (Sher 2005). Further, one can form a definition that commits to multiple deflated truth properties (Beall 2013). Finally,

Discourse domains have a crucial explanatory role in current pluralist frameworks.[4] As noted, pluralists of all persuasions tie truth-grounding properties, such as coherence and correspondence, to domains rather than to individual sentences. Consequently, the truth of different *types* of sentences is accounted for by their domain membership. In an optimal scenario, each truth-apt sentence belongs to a single unambiguously individuated domain governed by exactly one truth-grounding property. From this follows that, by knowing the domain membership of a sentence, one is able to account for its truth by inferring the property that grounds truth for the relevant domain. Without domains, explaining why a particular sentence is true in one way rather than another becomes difficult if not impossible (Wyatt 2013, 231-232). Even worse, without domains, some sentences end up being both true and false in pluralist frameworks, thus conflicting with the standard law of non-contradiction (Edwards 2018b, 85–86). As a result of such issues, domains are held as a safe haven that supposedly guard pluralists of all sorts from various issues with definitional ambiguity and indeterminacy.

In this paper, I argue that, like domain-free models, current domain-reliant pluralist frameworks generate similar issues with ambiguity and indeterminacy. This follows from the current pluralist neglect of addressing the issues that inherent natural language ambiguity generates in their frameworks. As I demonstrate later, because some truth-relevant components of sentences allow for different yet equally valid readings, these components end up assigning sentences to multiple domains with different truth-grounding properties, with the consequence of having one of these properties and lacking another. As a result, domain-reliant pluralist frame-

---

one can form a hybrid definition that allows for both inflated and deflated truth properties. In general, pluralists can utilize different monist theories, various inflated and deflated truth properties, and the logico-expressive definitions of the truth predicate, which are crucial components of deflationary theories.

[4] As Wyatt (2013, 228) notes, discourse is a more permissive category than a discussion. One can have a discussion about both equality of income and preservation of natural resources and still be under the same domain of ethical discourse.

works end up conflicting with both the standard laws of non-contradiction and identity. Against this backdrop, I argue that pluralists should re-consider their current aim of offering a complete, unambiguous, and determinate definition of truth for natural discourse. Finally, based on the findings, I explore some solutions to the issues noted and discuss the prospects of pluralist theories.

2. Truth Pluralism

Various forms of the general truth pluralist thesis have been endorsed in the literature (Edwards 2018a, 129; Kim & Pedersen 2018, 124). In general, these forms divide into strong (SP) and moderate (MP) categories:

> *SP*: there are many ways of being true, *none* of which is had by all true sentences
>
> *MP*: there are many ways of being true, *some* of which are had by all true sentences

The central difference between strong and moderate forms is that the former commit to radical *disunity* regarding truth, while the latter include both unifying *and* disunifying features. According to strong pluralism, truth is many but *not* one. There are independent ways of being true (T1, …, Tn), with no connection in between. According to moderate pluralism, truth is both one *and* many. Different sentences get to be true in different ways, but they are all true in some unifying sense. According to the truth pluralist literature, strong forms are not widely supported (Kim & Pedersen 2018, 108; Pedersen & Lynch 2018, 561) because moderate forms have ready answers to some of the objections faced by the strong forms. For example, strong pluralism has difficulty accounting for the *normativity* of truth, defining *validity*, and explaining *generalizations* via the truth predicate. Think about the normative aspect of truth as that which is *prima facie* correct to believe.[5] This is a unifying feature of *all* truths. Further, validity or logical consequence is standardly defined as the

[5] A further note concerns the value of truth. If strong pluralists hold that truth is valuable, they ought to explain whether different ways of being true entail variance in the value of truth.

preservation of truth over inference. The problem is that inference can be mixed, meaning that the premises can be true in different ways, assuming the basic pluralist premise that there are many ways of being true. The question, therefore, is what type of truth (T1, …, Tn) is preserved over mixed inference? Lastly, concerning generalizations via the truth predicate, statements such as "everything that the Pope said is (or was) true" present themselves as ambiguous in strongly pluralist frameworks. In which of the possible ways (T1, …, Tn) is everything that the Pope said true? Because of such issues, I restrict the discussion in this paper to moderately pluralist theories, though much of what will be said here also concerns strongly pluralist frameworks, especially insofar as they commit to using discourse domains as an explanatory resource.

As noted, the general thesis of moderate pluralism is that truth is both one and many. According to the standard explanation, truth displays unity on global, general, or language levels and disunity on local, domain, or sentence levels. According to the standard explanation, there is a general or elite way of being true. This is achieved through the possession of a *general truth property* F, which is denoted by the predicate "is true."[6] However, abiding by the general pluralist thesis of truth variability, discursively distinct types of sentences assume this property in different ways by possessing the relevant truth-grounding property of their domain. In other words, all true sentences are true in a general or unifying way, but the grounds of truth are many; depending on the domain, sentences possess the general truth property in different ways. This explanatory framework rests on two central commitments: a *platitude-based strategy* for defining the general truth property F and *domain reliance*, which accounts for the variability of the grounds of truth.

Starting with the first commitment, the general truth property F is commonly defined through a platitude-based strategy. According to this strategy, the general truth property inherits its nature from the concept of truth, which can be accessed through certain platitudes, intuitions, or folk beliefs

[6] Abiding by the law of symmetry, falsity is defined as the lack of said property.

about a notion. For example, Lynch (2009, 8–13, 2013, 24) commits to the following widely cited platitudes, translated in a way that makes reference to sentences:

> *Objectivity*: a sentence is true iff things are as the sentence says.
>
> *Norm of Belief*: it is *prima facie* correct to believe a sentence iff the sentence is true.
>
> *End of Inquiry*: other things being equal, true sentences are a worthy goal of inquiry.

A chosen set of platitudes are then used as a collective definition for the general truth property.[7] For example, Edwards notes that "[t]ruth is given as the property that is exhaustively described by the truth platitudes" (2018a, 126, 153). Simply put, moderate pluralists hold that the general truth property F is best characterized through specific platitudes about the concept of truth. How exactly one accounts for the metaphysical connection among the concept of truth, the platitudes about truth, the general truth property, and the truth-grounding properties will be largely overlooked in this article.[8] I will simply assume that some satisfactory explanation

---

[7] Note that the chosen set of platitudes need not be treated as an exhaustive definition of the concept of truth.

[8] When claiming that different types of sentences get to be true in different ways *because* they belong to distinct domains, the "because" relation between the concept of truth and the truth-grounding properties can be accounted for in many ways, some candidates being *grounding*, *manifestation*, *instantiation*, *entailment*, *determination*, and *conceptual necessity* (see Edwards 2018a, 122–141). For practical reasons, I commit to using grounding as the appropriate relation between the general truth-property F and truth-rendering properties. If one wants to remain neutral regarding a specific relation, then the term "truth-rendering" property is available. Thus, in my view, the truth of sentences belonging to different domains is *grounded* in a plurality of truth-grounding properties. However, most of what will be said here is independent of this question. Further, as the general truth property is a second-order property, the possession of which is determined by the ability of a sentence to possess the first-order truth-grounding property that is relevant to the domain it belongs to, truth-grounding properties can be called quasi-truth properties. As Pedersen notes, truth-grounding properties are "that in virtue of which propositions are true within specific domains, and so, locally behave very much

to this is available. The point of focus for the remainder of this paper is the second key commitment of pluralist frameworks to domain reliance, which plays a crucial explanatory role in accounting for the variability that truth displays across different regions of discourse.

3. Discourse Domains

According to domain reliance, truth-grounding properties such as coherence and correspondence vary by regions of discourse or discourse domains:

> Despite their different views on how to best articulate truth pluralism, strong and moderate pluralists share significant commitments. One such commitment is the commitment to *domains*. Domains are a crucial component of the theoretical framework of pluralism, as reflected by the fact that the core pluralist thesis is that the nature of truth varies *across domains*. (Pedersen, Wyatt & Kellen 2018, 6–8).

Further, Edwards (2018b, 85–86) makes an even stronger claim, arguing that domains ought to be treated as an inseparable feature of pluralist frameworks: "As a result, I think that [all] pluralists should take the notion of a domain seriously as a central aspect of the view" (see also Edwards 2011, 28, 41). Thus, there is no doubt that domains play a crucial explanatory role in current pluralist frameworks.

In general, discourse domains are taken as classes of sentences that are individuated by some semantic or ontological factor. As Kim and Pedersen (2018, 112) note, sentences belong to different domains because "they concern different subject matters or are about different kinds of states of affairs." According to a semantics-based strategy, sentences count as members of domains based on their *subject matter* or *aboutness*. For example, sentences that address ethical matters, or are composed of ethical concepts, belong to the domain of ethics and those addressing religious matters to the domain of religion. Ontology-based strategies distinguish between different types of *entities* referred to by the truth-relevant

---

like truth. They are quasi-truth properties because they only exhibit this behavior locally and, so, are distinct from truth" (2020, 356).

components of sentences (Edwards 2018a, 77–81, 2018b, 89-92). For example, sentences instantiating terms that designate extensional objects, or predicates that attribute representational properties, belong to a domain of realist speech, and those designating abstract objects or attributing non-representational properties belong to an anti-realist domain. In both cases, the goal is to individuate domains in a way that leads to them being unambiguous classes of sentences. Based on the desire to achieve this result, pluralists aim to account for the truth of different types of sentences based on their domain membership. As Edwards (2011, 31) writes: "According to the alethic pluralist, there will be a robust property in virtue of which the propositions expressed by sentences in a particular domain of discourse will be true, but this property will change depending on the domain we are considering." Similarly, Lynch (2009, 77) notes that: "Propositions about different subjects can be made true by distinct properties each of which plays the truth-role [for the relevant domain]."[9] Finally, based on this somewhat heavy metaphysical framework consisting of both the platitude-based strategy of defining the general truth property and the domain reliance that accounts for truth-variability, domain-reliant moderate pluralists argue that they can offer an unambiguous and determinate definition of truth, including for the grounds of truth, which scales from the concept of truth to the full range of truth-apt discourse in the context of natural languages.

However, according to the literature, domain reliance introduces its own array of definitional issues: "[t]he notion of a domain has been both a key and controversial aspect of pluralist theories" (Edwards 2018b, 103; Wyatt 2013). Some of these issues deal with the metaphysically challenging task of individuating domains. For others, ambiguity is generated by discourse bearing mixed content from various domains. Based on these challenges, some have expressed skeptical remarks about the very possibility of achieving a satisfactory

[9] It is worth emphasizing that domains rather than individual sentences play the adequate truth-bearing role in domain-reliant pluralist frameworks. What is relevant for sentence-level truth-grounding is their domain membership.

pluralist account (David 2013, 49).[10] In what follows, I explore certain issues with current domain-reliant pluralist models caused by inherent natural language ambiguity.

4. Issues with Defining Domains

Surprisingly, not much work has been done in exploring the nature of discourse domains in the standard pluralist frameworks: "Despite the central role that domains play within the standard pluralist framework not much systematic work has been done on their nature" (Kim & Pedersen 2018, 111). Direct studies can be found from Edwards (2018a, 77–82; 2018b, 86–100) and Wyatt (2013, 225–236). In general, the now prominent domain-reliant pluralists bear the burden of defining discourse domains in addition to offering a definition of truth that utilizes the notion. However, as noted in the literature, defining domains is a cumbersome and complex task (Lynch 2018, 66–67; see Blackburn 2013, 265; Quine 1960, 131). More specifically, domain-reliant pluralists are pressured to offer an answer to at least the following questions, some more truth-theoretically relevant than others: What are the necessary and sufficient characteristics of each domain, and how are they distinguished from one another unambiguously? How is the domain membership of sentences accounted for? How is the domain membership of sentences bearing content—potentially counting as members of multiple domains—accounted for? How are truth-grounding properties tied to the relevant domains?[11] Can a single domain have more than one truth-grounding property?[12] How can truth-

[10] Despite this, and perhaps surprisingly, the literature exploring alternative approaches such as domain-free models is sparse.

[11] Why is P1 and not P2 the truth-grounding property of D1? Further, it can be argued that the truth of some sentences, such as "water is H20," is based on multiple properties because it includes terms that refer to both mind-dependent and -independent entities. Thus, whether or not it is indeed true is dependent on both correspondence with actual states of affairs and coherence with the system of true beliefs that gives meaning to its terms.

[12] Wyatt (2013, 234) argues for an alternative approach where sentences belong to multiple domains: "truth pluralists should not presuppose that every atomic proposition belongs to one and only one domain." Lynch

apt sentences be separated from non-truth-apt sentences in the context of domains?[13] Can some sentences, such as necessary truths, be members of multiple domains, or does each domain include its own subset of necessary truths?[14] While the resolution of some of these issues is underway, no simple answers are forthcoming.[15]

Perhaps the most researched issue concerning domains is a set of problems labeled *mixed discourse* (Bar-On & Simmons 2018, 38). The general idea of mixed discourse is simple. Take two sentences, "snow is white" and "snow is beautiful," from the distinct domains of speech regarding extensional and aesthetical properties, individuated by the extensional predicate "is white" and the aesthetical predicate "is beautiful." Assuming that both sentences are true and that the truth of speech about extensional entities is grounded in correspondence, and that of aesthetics in coherence, one can form simple mixes of sentences, compounds, and inferences where both extensional and aesthetical speech are present. The predicate-emphasizing approach to domain membership allegedly solves the problem of mixed atomics, but the issues with mixed compounds and inferences remain persistent.[16] For example, it is not clear whether the truth-grounding property

---

(2013, 33-34) presents a similar case where "there is no need for the pluralist to sort propositions into strict domains." Does this generate ambiguity? According to Wyatt (2013), no, for we can still hold that sentences that belong to multiple domains have only one truth-grounding property. One can find a reply to Wyatt's argument in Edwards (2018b, 95), who disagrees with both Wyatt's and Lynch's approaches.

[13] For example, take two sentences from the domain of ethics: "killing innocent people is wrong" and "eating meat is wrong." While the former is obviously true, things are not so simple for the latter, since, for example, we now have artificial meat.

[14] Pluralists have largely overlooked the question of how one can account for the domain membership of necessary truths. This subject ought to be explored independently.

[15] Solutions to some of these issues are actively sought in the literature (see Wyatt 2013, 230; Edwards 2018a, 77, 2018b, 85; Lynch 2018, 66).

[16] Lynch (2009, 80) notes that the idea of mixed atomics is self-refuting: "belonging to a particular domain is a feature an atomic proposition at least, has in virtue of being the sort of proposition it is. Propositions are the kind of propositions they are essentially; therefore, belonging to a particular domain is an essential fact about an atomic proposition."

of "snow is white and snow is beautiful" is either correspondence coherence or both.[17]

Mixed discourse provides a suitable case study for illustrating the threat that natural language ambiguity poses for domain-reliant pluralist frameworks. As pluralists seek to offer a definition of truth for natural discourse, and this discourse manifests content mixing in various ways, solutions for clarifying matters will be required if one relies on the notions of domains and domain membership to help achieve an unambiguous and determinate definition of truth. While domain-reliant pluralists have proposed various solutions to the problems involved with content mixing in the context of truth-apt sentences, they have generally neglected a separate yet related issue that follows from the inherently ambiguous nature of certain truth-relevant terms, namely natural language predicates. More specifically, because some of these predicates encompass inherent ambiguity, as is the case, for example, of homonyms, this ambiguity risks carrying over to the pluralist frameworks. To emphasize, insofar as pluralists seek to offer an unambiguous and determinate definition of truth for natural discourse, the inherent ambiguity of some natural language terms should be adequately addressed. Thus far, pluralists have failed to satisfy this requirement, for they have largely circumvented this issue.

In what follows, I use Edwards' (2018a, 78–79) predicate-emphasizing approach to domain membership as a case study to illustrate a strategy that goes beyond the issue of mixed atomics.[18] Thereafter, I show how this approach leads to the above-noted problems with ambiguity and indeterminacy, ultimately conflicting with the standard laws of non-contradiction and identity. According to Edwards, one solution to the problem of mixed atomics is to account for the domain membership of sentences by *predicate kinds*. When

[17] One proposed solution to this issue can be found in Edwards (2018b, 100).

[18] A more general problem emerging from the discussion of this paper, and from the discussions had by various pluralists, is that if one aims for a theory of *truth*, and not only a theory of truth for *atomic sentences*, then the different ways in which *all* types of truth-apt sentences can be assigned to domains should be accounted for. Thus far, the literature focuses heavily on atomic sentences specifically.

dealing with atomic sentences of the form "a is F" (snow is white), where "a" (snow) is a singular term that designates a range of objects, and "is F" (is white) is a predicate that attributes a property onto the objects that the sentences are about, it is always the predicate that determines the domain of sentences:

> I will suggest that it is the predicate that determines the domain [of atomic sentences]. We can distinguish between two things: what a sentence is about, and what is said about the thing the sentence is about. A sentence is about its object […] But what makes these things sentences is that there is more: there is something that is said about the things that the sentences are about. […] It is this aspect—the attribution of a property to an object—that makes these kinds of sentences sentences in that they are bearers of content. So, it is not what a sentence is about that we should be considering [when assigning them into domains,] it is rather what is said about the thing the sentence is about. (Edwards 2018a, 78–79; see 2018b, 97)[19]

Thus, according to Edwards, while atomic sentences are always about their objects, the question of truth emerges only after something is said about these objects or a property is attributed to them. In this sense, it is the attribution of a property to an object that renders these sentences truth-apt, and because of this, the predicate ought to be treated as the domain-determining factor. From this, one can argue for the ideal situation where each predicate kind is tied to a specific domain of sentences. Thus, by instantiating a predicate kind, truth-apt sentences belong to distinct domains to which the adequate truth-grounding properties are tied. In general, the method of choosing either the singular term or the predicate kind as the domain-determining factor of sentences offers an answer to the following questions:

i. How are sentences and domains individuated?

ii. What are the necessary and sufficient criteria for accepting and rejecting sentences into domains?

---

[19] Edwards (2018a, 79) continues, claiming that "the singular term is not relevant to domain individuation."

However, choosing either the singular term or predicate kind as the domain-determining factor leaves the following question unanswered:

I. How can the domain membership of sentences that instantiate ambiguous singular terms or predicates be accounted for?

In what follows, I argue that, because of the inherently ambiguous nature of some natural language predicates, the domain membership of some sentences ends up being ambiguous and indeterminate in the standard domain-reliant pluralist frameworks.[20] The core of my argument is that, because of the inherently ambiguous nature of some predicates, some sentences end up counting as members of multiple domains with different truth-grounding properties, thus generating confusion regarding the grounds of their truth. More specifically, if there is no clarity on whether a sentence S1 belongs to the domain of D1 or D2 or both, with distinct truth-grounding properties P1 (D1) and P2 (D2), then there can be no determinate answer as to the property in which the truth of S1 is grounded. As I later demonstrate, subsequent problems emerge.

5. Issues with Ambiguity and Indeterminacy in Domain-reliant Frameworks

Domain-reliant truth pluralist frameworks rely on strategies of domain-individuation and account for the domain membership of sentences. As demonstrated earlier, a prominent strategy relies on predicate kinds. Each predicate kind assigns sentences to a specific domain governed by a distinct truth-grounding property. Here, the term "predicate kind" can be understood in two ways. First, predicate kinds can be individuated on semantic grounds, such as subject matter or aboutness. The predicate "is right" denotes a distinctively normative property, rendering sentences about things that are right or wrong, etc., thus assigning them to a specific do-

[20] For practical reasons, I restrict the discussion to those approaches that commit to the predicate-emphasizing approach to domain membership, but the arguments provided should carry over to other approaches, such as those that commit to the relevance of singular terms for domain membership.

main, a viable candidate being that of ethics. Other predicates denote extensional properties, rendering sentences that instantiate them about things that have representational or objective properties, hence assigning them to an appropriate domain, such as physics. Second, predicate kinds can be individuated on ontological grounds, relying on the ontological status of their referents. As the ontological status of the property denoted by "is right" is abstract, the non-extensional, non-representational, projected, non-natural, abundant, etc., sentences instantiating it belong to a domain that covers this type of anti-realist speech. Other sentences have predicates such as "is liquid" that denote extensional, representational, objective, natural, or sparse properties, etc., thus assigning them to a domain that covers this type of realist speech.

As expected, both of these strategies have their strengths and weaknesses. The first strategy is intuitive, but it involves the cumbersome task of individuating predicate kinds on thematic grounds. There is no shortage of natural language predicates, and assigning each of them to some of the numerous thematically individuated domains without ambiguity is a complicated task, especially bearing in mind that, in the optimal scenario, each domain is governed by a single truth-grounding property. For example, distinguishing between moral and religious discourse can be difficult; the same applies to speech about objective properties and aesthetics. In what way does the predicate "is bad" differ from "is sinful," and does the predicate "is a mosaic" assign sentences to the domain of aesthetics, even though it attributes a representational and objective property? The ontology-based strategy suffers less from this issue because it requires only two domains: one for the realist discourse and the other for the anti-realist discourse. For example, predicates that attribute sparse, concrete, representational, extensional, natural, or causally effective properties assign sentences to a realist domain governed by an appropriate truth-grounding property, such as correspondence, while those attributing abundant, abstract, non-representational, non-extensional, or non-causal properties assign them to an anti-realist domain governed by another truth-grounding property, such as coherence or superwarrant. Regardless of the strategy, the preferred outcome remains the same. To avoid ambiguity, each sentence

must belong to a distinct domain with a single truth-grounding property.

One issue with the predicate-emphasizing approach to domain individuation and membership that plagues both semantic and ontology-based strategies follows from the inherently ambiguous nature of some natural language predicates. This ambiguity comes in two kinds. First, some predicates are thick, meaning that they play both evaluative and descriptive roles. For example, courageousness ("is courageous") can be interpreted as a virtuous property with clear moral or prescriptive implications. Conversely, courageousness implies a tendency to act in the world, which is a causally relevant property. Thus, it is not obvious whether sentences such as "Charlie is courageous" are subject to a realist (correspondence) or anti-realist (coherence) criterion for truth (see Edwards 2018a, 79–80). Second, and more central to the discussion at hand, some predicates allow for multiple readings. Even a simple predicate such as "is white" is open to different readings because it encompasses a degree of ambiguity. It can be read as denoting the extensional property of having a certain *color* ("snow is white") or perhaps the social property of belonging to a specific *social class* ("Charlie is White"). From this homonym-based ambiguity follows that one and the same predicate potentially assigns sentences into the distinct domains of physical and social speech or speech about extensional and non-extensional properties. Take the following atomic sentence as instantiating said predicate:

Ambiguous: "Donald Trump is white"

Assuming this to be a truth-apt sentence, there seems to be no initial way of telling whether it is about Trump's physical color or the social class to which he belongs. Another way to illustrate this ambiguity is to use the notions of literal and implicit readings. Let us assume that the literal reading of *Ambiguous* is the physical reading and that the social reading is implicit. According to this strategy, *Ambiguous* claims that Trump is physically white, and it is implied that he belongs to the appropriate social class of White people. However, these are radically different understandings of one and the same sentence, with the only similarity being that they are both about Trump. What about a person of native African descent

who suffers from albinism, rendering their skin color white? Here, a literal claim of them being white cannot imply that they belong to the analogous social class. While the literal reading would be true, the implied reading would be false. Further, in the case of *Ambiguous*, the readings can just as well be the reverse. Nothing in the sentence itself indicates what the possible readings are and which of them ought to be treated as correct or primary from the perspective of domain membership. Of course, the utterer knows what they mean by a given sentence, but this is not necessarily evident to anyone beyond them, not to mention the independent issues that plague approaches that commit to treating utterances as truthbearers.

One problem that the *Ambiguous* example generates in the standard domain-reliant pluralist frameworks is that the truth-grounding property for the domain of physical or realist speech is *different* from that of social or anti-realist speech. It is widely held that speech about physical or extensional states of affairs is governed by a correspondence criterion. "Snow is white" is true iff the object designated by "snow" has the property predicated by "is white." Here, truth depends on the connection that linguistic entities have with the relevant objective states of affairs. Speech about social properties is not governed by the same criterion. For example, correspondence does not exhaust why a person belongs to a specific social class. As illustrated in the example of the native African with albinism, one's skin color does not determine their membership to a particular social class. Rather, it is a matter of coherence with other true beliefs regarding one's identity, culture, heritage, and opinions that contributes to their inclusion in or exclusion from these types of classes. This indicates that speech about social properties is governed by something other than a correspondence criterion, the viable alternative being coherence.

However, from this two-way ambiguity of physical and social readings follows a more serious problem for domain-reliant pluralists. If *Ambiguous* belongs to the domain of physical or realist speech that is governed by the truth-grounding property of correspondence, then it fails to be true. This is because Trump is physically *orange*; therefore, the sentence fails to correspond. Nevertheless, if this sentence belongs to

the domain of speech about social properties that is governed by an anti-realist criterion of coherence, then it turns out to be true, for Trump, indeed, belongs to the appropriate social class. Is this ambiguity harmless? There are a couple of reasons for thinking that the answer is negative. Take the standard law of non-contradiction that many see as a necessary condition for *any* truth definition:

Law of non-contradiction: No sentence is both true and false.

The *Ambiguous* sentence turned out to be both true and false in the standard domain-reliant pluralist frameworks. The reason is that the predicate "is white" allows for multiple readings, assigning the same sentence to distinct domains of speech about physical and social properties, whereas by possessing one of the relevant truth-grounding properties and failing to have the other, simultaneous truth and falsity emerge. Note that correspondence and coherence are both distinct truth-grounding properties, and they ground truth separately for the relevant domains. Because lacking the relevant truth-grounding property for the domain that a sentence belongs to constitutes falsity, *Ambiguous* emerges as both true and false. It is worth emphasizing that the truth and falsity of sentences is dependent on their ability to possess the relevant truth-grounding properties because the possession of the general truth property F is determined by the ability of the sentence to possess the relevant truth-grounding property. According to pluralists, the grounds of truth are many, a claim that ought to be taken seriously. The unfortunate result seems to be that, for some sentences, ambiguity emerges regarding the grounds of their truth. Finally, it is important to realize that the noted issue with simultaneous truth and falsity concerns both semantic and ontology-based individuation strategies. The ambiguous predicate "is white" (white in color) can attribute an extensional or representational property, thus assigning a sentence to a realist domain of speech about extensional states of affairs. However, the same predicate "is white" (member of social class) can predicate a non-extensional or non-representational property, assigning a sentence to an anti-realist domain. Assuming that these domains are governed by distinct truth-grounding properties, the *Am-*

biguous sentence once more emerges as both true and false, even according to the ontology-based strategies.

Interestingly enough, the troubles for domain-reliant pluralists do not end here. It also follows that the fundamental law of identity becomes contradicted in the standard domain-reliant pluralist frameworks when supplemented with ambiguous predicates. Take the standard law of identity:

> *Identity*: S is identical to S

From which we can trivially infer that:

> *Identity schema*: If "S" (sentence) is true, then "S" (sentence) is true.

or

> *Identity schema instance*: If "Donald Trump is white" is true, then "Donald Trump is white" is true.

Furthermore, the latter inference emerged as false in the domain-reliant scheme, for the left- and right-hand sentences allowed for different readings, assigning one and the same sentence to distinct domains with different truth-grounding properties and, at the same time, having one of these properties and lacking the other. Thus, in addition to conflicting with the standard law of non-contradiction, even the fundamental law of identity becomes compromised in the standard domain-reliant pluralist frameworks when supplemented with the inherently ambiguous natural language predicates. In what follows, I discuss these results.

6. Discussion

What options are there to resolve the above-mentioned issues? The initial option is to simply accept that ambiguous predicates assign sentences to multiple domains. However, this leads directly to the issue of mixed atomics, compromising the goal of an unambiguous and determinate definition of truth. If some sentences belong to multiple domains with different truth-grounding properties, or there is no clarity as to which of the possible domains they ought to be read as belonging to, then no determinate answer can be given to the question regarding the grounds of their truth. Simply put, if a

predicate assigns a sentence to the distinct domains D1 and D2 with different truth-grounding properties, then the question emerges as to which of these domains ought to be treated as primary from the perspective of truth-grounding. No simple answer is forthcoming.

Another option is to treat sentences with ambiguous predicates not as single sentences but as compounds. These types of ambiguous sentences can be treated as conjunctions or disjunctions of sentences rather than individual sentences. The sentence "Donald Trump is physically white and Donald Trump is socially white" would be false, while the sentence "Donald Trump is physically white or Donald Trump is socially white" would be true. Here, a crucial step has been taken regarding the disambiguation of the original *Ambiguous* sentence. There is no guarantee that, in the case of natural discourse, this step is taken, and if this is assumed, then there are good grounds to argue that we are no longer operating in the domain of natural discourse. Rather, we are speaking about some regimented or disambiguated subsection of natural discourse, and thus, the goal of offering a complete definition of truth for natural discourse is not met. In any case, it seems that solving the issue of ambiguous predicates with the help of conjunction- or disjunction-based strategies rests on the assumption that the ambiguous predicates can be, or are, disambiguated.

Indeed, if the pluralists were to adopt a regimentation or disambiguation strategy, then they would have to re-frame their program as offering a definition of truth for a regimented subsection of natural language. However, this conflicts with one of the major commitments of current pluralist frameworks. Recall the platitude-based strategy for defining the general truth property F that all true sentences have and all false sentences lack, which is denoted by the predicate "is true." According to this strategy, the general truth property inherits its nature from our *common-sense beliefs* and *intuitions* about the concept of truth. Thus, the platitudes are aimed at capturing our pre-theoretical and "naturally" emerging concept of truth. According to pluralists, our pre-theoretical conception of truth is accessible through certain platitudes about the notion that we use as a collective definition for the general truth property. In this sense, pluralists are not talking

about a regimented conception of truth or a restricted understanding of what it means to be a true sentence. If one wants a definition of truth for natural discourse, then it ought to be consistent with the natural or pre-theoretical ways in which truth appears in our cognitive lives. Thus, regimenting the scope of truth-apt sentences generates conflict with one of the major commitments of the pluralist program in seeking a definition of truth that is consistent with its pre-theoretical nature, that is, given by common-sense platitudes.

Of course, one could argue that the issues regarding natural language ambiguity are not only a problem for pluralist or domain-reliant pluralist frameworks but for the entire range of definitions of truth for natural language discourse. One issue with this counter-argument is that, while it is indeed the case that natural language ambiguity generates problems for various types of truth definitions, many of them seek to resolve these issues by regimenting the target language and ruling out ambiguous terms. For example, one might adopt a position of truth-apt minimalism, according to which the units of truth are restricted in a way that suspicious sentences, such as those with ambiguous predicates, are cast out of the question regarding truth or falsity. This type of project can be found in Quine (1992, 78–79), according to whom only eternal sentence tokens are to be treated as truthbearers. These types of sentences are not permitted to include troublesome terms, such as ambiguous predicates. Again, however, from the perspective of the pluralist program, the problem with accommodating the Quinean approach is that we do not commonly see *only* eternal sentence tokens as truthbearers. The sentence "Donald Trump is white" is surely not an eternal sentence, and both of the senses in which it can be interpreted are truth-apt in common discourse. Entities can possess distinct colors and can belong to distinct social classes. The problem is that we do not always know the ways in which all truth-apt sentences should be interpreted, and this ambiguity is very much in line with the richness of meaning that is an inherent feature of natural discourse. Semantic richness is one of the reasons why natural languages are such useful communication systems in the first place, enabling a wide range of expressive and descriptive functions. Insofar as a definition of truth is directed at natural discourse, as the

pluralist program surely seems to be, then the potential issues with ambiguity should be a top priority for examination. However, pluralists have hitherto said very little about the inherent ambiguity of natural discourse and the problems it generates for their definitions, even while setting the goal of achieving an unambiguous and determinate definition of truth for said discourse.

Finally, I want to make a brief note about an approach to defining truth that shows promise in avoiding the already noted issues generated by natural language ambiguity, albeit still retaining the virtue of enabling the accommodation of both realist and anti-realist intuitions. One could aim to construct a Tarski-style truth definition for a regimented subsection of natural discourse that would obviously be incomplete because of the paradoxes and infinite semantic ascent. Beyond this, however, as given by the Tarskian paradigm, one would end up with a definition that gives general and scaling criteria for the truth of all truth-apt sentences. Take the Tarskian T-schema where each sentence provides its own conditions for being true:

> *T-schema*: X is true iff p[21]

or

> *T-schema instance*: "Donald Trump is white" is true iff Donald Trump is white.

Indeed, the Tarskian paradigm allows for both coherence and correspondence readings. As such, there is no in-principle reason for why it could not be used to construct a definition that allows for both realist and anti-realist ways of being true. In this sense, supplementing it with a distinctively pluralist thesis is a worthy path of inquiry.

Of course, there are central differences between the Tarski-based approach and current domain-reliant pluralist frameworks. One important difference is that Tarski's account does not commit to using domains as an explanatory resource for

[21] Tarski's (1944, 344) explication of the T-schema reads: "We shall call any such equivalence (with 'p' replaced by any sentence of the language to which the word 'true' refers, and 'X' replaced by a name of this sentence) an equivalence of the form (T)."

defining truth. Because it treats individual sentences as truthbearers, no commitment to discourse domains is required. From this follows that the Tarskian approach does not fall victim to the noted ambiguity issues emerging in the domain-reliant frameworks. Independent of this, the project of defining domains is strictly non-truth-theoretical in the first place, and thus, there is no in-principle reason why a definition of truth should commit to it. Of course, as domains can be understood as simple classes of sentences, avoiding them altogether seems unnecessary. Indeed, even acknowledging different ways of being true would constitute domains. One key difference between the domain-reliant pluralist models and the Tarski-inspired approach is that one can either accept that a definition results in the existence of domains or that a definition can utilize the notion of domains in accounting for the truth of sentences. As demonstrated throughout this paper, there are reasons for being suspicious about the latter path. Because of space limitations, I shall delay further discussion on the prospects of forming a domain-free pluralist definition in the spirit of Tarski's semantic conception of truth.

Finally, one note from the perspective of an unambiguous and determinate pluralist definition of truth arising from the comparison of current pluralist models and the Tarskian approach is that many of the issues with natural language ambiguity that pluralists face follow from their confidence in committing to a strict grounding claim. Pluralists are not only satisfied with offering general criteria for the truth of sentences; they seek to offer a scaling, unambiguous, and determinate definition of the grounds of truth on the level of natural discourse. The Tarskian approach simply provides general criteria for the truth of each sentence. There is no direct answer to the question of in what is the truth of each true sentence grounded in. Thus, the Tarskian approach is satisfied with a less specific definition, and for good reason. Tarski was well aware of the problems involved with offering a complete definition of truth for natural discourse, one reason being the inherent ambiguity and vagueness of natural language terms. Indeed, in this sense, Tarski can be interpreted as giving a reason why a determinate and scaling definition on the grounds of truth for natural language sentences

cannot be given. Indeed, in light of our discussion, the issues generated by natural language ambiguity for definitions of truth in general seem to intensify the more a definition of truth commits itself to explaining. A criterial definition that makes strict grounding claims is faced with the issue of natural language ambiguity if it subjects itself to offering an unambiguous and determinate definition of truth. Other less ambitious definitional paths seem to face this issue to a lesser degree, but exploring the full scope of this idea deserves an independent study. I hope that at least some of the current findings will aid future examinations.

University of Jyväskylä

References

Bar-On, D. & Simmons, K. (2018), "Truth: One or Many or Both?", in J. Wyatt, N. Pedersen, and N. Kellen (eds.): *Pluralisms about Truth and Logic,* London: Palgrave Macmillan, pp. 35-61.

Beall, J. (2013), "Deflated Truth Pluralism", in N. Pedersen and C. Wright (eds.): *Truth and Pluralism: Current Debates*, Oxford: Oxford University Press, pp. 323–338.

Blackburn, S. (2013), "Deflationism, Pluralism, Expressivism, Pragmatism", in N. Pedersen and C. Wright (eds.): *Truth and Pluralism: Current Debates*, Oxford: Oxford University Press, pp. 263–277.

David, M. (2013), "Lynch's Functionalist Theory of Truth", in N. Pedersen and C. Wright (eds.): *Truth and Pluralism: Current Debates*, Oxford: Oxford University Press, pp. 46–68.

Edwards, D. (2011), "Simplifying Alethic Pluralism", *The Southern Journal of Philosophy* 49(1), pp. 28–48.

Edwards, D. (2018a), *Metaphysics of Truth*. Oxford: Oxford University Press.

Edwards, D. (2018b), "The Metaphysics of Domains", in J. Wyatt, N. Pedersen, and N. Kellen (eds.): *Pluralisms about Truth and Logic,* London: Palgrave Macmillan, pp. 85–106.

Kim, S. and Pedersen, N. (2018), "Strong Truth Pluralism", in J. Wyatt, N. Pedersen, and N. Kellen (eds.): *Pluralisms in Truth and Logic*, London: Palgrave Macmillan, pp. 107–131.

Lynch, M. (2009), *Truth as One and Many*, Oxford: Oxford University Press.

Lynch, M. (2013), "Three Questions for Truth Pluralism", in N. Pedersen and C. Wright (eds.): *Truth and Pluralism: Current Debates*, Oxford: Oxford University Press, pp. 21–41.

Lynch, M. (2018), "Truth Pluralism, Quasi-Realism, and the Problem of Double-Counting", In W. Jeremy and N. Pedersen and N. Kellen (eds.): *Pluralism in Truth and Logic*, London: Palgrave Macmillan, pp. 63–84.

Pedersen N. and Wright C. (2013), *Truth and Pluralism: Current Debates,* Oxford: Oxford University Press.

Pedersen, N. and Wright, C. (2013), "Pluralism about Truth as Alethic Disjunctivism", in N. Pedersen and C. Wright (eds.): *Truth and Pluralism: Current Debates,* Oxford: Oxford University Press, pp. 87-112.

Pedersen, N. and Lynch, M. (2018), pp. 87–113, "Truth Pluralism", in M. Glanzberg (ed.): *The Oxford Handbook of Truth,* Oxford: Oxford University Press, pp. 543–575.

Pedersen, N., Wyatt, J. and Kellen, N. (2018), "Introduction", In J. Wyatt, N. Pedersen, and N. Kellen (eds.): *Pluralism in Truth and Logic,* London: Palgrave Macmillan, pp. 3–35.

Pedersen, N. 2020, "Moderate Truth Pluralism and the Structure of Doxastic Normativity", *American Philosophical Quarterly* 57(4), pp. 355–376.

Quine, W.V.O. (1960), *Word and Object*, Cambridge, Mass.: M.I.T. Press.

Quine, W.V.O. (1992), *Pursuit of Truth*, Cambridge, Mass.: Harvard University Press.

Sher, G. (2005), "Functional Pluralism", *Philosophical Books* 46(4), pp. 311–330.

Tarski, A. (1944), "The Semantic Conception of Truth: and the Foundations of Semantics", *Philosophy and Phenomenological Research* 4(3), pp. 341–376.

Wright, C. (1992), *Truth and Objectivity*, Cambridge, Mass., Harvard University Press.

Wyatt, J. (2013), "Domains, Plural Truth, and Mixed Atomic Propositions", *Philosophical Studies* 166(1), pp. 225–36.

Wyatt, J. and Lynch, M. (2016), "From One to Many: Recent Work on Truth", *American Philosophical Quarterly* 53(4), pp. 323–340.

Wyatt, J., Pedersen, N. and Kellen, N. (2018), *Pluralism in Truth and Logic*, London: Palgrave Macmillan.

The Concert of Forces

JAN HAUSKA

Introduction

Although it is universally accepted in Newtonian physics, the principle of the composition of forces has experienced a certain reversal of fortune over the past century or so. Whereas the close of the 19th century saw persistent attempts to show *why the principle is true* by deriving it from more basic tenets (cf. Lange 2009, 397–414), nowadays the question is *how it can be true,* or what is supposed to make it true (e.g. Cartwright 1983, 54–73; Creary 1981; Massin 2017, 808). The principle, also known as the parallelogram law, says that when a number of forces act on a body, they add vectorially. In particular, when two such forces, called *component* forces, differ in direction, the *resultant* force (i.e. their vectorial sum) is represented by the diagonal of a parallelogram whose two sides stand for the component forces, as shown in Figure 1. The direction of the acceleration imparted to the body is then the same as the direction of the resultant force, and the magnitude of the acceleration is given by the second law of motion ($F = ma$) applied to the force.

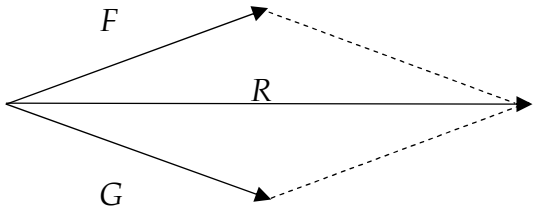

Figure 1. The parallelogram of forces.

Taken at face value, the parallelogram law is committed to the reality of all the forces it mentions. But then, it has been argued (Creary 1981, 151–153; Wilson 2009, 536–540; cf. Massin 2017, 811–812), their effect – the acceleration of a body – would be overdetermined: both the component forces and the resultant would be sufficient to bring it about. The objection presumes that the component forces would produce the same acceleration as the resultant force (Wilson 2009, 533–534; Massin 2017, 818–820), which might be called into question. If they caused their distinct accelerations, however, then the body would have more than one acceleration, which is impossible. Furthermore, if the resultant force acted on the object in conjunction with the component forces, it would have to be added to them, which would yield another resultant force, giving rise to a vicious regress (Hüttemann 2004, 105; see Massin 2017, 812).

Various routes have been taken in response to these difficulties by philosophers who recognize the existence of forces (as I will provisionally do here). It was claimed, at first approximation, that there are only resultant forces (Cartwright 1983, 54–73; Wilson 2009), that there are only component forces (e. g. Creary 1981; Molnar 2003, 194–198), and that both kinds of forces manage to coexist peacefully (Massin 2017, 828–843). In what follows, I will examine the last response[1] before attempting to show that if forces are taken to be powers, a plausible picture of their composition emerges.

Residualism and Remaining Difficulties

Massin's residualism (Massin 2017, 829–30 and 840–42) attributes two essential causal powers to forces. The powers are taken to be dispositional properties individuated – in conformity with the so-called conditional analysis of this kind of property (see e.g. Molnar 2003, 83–94) – by some characteristic events (their *manifestations*) and the circumstances (their *activating conditions*) which contribute to producing the events.[2] Each force has the power to "bring about accelera-

[1] For an excellent discussion of the other accounts, see Massin (2017, 808–828).

[2] What Massin says about the dispositions is compatible with a few variants of the conditional analysis, perhaps most naturally with the Causal

tions of the body it acts upon". This *kinematic* power is invariably accompanied by a *static* one – the power "to prevent antagonistic forces (same magnitudes, opposite directions) from causing the acceleration of the body it acts upon". Thus, the static power of a force is triggered by the presence of an antagonistic force, and its manifestation consists in preventing the antagonistic force from causing an acceleration. The activating condition of the kinematic power, on the other hand, is said to be the absence of an antagonistic force. It follows that when the trigger of one of the powers occurs, the trigger of the other one does not, and *vice versa*. Accordingly, forces "necessarily exert one and only one of their two powers".

The twin dispositional properties of a force are not assigned a role in bringing about their manifestations.[3] That role is said to be played by the force in conjunction with the activating conditions of the dispositions (Massin 2017, 810, 829–830 and 840–841). The metaphysical profile of forces is not drawn in perfect detail, however. While denying that they are dispositions (Massin 2017, 810), Massin maintains that they are symmetric relations between (for the most part) bodies (Massin 2017, 810; 2009, 581)[4], relations whose causal involvement does not mean that they are a species of causal connection (Massin 2009, 582–587). As they are not regarded as spatio-temporal relations either (Massin 2017, 810; 2009,

---

Conditional Analysis. According to this version (see Molnar 2003, 89–90), an entity has a disposition at time *t* to give rise to manifestation *m* in conditions *C* if and only if it has some property which – were the entity to be in conditions *C* at time *t* – would join with the conditions to produce event *m*. Thus, a vase has a disposition at time *t* to break when struck with moderate force (or is fragile) just in case it has a property which – were the vase to be struck with moderate force at time *t* – would combine with the striking to produce a breaking of the vase.

[3] The claim that a disposition is not involved in producing its manifestation is fully consistent with the conditional analysis, which is silent on whether the complete cause of a manifestation includes the disposition (as opposed to a closely associated property – its distinct *causal basis*).

[4] The reason for which Massin holds that forces are relations is that they have direction (Massin 2009, 565–574). Their symmetrical character is in turn said to follow from their involvement in the third law of mechanics (the action-reaction law) (574–582).

559), it is unclear what exactly their positive nature is supposed to be, and in particular whether they are taken to form a *sui generis* metaphysical category.

When only antagonistic forces act on a body, which leaves its movement (or lack thereof) unchanged, each of them is said to counteract the other completely, that is, to prevent it from displaying its kinematic power. No resultant force is then present (Massin 2017, 830–831). When the forces that act on a body are or include ones which are not antagonistic, at least some of them are not disarmed in this manner: they remain active and are considered to be partly identical (or identical *tout court* if there is only one of them) with the resultant force. This is captured by the thesis, central to residualism that "the resultant force acting on a body is identical to the (sub-) component force or forces that do not prevent each other from bringing about the acceleration of the body. The resultant force is then said to be a residue of the forces pertaining to a body (829). A kind of concurrent existence of both component and resultant forces is thus espoused (824–825).

When only a single force acts on a body, its identity with the resultant is straightforward (Massin 2017, 831). Similarly, maintains Massin, when two forces with the same direction influence an object, then the resultant force "is nothing but the component forces together". The mode of their addition is said to be either standard mereological composition (for the forces have the same direction and "their summation only concerns their magnitudes") or primitive vectorial composition (831–832). (Standard vectorial composition is ruled out as giving rise to the problems mentioned above.) Further, when opposite forces (opposite directions, different magnitudes) act on a body, the smaller one is taken to prevent a part of the larger one from exercising its kinematic power. The other part is then left unimpeded, causes an acceleration, and thus amounts to the resultant force. Put another way, the larger force is supposed to split into co-directional sub-component forces here, of which one prevents the smaller force (and is also prevented by it) from causing an acceleration. The mode of decomposition is, again, regarded as either primitively vectorial or standardly mereological (832–833). Finally, the case of non-colinear component forces, usually depicted by means of the (non-flat) parallelogram of forces, is dealt with by as-

serting that there is only one natural decomposition of the component forces. Some of the sub-component forces are then antagonistic whereas the others turn out to be co-directional. The sum of the later is the resultant force, as shown in Figure 2 (833). (This approach is also applied in the case of more than two non-colinear forces (Massin 2017, 833–835).)

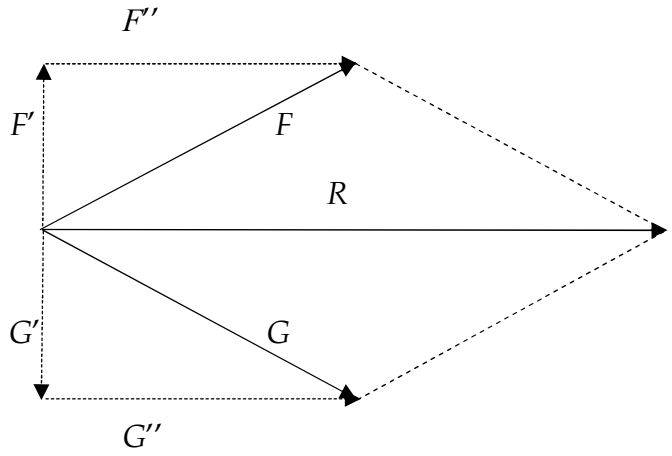

Figure 2. Composition of non-linear forces: the sum of *F''* and *G''*. (Source: Massin 2017, 835)

As just seen, residualism regards the presence of an antagonistic force as the activating condition of the static power of a force. But what is the manifestation of the power? The power is supposed to prevent the display of the kinematic power of the antagonistic force, and yet there is no indication that it would do so by *acting* in any way – for example, by removing a condition, causing an event which interferes with a causal process, or producing an event incompatible with the manifestation of the kinematic power. It appears that the mere presence of the force would be sufficient to render the kinematic power of the antagonistic force inoperative. But then the behaviour of a force would be fully accounted for by the kinematic power, which would cause an acceleration in the absence of an antagonistic force and be idle otherwise. It is therefore questionable whether there is any good reason for

positing the static power of forces. (At one point Massin (2017, 841) contends that "the manifestations of the static power of forces amount to stresses (pressures and tensions), which can be felt". One could then claim that the static power prevents the display of its kinematic cousin by producing an event incompatible with the manifestation of the latter, that is, with an acceleration. However, the proposition that stresses are non-kinematic effects of forces seems tantamount to supplementing the second law of motion with another dynamical principle. Such a philosophical revision of physics should be eschewed, especially since pressures and tensions appear to be amenable to description in terms of the second law. A pressure, for example, occurs when a force acts on some molecules, pushing them – in accordance with the law – closer to their neighbours, and it continues to be present when the force becomes balanced. Thus, the pressure is a kinematic effect of the force and then it remains in place not because the force has some other effect, but because no force brings about a movement of the molecules towards decompression.)

One of the difficulties of residualism addressed by Massin (2017, 839) arises when a force, F, which first acts alone on a body, is joined by a non-colinear force, G, at time t_1. As the only force on the scene before t_1, F is both component and resultant and "is not composed of any actual sub-forces". According to residualism, however, after t_1 the original force has orthogonal sub-component forces, F' and F'' (as shown in Figure 2). Owing to their distinct directions, the sub-component forces cannot compose in the standard mereological manner. Their composition would then be primitively vectorial, which Massin (2017, 825) takes to entail that force F would be nothing over and above them. It therefore appears to be a ramification of residualism that whether a force has components and what the components are depends on the presence of concurrent forces. Yet "how can a force remain the same force if its essential components change over time"?

Massin (2017, 839–840) attempts to escape the difficulty by embracing the view that the sub-component forces recognized by residualism arise from a *sui generis* breakage of component forces. The body on which non-colinear component forces act would make them break into sub-forces: each

of the them would split into orthogonal forces and go out of being. (The antagonistic sub-component forces, F' and G', would then counteract each other while the co-directional ones, F'' and G'', would produce an acceleration.) As the component forces would no longer exist, they would not be constituted by the sub-component forces: at no time would force *F*, which is supposed to act alone at first, consist of the sub-component ones, for it would be *replaced* by them at breakage. Thus, it would not be true that one force could have different essential components over time depending on context.

As Massin (2017, 840) recognizes, this response faces a dilemma. If the breakage of force *F* occurred *after* force *G* begins to act on the same body, then there would be a time at which both forces would act without composing with each other: as the sub-component forces would not yet be present, each component force would act on its own. In Massin's view, it would be quite speculative to embrace the reality of such a state "just on theoretical grounds". If the breakage occurred at the very time of the arrival of *G* (i.e. if it were *instantaneous*), however, the component forces *F* and *G* would never act on the body together: the sub-component forces into which they split would do so. The forces which would then compose with each other would not be the ones that we first thought to be involved.

The dilemma poses a greater threat to residualism than Massin seems to appreciate. If its first horn were embraced, then at some time a body would be acted upon by forces which would not compose. The kinematic powers of the forces would thus be displayed, imparting on the body two distinct accelerations, which cannot be (unless accelerations compose, which is even more problematic than force composition (Wilson 2009, 533–534; Massin 2017, 819–820)). On the other hand, the idea that forces split instantaneously, which Massin seems to favour, runs into a contradiction. In general, an instantaneous effect cannot comprise the demise of its cause, for then the latter would have to exist and cease existing at the same time. It follows that a force and the result of its splitting cannot be simultaneously present, especially since the idea of breakage is explicitly introduced to deny the continued identity of forces.

Residualism maintains that the difficulties facing the recognition of both component and resultant forces are circumvented if one appreciates their powers. In some cases, only a single force or co-directional forces are present, which is taken to mean that their kinematic power would be displayed. On other occasions, the static power of some of the forces pertaining to an object is taken to neutralize antagonistic forces, and again only a single force or co-directional forces would be left to exercise their kinematic power. Thus, co-directional forces figure prominently in Massin's theory. They also give rise to a serious difficulty, though.

Whenever co-directional forces are present and none of them is counteracted, they are said to be parts of the resultant force. The mode of their composition, to reiterate, would then be either primitively vectorial or standardly mereological (Massin 2017, 831–833). But even if it were granted that the forces would compose in one of these ways, it would remain true that they would retain their identity. Essential to the latter would be their kinematic power to impart an acceleration to the object they act on. Again, however, the object cannot have two or more accelerations. The difficulty goes to the heart of residualism since it is co-directional forces which are supposed to embody the idea that component forces are partly identical with the resultant. As co-directional forces cannot merge in a way which would dissolve their identity – for then only the resultant force would be present – they have to display their kinematic powers, which lands residualism in trouble.[5]

[5] It may be worth noting some of the difficulties facing the two modes of composition which Massin regards as plausible candidates for applying to co-directional forces. First, it is not clear what primitive vectorial composition is supposed to be metaphysically and what laws it would obey. In an echo of the controversial claim that the composition of material objects is a form of identity, primitivism has it that resultant forces are "nothing over and above the component forces" (Massin 2017, 825). At the same time, the view is said to take the parallelogram of forces at face value (826). Yet in many cases the parallelogram unambiguously represents the resultant as distinct from the component forces. The inadequacies of primitivism acknowledged by Massin (826–828) aside, this suggests that the positing of primitive vectorial composition in the case of co-directional forces (where the inadequacies seem to be absent) would provide no reason for

Powerful Composition

The debate about the existence and nature of forces has recently witnessed the emergence of the idea that forces are causal powers (e.g. Cartwright 1999, 52). When invoked in this connection, powers are still supposed to be dispositional properties whose basic traits are captured by the conditional analysis. In contrast to Massin's notion of them, though, they are now ascribed a role in causing their manifestations. The claim is that if a power were in its activating conditions, then the power itself, rather than some closely associated property (or relation), would join with them to bring about its manifestation.[6] If forces are such powers, their profile is fairly precisely delineated, and it is clear that they do not form a new *sui generis* metaphysical category, which amounts to an advantage in economy over residualism. As I am going to argue now, the view also furnishes the parallelogram law with a metaphysical underpinning, issuing in a superior account of force composition. (I will address the question of the justification of the view in the conclusion.)

In the framework of Newtonian physics, when a force acts alone on an object endowed with inertial mass, the result is an instantaneous acceleration of the object (see e.g. Lange 2005, 434). This means that if the mass of the object is a

---

taking the forces to be partly identical with the resultant. Second, as standard mereology is extensional (the same parts can only compose a single whole), it is questionable whether it can be applied to properties and relations if they are universals (e.g. Azzano 2021, 4321–4322). It is also held, though, that tropes would escape this difficulty if they were non-transferable (i.e. if their inhering in their bearer were part of their identity) (Azzano 2021, 4322–4327). This leaves a way out for Massin, who may not embrace universals. Still, to understand what the composition of forces might amount to, one would need to have a clear idea of what it means metaphysically for a relation to be a vector and therefore have a direction. The matter is fairly mysterious, particularly if one holds (Massin 2009, 274–279) that having a direction is different from, and even independent of, being non-symmetrical. (I owe some of the points made in this footnote to a referee.)

[6] The identity of powers is then supposed to be given by their causal aspect (e.g. Bird 2016, 345–346).

power,[7] then the force will be its activating condition and the acceleration will be its manifestation. Moreover, the object could be acted upon by a force of a different strength, and it commonly happens that disparate forces actually act alone on various objects of the same mass. A determinate mass – for example, the mass of one kilogram – will thus have multiple possible activating conditions and accordingly multiple possible manifestations, which is to say that it will be a multi-track power (Bird 2007, 21; see also Vetter 2015, 39–43 and 50–53). One prominent view of the nature of multi-track powers (Bird 2007, 22–24) takes them to be conjunctions of finer powers, each with very specific activating conditions and manifestation. The mass of one kilogram would then comprise the power for its bearer to be accelerated by 1 m/s^2 when acted upon with the force of one Newton, the power to be accelerated by 2 m/s^2 when acted upon with the force of two Newtons, and so on. (I discuss the question of the metaphysical character of multi-track powers in more detail below.)

An analogous reasoning applies to forces. The mass of an object on which a force is impressed will amount to the activating condition of the force, and the resulting instantaneous acceleration will be its manifestation. As the force could act on objects that differ in mass, it will have multiple possible activating conditions and manifestations, and thus be a multi-track power. It would then be natural to regard it as a conjunction of finer powers.

The force acting on an object and the mass of the object are symmetric in an important way: they are powers which activate each other and jointly bring about an acceleration. In the light of this parity in producing a common manifestation, they are regarded as "dispositional partners" (e.g. Martin 1994). Now, their partnership is affected when another force acts on the body and thus plays a role in causing the acceleration. How is this to be accounted for? Clearly, one cannot say that each force separately produces its specific acceleration, since then the object would fall victim to the curse of many distinct accelerations (followed immediately by many distinct

[7] The proposition that inertial mass is a power is supported by the main argument for the view that the properties involved in (fundamental) laws of nature are powers, which I address in the Conclusion.

velocities and positions). The forces cannot give rise to a wholly distinct third force either, for that would only aggravate the metaphysical disorder. Nor – for the reasons given in connection with the idea of force breakage, which have to do with the temporal relation between cause and effect – can the forces be said to lose their identity and merge into another one (the resultant). It seems that the difficulty would be avoided only if the forces preserved their identity while in some way participating in the resultant. But that may strike one as a steep hill to climb: after all, residualism is a sophisticated attempt to describe such a resultant force, and it runs into trouble even though in some cases it trims the component forces (as some of the sub-component forces into which they split counteract each other) before incorporating them into the resultant.

The prospects of the idea that component forces acting on an object participate in the resultant become brighter, however, when all of them are recognized as dispositional partners. On this approach, the mass of the object is a multi-track power which partners with the forces acting on it to produce its instantaneous acceleration. The magnitude and direction of the acceleration will then depend on the mass as well as on the number of the forces and their magnitudes and directions. Since the number could vary, mass is a multi-track power of a multigrade variety. If multi-track powers are regarded as conjunctions of finer powers, then a determinate mass will consist of a great many powers, each of them to partner with a specific number of forces with specific directions and magnitudes to produce a specific manifestation.

A force, on the other hand, is a power which partners with the mass *and the other forces impressed* on the object to cause its acceleration. The magnitude and direction of the acceleration will depend on the force as well as on the magnitude of the mass and the number, magnitudes and directions of the other forces. As the number could vary, the force is a multi-track power of a multigrade variety. If multi-track powers are regarded as conjunctions of finer powers, then a determinate force will consist of a great many powers, each of them to partner with the mass and a specific number of other forces with specific directions and magnitudes to produce a specific manifestation.

When single forces act on bodies, accelerations are brought about in a systematic way: changes in the magnitude of a force or in the mass of the body it acts upon issue in proportional changes in the acceleration of the body. The systematicity is captured by the second law of motion. By the same token, when a number of forces (the "component forces") act on the mass of an object to cause its acceleration, shifts in their magnitudes and directions issue in systematic changes in the acceleration. On the account advocated here, this means that the shifts affect the team of powers which partner to accelerate the body – the team which is referred to as the "resultant force" and whose action is represented by arrow *R*. This systematic relationship is captured by the parallelogram law.

The thesis that powers are individuated by activating conditions and manifestations loses its simplicity in the case of multi-track powers, which in a sense have many activating conditions and manifestations. The force that would produce the acceleration of 1 m per s^2 if it acted on a body of 1 kilogram would also produce the acceleration of .5 m per s^2 if it acted on a body of two kilograms, and the acceleration would be different still if the force partnered with another. This gives rise to difficulty with identifying multi-track powers. In the case of powers which behave in a highly systematic way, the difficulty is circumvented by focusing on a single activating condition and the corresponding manifestation. Thus, forces are identified by specifying what acceleration they would produce if they acted *alone* on a body whose mass is one kilogram.

It is forces identified in this way which are represented as the component forces in the parallelogram. Thus, the magnitudes and directions of them represented by the diagram are not manifested in the case to which the diagram refers. (They are manifested when the forces act solo.) The parallelogram shows the manner in which a number of powers that are forces, specified by how they would act alone on a body of one kilogram, would behave in different circumstances, namely when they act on such a body in concert. In other words, the parallelogram law tells us how the magnitude and direction of the action of the whole team of forces (i.e. of the "resultant force") relates to those of its members (the "com-

ponent forces"), where the members are described by how they would act in isolation. In keeping with the approach to the member forces, the action of the team, represented by arrow *R*, is specified by reference to a body whose mass is one kilogram. (It is a crucial feature of multi-track powers that, depending on their partners, they may produce disparate manifestations. A body of a certain mass, for example, can be made to accelerate at dissimilar rates. It is not therefore surprising that the action of a team of forces may differ considerably from the 'solitary' behaviour of some or even all of its members.)

This approach seems capable of providing a detailed explanation of antagonistic forces. If there were a property of zero acceleration (see Balashov 1999, 260–276), the forces would combine to produce it. If, on the contrary, such a property did not exist (e.g. Massin 2017, 827–828), then antagonistic forces would not produce any event. Unless one embraces the immensely controversial claim that absences can be effects, this would imply – given the lack of any candidates for interfering factors – that the activating conditions of the forces would not then obtain. Accordingly, a force would be displayed if it acted alone on an object or if the other forces present did not include its antagonistic force. In other words, the antagonistic force would inhibit the display of the force by its mere presence, without bringing about any event. Notice that the description does not run into the trouble which undermines Massin's thesis that the static power of a force prevents the kinematic power of its antagonistic counterpart from manifesting. This is because the manifestation of the force is not supposed to be such a prevention – in other circumstances, the force would play a role in causing acceleration and thus would be displayed.

Satisfying desiderata

The debate about the composition of forces has crystallized a number of desiderata which a successful account should satisfy. It is maintained, to begin with, that such an account ought to entail semantic systematicity of the term "force". The contention is that the term is to be employed in the same manner when it occurs in various nomic statements of New-

tonian physics (Wilson 2009; Massin 2017, 812–813). Otherwise, the concept at the heart of the statements, and *ipso facto* the statements themselves, would not be provided with a uniform interpretation. This means that if component forces, which are referred to in the formulae expressing special force laws, are recognized as real (as opposed to fictional), then the same should go for resultant forces, which are referred to in the statement of the second law. It is not difficult to see that the desideratum is met by the dispositionalist account of force composition. Since the account clearly recognizes the existence of component forces and takes them to serve as members of the team which amounts to the resultant force, it is committed to the reality of both.

It is a plausible constraint on an account of force composition that it should provide a "metaphysical answer as to why a given unitary acceleration follows from a chaotic swarm of component forces" acting on a body (Massin 2017, 818). Otherwise, the employment of the parallelogram to predict the magnitude and direction of the acceleration would just be an epistemic trick with no ontological grounding. The dispositionalist account satisfies this requirement by identifying forces with multi-track powers and regarding them as dispositional partners. This provides an explanation of how a multitude of forces which have diverse directions when acting alone, act in one direction when they team up, or how the component forces are related to the resultant one. Thus, while embracing the idea that the dependence of the resultant on the component forces is described by the parallelogram of forces (and *ipso facto* accepting vectorial composition), the account points to the metaphysical mechanism of the dependence. It is a further virtue of the account that it does not go much beyond what is largely accepted concerning multi-track powers.

Newtonian physics apportions causal responsibility, which is to say that it specifies which of the forces acting on an object play a greater role in producing its acceleration and in what proportion. It is therefore maintained (Massin 2017, 820–823) that an account of force composition should be able to tell what causal contribution is made by a force when it acts on an object together with other forces. The dispositionalist account measures up by embracing the parallelogram of

forces. It can then be shown how a change in a component force affects the resultant, which is tantamount to apportioning causal responsibility. (Given the linearity of their relationship, if a 50 percent diminution of the component force translates to a 10 percent diminution of the resultant, it can be inferred that the causal contribution of the former to the latter amounts to 20 percent.)

It has been argued (Wilson 2009, 535–536 and 546–547) that it is resultant forces, rather than component forces, which are experienced in cases where one might *prima facie* think that many forces act on a body. The claim is that even when there are "multiple influences" (i.e. when many force laws are in operation), we experience "forces associated with a single magnitude and direction, that directly result in our accelerations". For example, when a magnet held in a hand is attracted by another magnet, we experience a single force rather than separate magnetic and gravitational ones. This is taken to speak against the existence of component forces and therefore against the accounts which recognize them.

Wilson's description may be adopted by the dispositionalist account of force composition even if it is granted that what one experiences is forces rather than accelerations or their effects (i.e. displacements in one's body brought about by a velocity which is in turn caused by an acceleration[8]). The account implies that we would then experience only one cause of an acceleration (and ultimately of a bodily pressure or tension). This is because component forces produce 'their own' manifestations only when they act alone. By contrast, when they act in concert, they are united in the manner characteristic of a single cause and produce just one acceleration. They would not then be experienced as separate forces.

Multi-track Trouble?

As already noted, the account of forces as powers presumes that their basic traits are captured by the conditional analysis of dispositional properties. At first approximation, a force that is not accompanied by other forces would be described by the causal version of the analysis (see Molnar 2003, 89–90)

[8] For more on the causal roles of acceleration and velocity in Newtonian physics, see Lange (2005, 434 and 452–461).

as a property which – were it in the presence of an object endowed with inertial mass (i.e. in its activating condition) – would join with the mass to produce the acceleration of the object (i.e. the manifestation). The conditional analysis of dispositions has come under fire (e.g. Martin 1994; Bird 2007, 27–29), however, which threatens the views that embrace it. Various modifications of the analysis have been put forward (e.g. Manley and Wasserman 2008, 73–82)[9], but they cannot be adopted in the case of forces. This is because they go beyond the causal version of the analysis by introducing requirements which have no place in the operation of laws of nature, requirements which are thus extraneous to how forces are to play their nomic role.[10] Put another way, the causal analysis of forces is parallel to relevant nomic statements, and if the analysis is in trouble, then so is the idea that forces are powers. I will first address an objection which attempts to undermine the conditional analysis, including its causal variant, by focusing on its ramifications in the case of multi-track powers. Then I will indicate how the causal analysis can deal with some more entrenched difficulties.

The focus of a recent argument against the conditional approach, put forward by Barbara Vetter (2015, 54–59), is the combination of the theses that properties are (mostly) powers, that "all except the maximally specific [powers] are multi-track", and that the basic traits of powers are captured by the conditional analysis (54). If the nature of multi-track powers "is best or adequately characterised by conditionals", contends Vetter, "then it will be infinitely complex, for it requires an infinity of conditionals". The conditional analysis is thus taken to play a crucial role in underwriting the view that determinate properties (such as inertial mass of a certain magnitude) are conjunctions of maximally specific powers, or that

---

[9] An analysis which I find promising says that an object has an intrinsic disposition at time *t* to give rise to manifestation *m* in conditions *C* if and only if it is nomologically possible that (*i*) the object retains all the intrinsic properties it actually has at time *t* and (*ii*) some of the properties join with conditions *C* to produce event *m*.

[10] This contrasts with one of the reasons for which the accounts of powers which deny that they have activating conditions (e.g. Vetter 2015, Ch. 3; Aimar 2019) cannot be applied to forces. The accounts fail to reflect the involvement of the conditions in the operation of laws of nature.

they are built up from single-track powers. And since "simpler building-blocks are, in whatever area, more fundamental than their complex compounds", the maximally specific powers then have to be recognized as more fundamental than the determinate properties (55).

Aiming to show that this ramification is false, Vetter (2015, 56–57) considers determinate properties with reference to statements of functional laws of physics, which concern mathematical relationships between quantities. Some of the statements, in particular Coulomb's law, are "our best bet" at being near-to-fundamental. The statement succeeds in its explanatory role because it relates determinate properties of electric charge to determinates of distance and force. A nomic statement which would invoke a maximally specific (i.e. single-track) power would explain the regularity which amounts to the occurrence of the manifestation of the power in its activating conditions. And the totality of such statements would explain an infinity of such narrow regularities. Yet, maintains Vetter, they would fail to explain the "much more striking regularity" – the systematic link between these maximally specific regularities. In other words, it would then be inexplicable why the exerted force always stands in the same mathematical relation to the charges and distance involved. As Coulomb's law explains the more striking regularity, it is more fundamental than the maximally specific nomic statements in question. And since "the more fundamental properties figure in the more fundamental laws", determinate multi-track powers are more fundamental than single-track ones. As just seen, the proposition that single-track powers are fundamental is regarded by Vetter as a consequence of the trio of theses mentioned above. If the proposition is false, the theses cannot all be true. Since Vetter embraces dispositionalism and the ubiquity of multi-track powers, she pins the blame for the falsehood of the proposition on the conditional analysis.

Focusing on properties themselves rather than the relevant laws of nature, Vetter's (2015, 57–58) second argument departs from the proposition that "[i]nstantiations of the more fundamental properties ground, or 'fix', the instantiation of the less fundamental ones". But, she contends, determinate electric charges fix facts "that specific [powers], even all taken

together, do not fix" (57). For instance, an electric charge fixes the fact that if an object has one of the specific powers pertaining to the charge, then it also has all the others. This co-instantiation of specific powers is easily explained if the determinate charges are fundamental: the powers keep together because they "are consequences of one and the same fundamental [power], electric charge" (58). By contrast, while each of the maximally specific powers, along with its activating conditions, fixes the value of the force which would be exerted, it does not fix facts about other powers of its ilk, in particular their co-instantiation. And if so, then determinate properties are more fundamental than maximally specific powers, which Vetter again takes to contradict a ramification of the conditional analysis and thus the analysis itself.

If determinate properties were built up entirely from maximally specific powers, then each of the powers would – by virtue of its character – underpin a single narrow regularity related to the involvement of a property in a law of nature. But the common features of the narrow regularities, and *ipso facto* some more general aspects of the behaviour of the property, would be explained by the powers taken together. Thus, the constancy of the mathematical relation between a determinate electric charge, other electric charges, and their distance (on the one hand) and the exerted force (on the other) would be explained by the causal profiles of the specific powers in determinate charges and their collective presence in the charges. In other words, this "much more striking regularity" would stem from the fact that the powers which would be the building-blocks of a charge would be mathematically related to their manifestations in the same way.

This conclusion would not be affected if the maximally specific powers that would be the building-blocks of a determinate property were cemented by, say, a relation. (The existence of such a unifying component is perhaps presumed by the adherents of the view that determinate properties are conjunctions of single-track powers. The component would account for the remarkable cohesion of the properties, which neither shed their powers nor acquire new ones). While the relation would underwrite the stability of the causal profile of the property over time, the profile itself would still be delineated by the natures of the maximally specific powers and

their collective presence in the property. These two factors would thus indirectly ground the behaviour of the property, including any broad regularities in which it would be involved.

A similar response can be given to Vetter's second argument for the claim that determinate properties are more fundamental than the relevant maximally specific powers. The argument, to reiterate, rests on the proposition that the properties ground (i.e. fix, or guarantee) facts about the co-instantiation of the powers, facts which are not grounded by powers themselves, even "taken together". The proposition is in turn said to be supported by the fact that if an object has one of the specific powers corresponding to a property, then it also has all the others. This fact would obtain, however, if the property were a conjunction of the powers, or if each of the powers were a building block of the property. *Contra* Vetter, the fact would then be fixed by the specific powers taken in conjunction. Again, there would be little difference here if the powers were bound together by a relation underwriting their cohesion and thus the stability of the property. While the relation would then contribute to fixing the fact in question, the role played by the specific powers would be crucial. Indeed, it appears that the property would fix the fact in part *because* it would comprise the specific powers. And this means that the case would not show that the property would be more fundamental than the powers.[11]

---

[11] I should like to thank an anonymous referee for encouraging me to address Vetter's criticism of the conditional analysis. Space precludes discussion of the proposition that if a property is best or adequately characterised by a number of sentences which capture the natures of maximally specific powers, then the property is mereologically complex. One might also enquire whether Vetter, who embraces the proposition and recognizes multi-track powers, is not committed to the view that determinate properties are conjunctions of maximally specific powers.

Conclusion

Laws of nature are the focus of one of the central arguments for the proposition that all properties, or at least the fundamental ones, are powers (Swoyer 1982, Bird 2007). The nub of the argument is that this view of properties "provides an explanation of why there are laws, while avoiding problems" that bedevil competing accounts of lawhood (Bird 2016, 346–47). Versions of this approach differ slightly with respect to what exactly laws of nature are. (The leading version (Swoyer 1982; Bird 2007, 64 and 200–202) has it that a law is a relation between properties, namely a power (*cum* the properties at the core of its activating conditions – its dispositional partners) on the one hand, and the property which defines its manifestation on the other.) What all the versions agree about, though, is that the nature of powers is the source of laws.

Since forces are involved in laws of nature on a par with such properties as mass or charge, the nomic argument for the reality of powers speaks in favour of the thesis that forces are a category of them. It is therefore a ramification of the argument that a solution to the conundrum of force composition should rest on the recognition of forces as powers. In the absence of such a solution the argument would have a loose end. By contrast, if the solution put forward above succeeds, the argument will be strengthened. The view that forces are powers will then be supported both by its role in the solution and by the argument. This does not mean that the view will be home and dry, for it faces significant difficulties (which, for reasons of space, can only be touched on here). Chief among them is a worry which springs from the observation that displays of powers are susceptible to interfering factors, particularly so-called finks and masks.[12] The involvement of forces in, say, the second law of motion would then mean that the law is not exceptionless, which is a position that has come under heavy fire (e.g. Earman *et al*. 2002). I am inclined

[12] One would also have to show that the view can deal with cases of force composition which *prima facie* challenge it. Thus, the view would need to explain away the appearance that when two horses walking in the same direction on the opposite banks of a river pull a barge by means of ropes, the vessel is subject to forces acting along the ropes.

to think that the response to this concern should for the most part focus on a peculiar feature of the manifestation of forces, that is, on the *instantaneity* of the acceleration which they produce.[13] For the time being, though, the jury is out on the question of whether forces are a kind of power, and the argument of this essay speaks in favour of caution.[14]

Jagiellonian University at Krakow

References

Aimar, S. (2019), "Disposition ascriptions", *Philosophical Studies* 176, pp. 1667–1692.

Azzano, L. (2021), "Structural properties, mereology, and modal magic", *Synthese* 198, supplement issue 18, pp. 4303–4329.

Balashov, Y. (1999), "Zero-Value Physical Quantities", *Synthese* 119, pp. 253–286.

Bird, A. (2007), *Nature's Metaphysics: Laws and Properties*, Oxford University Press, Oxford.

Bird, A. (2016), "Overpowering: How the Powers Ontology Has Overreached Itself", *Mind* 125, pp. 341–383.

Cartwright, N. (1999), *The Dappled World*, Cambridge University Press, Cambridge.

Creary, L. G. (1981), "Causal Explanation and the Reality of Natural Component Forces", *Pacific Philosophical Quarterly* 62, pp. 148–157.

Earman, J., J. Roberts and S. Smith (2002), "*Ceteris Paribus* Lost", *Ekenntnis* 57, pp. 281–301.

Hüttemann, A. (2004), *What's Wrong with Microphysicalism?*, Routledge, London.

Johnston, M. (1992), "How to Speak of the Colors", *Philosophical Studies* 68, pp. 221–263.

---

[13] In order to deal with masks (conceived of in the way suggested by Johnston (1992, 233)), one might need to argue that the mere presence of a property cannot prevent the display of a power. If so, then one of the approaches to antagonistic forces sketched above will have to be reconsidered.

[14] I would like to thank Jerzy Gołosz, Mirosław Janusz, Joanna Luc, Tadeusz Szubka, and especially the referees for this journal for helpful comments. Research for this paper was supported by the National Science Centre, Poland (grant 2017/25/B/HS1/00897). I am grateful to the Centre for the support.

Lange, M. (2005), "How Can Instantaneous Velocity Fulfill Its Causal Role?", *The Philosophical Review* 114, pp. 433–468.
Lange, M. (2009), "A Tale of Two Vectors", *Dialectica* 63, pp. 397–431.
Manley D. and R. Wasserman (2008), "On Linking Dispositions and Conditionals", *Mind* 117, pp. 59–84.
Martin, C. B. (1994), "Dispositions and Conditionals", *The Philosophical Quarterly* 44, pp. 1–8.
Massin, O. (2009), "The Metaphysics of Forces", *Dialectica* 63, pp. 555–589.
Massin, O. (2017), "The Composition of Forces", *The British Journal for the Philosophy of Science* 68, pp. 805–842.
Molnar, G. (2003), *Powers: A Study in Metaphysics*, Oxford University Press, Oxford.
Swoyer, C. (1982), "The Nature of Natural Laws", *Australasian Journal of Philosophy* 60, pp. 203–223.
Vetter, B. (2015), *Potentiality: From Dispositions to Modality*, Oxford University Press, Oxford.
Wilson, J. (2009), "The Casual Argument Against Component Forces", *Dialectica* 63, pp. 525–554.

The following back volumes of ***Acta Philosophica Fennica*** are available from Bookstore Tiedekirja, Snellmaninkatu 13, FI-00170 Helsinki, Finland, tel. +358−9−635 177, email: tiedekirja@tsv.fi, www.tiedekirja.fi:

Fasc. LXXXVI (2009): SAMI PIHLSTRÖM AND HENRIK RYDENFELT (eds.): Pragmatist Perspectives. 295 pp.
Fasc. LXXXVII (2010): VIRPI MÄKINEN (ed.): The Nature of Rights: Moral and Political Aspects of Rights in Late Medieval and Early Modern Philosophy. 257 pp.
Fasc. LXXXVIII (2010): LEILA HAAPARANTA (ed.): Rearticulations of Reason. Recent Currents. 274 pp.
Fasc. LXXXIX (2012): ILKKA NIINILUOTO AND SAMI PIHLSTRÖM (eds.): Reappraisals of Eino Kaila's Philosophy. 232 pp.
Fasc. XC (2013): JAAKKO HINTIKKA (ed.): Open Problems of Epistemology – Problèmes ouverts en épistémologie. 207 pp.
Fasc. XCI (2015): GABRIEL SANDU: Logic, Language and Games. 139 pp.
Fasc. XCII (2016): GEORG MEGGLE AND RISTO VILKKO (eds.): Georg Henrik von Wright's Book of Friends. 250 pp.
Fasc. XCIII (2017): ILKKA NIINILUOTO AND THOMAS WALLGREN (eds.): On the Human Condition – Philosophical Essays in Honour of the Centennial Anniversary of Georg Henrik von Wright. 463 pp.
Fasc. XCIV (2018): JAAKKO KUORIKOSKI AND TEEMU TOPPINEN (eds.): Action, Value and Metaphysics. Proceedings of the Philosophical Society of Finland Colloquium 2018. 187 pp.
Fasc. XCV (2019): HENRIK RYDENFELT, HEIKKI J. KOSKINEN AND MATS BERGMAN (eds.): Limits of Pragmatism and Challenges of Theodicy – Essays in Honour of Sami Pihlström. 227 pp.
Fasc. XCVI (2020): ILKKA NIINILUOTO AND SAMI PIHLSTRÖM (eds.): Normativity – The 2019 Entretiens of Institut International de Philosophie. 228 pp.

Reasons and Responsibilities
Proceedings of the Philosophical Society of Finland Colloquium 2020

Edited by Inkeri Koskinen & Teemu Toppinen

This volume of the *Acta Philosophica Fennica* series collects articles based on papers delivered at the Philosophical Society of Finland Colloquium 2020 (FiPhi 2020). The diversity of the topics it covers - from environmental ethics to philosophy of physics, and from epistemic responsibility to theories of truth - reflects the diversity of philosophical research in Finland. The colloquium was held in Helsinki in January 2020.

INKERI KOSKINEN is University Lecturer at Tampere University.

TEEMU TOPPINEN is Associate Professor at Tampere University.

ISBN 978−951−9264−94−3
ISSN 0355−1792

Hansaprint Oy
Helsinki 2021