CROATIAN JOURNAL OF PHILOSOPHY

Vol. XXIII · No. 68 · 2023

Articles

Croatian Journal of Philosophy
Vol. XXIII, No. 68, 2023
https://doi.org/10.52685/cjp.23.68.1
Received: June 21, 2022
Accepted: February 6, 2023

The Problem of Perceptual Agreement

ELAY SHECH and MICHAEL WATKINS
Auburn University, Alabama, USA

We present the problem of perceptual agreement (of determinate color) and submit that it proves to be a serious and long overlooked obstacle for those insisting that colors are not objective features of objects, viz., non-objectivist theories like C. L. Hardin's (2003) eliminativism and Jonathan Cohen's (2009) relationalism.

Keywords: Color; objectivism; perception.

The philosophical literature on color is replete with arguments from perceptual variation. These arguments take various forms and reach different conclusions. Jonathan Cohen (2009), for instance, argues that perceptual variation supports the position that colors are highly relational features of objects; every object has many colors, relative to different observers and different viewing conditions. C. L. Hardin (2003) argues that perceptual variation commits us to eliminativism about colors since each color necessarily has a particular hue, and there is no fact of the matter as to what particular hue any object has.

We will not address these arguments here. Instead, we highlight a feature of color perception and our communication about it: we can agree when two objects are exactly the same determinate color. Whatever might be said about perceptual disagreement, we think that the problem of perceptual agreement that we highlight below proves to be a serious and long overlooked problem for those insisting that colors are not objective features of objects.

It is well known that we generally agree about the more determinable colors of objects. Otherwise, color vocabulary would not have earned its keep. It is also widely recognized that human color vision is fairly constant in how it sees an object's color across a wide range of lighting conditions, or at least that our judgments about an object's color will generally remain consistent even while viewing that object

across a wide range of lighting conditions. Objectivists about color, those holding that an object has its color independent of how it is experienced, often appeal to such agreements.[1] But these are not the agreements on which we focus here. We focus, instead, on determinate shades. This may come as a surprise, since it is well known that different observers under different circumstances will experience the color of an object differently and will even, at times, make different judgments about the colors of particular objects. For example, what any person sees as unique blue (as blue with no red or green in it) will be seen by most as having some red or green it. And we know that two objects that match in color might match only relative to an observer and a lighting condition. Metameric matches, objects that appear the same color for an observer under some lighting condition despite having different reflectance profiles, will sometimes appear very differently colored to some other observer or to the same observer under some different lighting condition.[2]

To appreciate the perceptual agreement that we wish to focus upon, imagine someone tasked with matching the color of some paint. This task is common enough. If we have painted part of a room and find we do not have enough paint, or we are repairing a painting or a car, finding exactly the right color might be very tricky. It will not be enough, for instance, for the new paint to match the old only under sunlight. A match in color will require that any observer (or at least any observer we care about) under any lighting condition (or at least any lighting condition in which the object might be viewed) will not see a difference between the new paint and the old.

Now imagine that the task was successfully completed. The wall painted with the new paint matches the wall painted with the old paint. Enter Susan and John. Susan sees the walls as slightly more purple than blue; John sees the walls as slightly more blue than purple. That's our old friend, perceptual variation, entering the stage for a brief moment. What Susan and John agree on, what they might verify by looking at the walls where they meet across a wide range of lighting conditions, is that the two walls are exactly the same color. Indeed, we might reasonably claim that a necessary condition for two objects having *exactly* the same determinate color is that no one can visually detect a color difference between those objects so long as those objects are viewed side by side and against the same background. Yet, even this might not be sufficient for a perfect match, since A and B might be indiscernible in color for any observer under any condition, and B and C might also be indiscernible in color, even though A and C are discernible in color. And so, by that visual test, we will have shown that A and B are ever so slightly different. But our goal here is not to

[1] See, for example, Keith Allen (2017).

[2] See Hardin (1988) for a scientifically informed discussion of perceptual variation.

give a full account of what it is for two objects to have exactly the same determinate color. Undoubtedly, just as the standards for two objects having the same length might vary depending on purpose, so will the standards for two objects matching in color.

It is important to be clear about what it is that Susan and John agree about. It is not that Susan and John agree about what color the walls are, although they very well might. Rather, what they agree on, what they might well have determined visually across a range of lighting conditions, is that the walls *match* in color.[3] That is what they visually determine, not by seeing the walls at any particular moment, but over a range of lighting conditions. Moreover, when Susan claims that the two walls match in color, she is not merely claiming that they match for her, or for her at the moment. Susan's claim commits her to its being the case that the walls match for everyone (or everyone relevant for the standard she is using) across all lighting conditions (or every relevant condition for the standard she is using).[4] A non-objectivist about color must, of course, explain how we often agree in our judgments about an object's color and why color language seems to ascribe objective properties to objects, and some have taken on that task (e.g., Cohen (2009) and Brogaard (2015)). Their success or failure is not relevant here, however. Our interest is not in how we might explain agreement in judgment. Our interest is in how to explain a particular kind of visual success, our ability to each recognize visually that two objects match in color, i.e., our ability to determine that two objects are indiscernible in color across all lighting conditions.

What it is for two objects to look alike is simply for them to be visually indistinguishable, and so what it is for two objects to look alike in color is for their colors to be visually indistinguishable across observers and lighting conditions. Of course, two objects might look alike in color under some lighting condition and not another. Or they might look to be different colors while against different backgrounds, but the same against the same background. But we assume that the common sense standard for visually determining whether two objects have the same color is by looking at them side by side, against the same background, and under various lighting conditions. Susan and John, employing this common sense standard, agree that the two walls look to have the same

[3] Susan may only care, of course, about human observers (and so not care about ultraviolet shades) or only the lighting conditions that are typically available to homeowners, including sunlight. For Susan, it's likely enough that no difference can be seen; she only needs the walls to match, not perfectly, but perfectly relative to her particular interests.

[4] The predicate "is the same color as" thus seems to work much as "is the same height as." And if, for instance, Susan claims of one wall that it is blue, she commits to treating as blue anything that matches that wall in color across observers and lighting conditions. In this way, at least, "is blue" would seem governed much as "is tall." It seems not to ascribe a relative or "centered" property, as, perhaps, "is tasty" might. An opposing view is suggested by Andy Egan (2007) and endorsed by Brogaard (2015). Also see Cohen (2009).

colors. And we can well imagine that Susan and John are not alone. We can well imagine that no one could see a difference in color between the two walls. The two walls appear to be (at least very nearly) a perfect match in color. Everyone agrees.

Larry Hardin tells us that any objectivist about colors should agree that "it is normally possible to determine what color a thing has by looking at it" (2003: 191). Due to perceptual variation and our inability to select the favored observers and conditions, he argues that objectivism should be abandoned. But we now turn this argument on its head. Every eliminativist about color should agree that, since nothing is colored, no two things can be colored the same. But it is normally possible to determine whether two things are differently colored or the same color by looking at them, at least over a range of lighting conditions, against the same background, and compared side by side. That is what Susan and John did. They determined that the two walls have the same color by looking at them. Susan and John visually determined that the walls are alike in color. For Hardin, this success is illusory. The two objects are not alike in color despite Susan and John seemingly seeing that they are and everyone else agreeing, and despite our having every reason to believe that those objects share physical properties that explain their agreement.

A relationalist like Jonathan Cohen might seem better placed to account for agreement. For Cohen, each object has many colors, but colors are highly relational features of objects. On Cohen's view, the color that you see an object as having in direct sunlight is not the same color that you see the object as having in shadow, and so you see a cup that is half in shade as having two colors. This, to many, is very counterintuitive. The cup, many will insist, appears uniformly colored, but partly in shade. Cohen's reply is that, although you will see two different colors, your judgment that the cup is uniformly colored

> is not a judgment to the effect that the regions are occurrently manifesting a common color, but rather to the effect that the regions share a color that one of them is not occurrently manifesting. That is, the subject judges that, although the unlit region looks different (in respect of color) from the region in shadow, the two regions would look the same (in respect of color) were they both viewed under sunlight. (2009: 56)

So Cohen might claim that when Susan judges that the two walls are colored the same, what she is saying is just that the two walls have all and only the same colors. John agrees. Agreement explained.

But this will not do. For Cohen, Susan is claiming that the two walls share a set of relational properties. John is claiming that the two walls share an *entirely different* set of relational properties. On Cohen's account, when Susan and John each claim that the walls have the same color, they are making radically different claims.

To illustrate how odd this situation is, let's look at a very different kind of case. Cohen tells us little about what it is for a property to be relational. He thinks that we can make do with paradigm examples

like *being a sister* (2009: 8). So imagine two detectives, Jake and Hank. Jake is hired by Evelyn, who is a sister of Laverne. Noah is their father. Hank is hired by Laverne. Jake concludes that Katherine, Evelyn's ward, is Evelyn's sister; and that Patricia, Laverne's ward, is Laverne's sister. Hank concludes that Katherine is Evelyn's daughter, and that Patricia is Laverne's daughter. It turns out that both are correct since the incestuous Noah fathered both Katherine and Patricia. Of course, Jake and Hank agree that Katherine is related to Evelyn just as Patricia is related to Laverne. But their agreement is accidental. Jake and Hank are equally correct and equally in the dark, but about different relations. That's *Chinatown*.

Cohen's account of colors puts Susan and John in positions similar to that of Jake and Hank. Susan and John agree, but not about what they thought they agreed about. But the case for Cohen is odder still, even if not nearly as disturbing. For not only do Susan and John agree that the walls share a color, but everyone does. And, presumably, what everyone agrees about is that the walls share a property in common. But it turns out, if Cohen is correct, no one (or hardly anyone) agrees about what it is that the walls have in common.

Nonetheless, one may object that, for the objectivist, concerns abound: What color is the color that the walls share, especially given that John and Susan won't typically agree on this issue? And what about problems having to do with perceptual variation that the objectivist must face? Such issues are beyond the scope of this short paper.[5] Instead, what is crucial here is that non-objectivists like Cohen and Hardin must contend with the problem of perceptual agreement and it isn't clear that a reasonable solution is forthcoming. The objectivist, on the other hand, has an easy and common-sensical solution: the walls share a common property; namely, they have exactly the same determinate color.[6]

[5] We take up these issues elsewhere. See Watkins and Shech (2022) and Shech and Watkins (unpublished). For a sample of other strategies, see Alex Byrne and David Hilbert (2004), Mark Kalderon (2011), and Allen (2016).

[6] The argument from perceptual agreement alone does not motivate any view of what colors are. Colors might be dispositions (e.g., McGinn (1983)), or properties that supervene on dispositions (e.g., McGinn (1996)), or physical properties (e.g., Byrne and Hilbert (2021)), or properties that supervene on physical properties (e.g., Joshua Gert (2021)). Moreover, for all we have said, an object might have different colors all over at the same time, at least at various determinable levels. What the argument is an argument for is that there must be some feature that objects share when they match in colors. Whatever feature that is might reasonably be a thought of as the determinate color of the object.

References

Allen, K. 2017. *A Naïve Realist Theory of Colour*. Oxford: Oxford University Press.

Brogaard, B. 2015. "The Self-Locating Property Theory of Color." *Minds and Machines* 25 (2): 133–147.

Byrne, A. and Hilbert, D. R. 2004. "Hardin, Tye, and Color Physicalism." *Journal of Philosophy* 101 (1): 37–43.

Byrne, A. and Hilbert, D. R. 2021. "Objectivist Reductionism." In D. Brown and F. Macpherson (eds.). *The Routledge Handbook of Philosophy of Colour*. New York: Routledge, 287–298.

Cohen, J. 2009. *The Red and The Real: An Essay on Color Ontology*. Oxford: Oxford University Press.

Gert, J. 2021: "Primitivist Objectivism." in D. Brown and F. Macpherson (eds.). *The Routledge Handbook of Philosophy of Colour*. New York: Routledge, 299–311.

McGinn, C. 1983. *The Subjective View*. New York: Oxford University Press.

McGinn, C. 1996. "Another Look at Color." *Journal of Philosophy* 93 (11): 537–553.

Egan, A. 2007. "Secondary Qualities and Self-Location." *Philosophy and Phenomenological Research* 72 (1): 97–119.

Hardin, C. L. 1988. *Color for Philosophers*. Indianapolis: Hackett.

Hardin, C. L. 2003. "A Spectral Reflectance Doth Not a Color Make." *Journal of Philosophy* 100 (4): 191–202.

Kalderon, M. E. 2011. "The Multiply Qualitative." *Mind* 120 (478): 239–262.

Shech, E. and Watkins, M. "The Metaphysics of Colors." Unpublished.

Watkins, M. and Shech, E. 2022. "Colors, Perceptual Variation, and Science." *Erkenntnis*: doi.org/10.1007/s10670-022-00574-2

Croatian Journal of Philosophy
Vol. XXIII, No. 68, 2023
https://doi.org/10.52685/cjp.23.68.2
Received: August 14, 2022
Accepted: April 3, 2023

Transitivity and Humeanism about Laws

ANDREJ JANDRIĆ and RADMILA JOVANOVIĆ KOZLOWSKI*
University of Belgrade, Belgrade, Serbia

Humeanism about laws has been famously accused of the explanatory circularity by David Armstrong and Tim Maudlin, since the Humean laws hold in virtue of their instances and, at the same time, scientifically explain those very instances. Barry Loewer argued that the circularity challenge rests on an equivocation: in his view, once the metaphysical explanation is properly distinguished from the scientific explanation, the circularity vanishes. However, Marc Lange restored the circularity by appealing to his transitivity principle, which connects the two types of explanation. Lange's transitivity principle has been widely discussed and criticised in the literature. In view of counterexamples, Lange refined both the principle, by taking into account the contrastive nature of explanation, and the requirement of prohibition on self-explanation. Recently, Michael Hicks has developed a new strategy for defending Humeanism about laws from the refined circularity challenge, critically appealing to the contrastive nature of both explanations and meta-explanations. We will argue that his strategy fails.

Keywords: Humean laws; explanatory circularity; transitivity; contrastive explanations.

* Our research has been supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (451-03-47/2023-01/200163). Andrej Jandrić's research has also been supported by the University of Rijeka, Croatia (uniri-human-18-239). Earlier versions of this paper were presented at the conferences in Dubrovnik (*Metaphysics* 2022) and Rijeka (*Contemporary Philosophical Issues* 2022). We would like to thank Boran Berčić, David de Bruijn, Sam Coleman, Miljana Milojević, David Pitt, Guy Rohrbaugh, Márta Ujvári, Michael Watkins and an anonymous reviewer for their valuable comments.

1. *Humeanism about laws and explanatory circularity*

According to Humeans, scientific laws are generalisations obtaining in virtue of the totality of facts in the global space-time Humean Mosaic[1] and nothing more.[2] In order to distinguish between accidental generalisations and lawful generalisations Humeans typically appeal to Lewis's Best System Account (BSA),[3] or what Psillos (2002: 8) calls "the web-of-laws view"– laws are those generalisations which are entailed by the ideal axiomatic system for our world, i.e. a system containing all the fundamental true propositions about the Mosaic which obtains the best balance between simplicity, informativeness and other desirable properties.[4]

The Humean account of laws has been confronted with many challenges, the crucial one being that the laws conceived in that manner are explanatorily futile. Namely, if laws are nothing but regularities derived from the Humean Mosaic, it is suspicious if such laws are adept to scientifically explain the very features of the Mosaic. It seems that the laws are (at least partly) explained by the Mosaic, parts of which they are expected to explain. The circularity challenge for Humeanism was raised by David Armstrong (1983: 40):

> Suppose, however, that laws are mere regularities. We are then trying to explain the fact that all observed Fs are Gs by appealing to the hypothesis that all Fs are Gs. Could this hypothesis serve as an explanation? It does not seem that it could. That all Fs are Gs is a complex state of affairs which is in part constituted by the fact that all observed Fs are Gs. 'All Fs are Gs' can even be rewritten as 'All observed Fs are Gs and all unobserved Fs are Gs'. As a result, trying to explain why all observed Fs are Gs by postulating that all Fs are Gs is a case of trying to explain something by appealing to a state of affairs part of which is the thing to be explained. But a fact cannot be used to explain itself. And that all unobserved Fs are Gs can hardly explain why all observed Fs are Gs.

[1] In Lewis's version of Humeanism, the Mosaic contains the totality of facts about the point-size distribution of natural properties and natural relations.

[2] The relation between a generalisation and the Mosaic is described by some relation of ontological dependence: originally it was supervenience, but in the more recent literature it is usually grounding.

[3] See Lewis (1983, 1986, 1994), Psillos (2003), Loewer (1996, 2012), Beebee (2000), Schrenk (2006), Cohen and Callender (2009), Bhogal and Perry (2017).

[4] The plea for BSA is far from being philosophically settled. Many concerns have been raised over the years: the question of criteria for the best balanced system, the issue of its uniqueness, the question of problematic mind-dependence of laws, the issue of the choice of language which would allow for comparison between the competing systems, the problem of justifying a preference for one system over the other if they both contain only true propositions, etc. For more, see Armstrong (1983), Carroll (1990), Maudlin (2007) and Roberts (2008), among others. Moreover, BSA is conceived by some as the *objectively* best system which may or may not be formulated yet (and if we are already in possession of the system, we have no way of knowing that, see Loewer (2012:20)): then it is hard to see how it can provide a standard by which to actually discern between laws and merely accidental generalizations. But while we agree that the idea of the best system is dubious in many respects, it will not be the focus of this paper.

And Tim Maudlin (2007: 172) gave a more succinct formulation of the challenge:

> If the laws are nothing but generic features of the Humean Mosaic, then there is a sense in which one cannot appeal to those very laws to explain the particular features of the Mosaic itself: the laws are what they are in virtue of the Mosaic rather than vice versa.[5]

Barry Loewer (2012) tried to meet the challenge by arguing that the alleged circularity results from the equivocation in the use of the term "explanation". Bottom-up explanations are metaphysical explanations: laws are thus *metaphysically explained* by their instances in the Mosaic, which does not preclude them from *scientifically explaining* their instances. The difference between metaphysical and scientific explanations Loewer (2012: 131) described as follows:

> Metaphysical explanation need not involve laws and the explanandum and explanans must be co-temporal (if the explanans is a temporal fact or property). Scientific explanation of a particular event or fact need not show that it is grounded in a more fundamental event or fact but rather, typically, shows why the event occurred in terms of prior events and laws.

Loewer's response provoked a very fruitful debate which continues until today.[6] He did not offer much in the way of a further clarification about metaphysical or scientific explanations, but the currently popular view is to link metaphysical explanations with grounding: laws, which typically have the structure of universal generalisations, are grounded in the total conjunction of their instances.[7] Loewer's proposal has been criticised from different perspectives, but one of the most interesting objections was raised by Marc Lange (2013).

2. *Transitivity*

Lange pointed out that even though the metaphysical and the scientific explanation are two different kinds of explanation, they are not completely unrelated: what connects them, in his opinion, is the *principle of transitivity*:

(T) If E scientifically explains [or helps to scientifically explain] F and D grounds [or helps to ground] E, then D scientifically explains [or helps to scientifically explain] F. (Lange 2013: 256)

[5] For similar arguments, see Bird (2007: 86) and Lange (2013: 256). Earlier, Dretske also contested the view that mere generalisations could have any explanatory power over their instances: "Subsuming an instance under a universal generalization has exactly as much explanatory power as deriving Q from P & Q. None" (1977: 26).

[6] See Lange (2013), Hicks and van Elswyk (2015), Marshall (2015), Miller (2015), Roski (2018), Shumener (2017), Marshall (2015), Dorst (2018), Emery (2019), Bhogal (2020), Hicks (2020), Kovacs (2020) and Duguid (2021).

[7] However, not everybody endorses that view—according to Emery (2019), it is the other way round: laws ground their instances.

In order to argue against Loewer's solution of the circularity challenge, Lange (2013: 258) also made explicit another important and highly plausible principle—that of the prohibition on self-explanation:

(PSE) A fact q cannot explain [or help to explain] itself.

Lange motivated (T) by evoking the actual scientific practice and offering several plausible examples in its favour, such as:

> suppose that a given balloon expands because of various laws and the fact that the pressure of the gas inside the balloon is greater than the atmospheric pressure outside of the balloon. Then since the fact that the internal pressure is greater than the external pressure is grounded in the value of the internal pressure and the value of the external pressure, it follows from the transitivity principle that the internal and external pressures help to scientifically explain why the balloon expands. That is also correct. The internal pressure, in turn, is grounded in the forces exerted by various gas molecules as they collide with the balloon's interior walls. By the transitivity principle, then, those forces help to scientifically explain why the balloon expands. (Lange 2013: 257)

The principle of transitivity immediately restores the circularity of explanation: if the law L is partly grounded in its instance I, and L partly scientifically explains I, then, according to (T), I partly scientifically explains itself, which violates (PSE). Lange (2013: 257) illustrates such circularity with the following example:

> [A] coin's chance of landing heads explains its actual relative frequency of landing heads, so if the chance were grounded in the actual relative frequency, then […] the actual relative frequency would have to explain itself, which it cannot do.

The general validity of the transitivity principle was immediately questioned. Elizabeth Miller (2015) and Michael Townsen Hicks and Peter van Elswyk (2015) have offered a number of counterexamples to it. All these counterexamples roughly follow the same pattern: an instantiation of a higher-level multiply realizable property P (typically a biological or a psychological one) is considered as an explanation of some observed phenomenon F; the instantiation of P is grounded in one of P's micro-structural realizers, M; it is then argued that M does not explain F, since F might have occurred even if M had been missing—if P, for instance, were realized by a different realizer. Here is an example of Hicks and van Elswyk (2015: 438):

> The position of electron e partially metaphysically explains the position of lion L. The position of L scientifically explains the number of prey animals in region R. But the position of electron e does not explain the number of prey animals in region R. For if the electron were elsewhere, L would still be warding prey animals out of R.

It should be noted that this way of defending Humeanism has rather dubious effects: all that can be achieved with the counterexamples, like the one cited, is to show that (T) is not a universally valid principle. However, Lange's principle of transitivity *need not* hold universally in

order to raise a challenge for Humeanism. The Humean account of laws is envisaged as having the most general scope, i.e. as being a metaphysical account of *all* laws: it is the thesis that all laws are grounded in the Mosaic. If (T) were true only of *some* laws and their instances, the Humean account would still render the explanation of *those* instances circular, which is enough of a problem already. While it is perfectly adequate to criticise Humeanism about laws by producing counterexamples to it, it does not seem to be nearly as effective as a strategy against Lange's criticism of Humeanism.

Nevertheless, Lange (2018) himself answered these counterexamples by refining his transitivity principle and by bringing into play the contrastive nature of explanations.[8] In order to restore the circularity challenge for Humeanism, Lange appealed to the fact that scientific explanations typically contain hidden contrasts.[9] According to this view, an explanation does not simply connect an explanandum with its explanans: what it combines instead is a specific difference-maker in the explanandum with the appropriate difference-maker in the explanans. Instead of regarding an explanation as a two-term relation, as we are accustomed, we would do more justice to its nature if we considered it, so to say, as holding between four relata: that *A explains B* is thus to be regarded as an abbreviated form for the claim that *A rather than A' explains why it is the case that B rather than B'*. Contrasts are mostly left implicit, as they are determined by the context of an explanation. By stating contrasts explicitly, Lange (2018: 1341–1342) formulated the refined transitivity principle:

(RT) If the fact that E rather than E' scientifically explains [or helps to scientifically explain] the fact that F rather than F', and if the fact that D rather than D' grounds [or helps to ground] the fact that E rather than E', then the fact that D rather than D' scientifically explains [or helps to scientifically explain] the fact that F rather than F'.

When the relevant contrasts are disclosed in the abovementioned example with a lion, we can easily see that it presents no counterexample to the refined transitivity principle (RT): although the explanandum in the metaphysical explanation (in the first premise) seems *prima facie* identical with the explanans of the scientific explanation (in the second premise)—i.e. the position of lion L—the implausible conclusion about the number of prey animals in region R being explained by the position of electron e does not follow by (RT) from the premises since the contrast implicit in the explanandum of the first premise does not match with the contrast implicit in the explanans of the second premise. According to

[8] The idea of contrastive nature of explanations is defended by van Fraassen (1980), Hitchcock (1996), Barnes (1994), Schaffer (2005) and Hicks (2021), among others.

[9] The idea of using contrastive explanations as a strategy for non-Humeans was suggested by Hicks and van Elswyk (2015)—Lange (2018) accepted the challenge.

Lange (2018: 1342–1344), what the presence of the picked out electron e rather than its absence explains is the occurrence of a particular "leonine configuration"—L—*rather than the occurrence of some other leonine configuration*—L minus e—in region R, while, in the second premise, it is the presence of a leonine configuration L in R, *rather than the absence of any leonine configuration* in R, that explains the number of prey animals there. Hence, the true explanandum in the metaphysical explanation, when the contrasts are taken into account, is a different difference-maker than the explanans of the scientific explanation, and, consequently, the explaining is not transferred by transitivity from the first premise to the second, and the untenable conclusion cannot be derived. By appealing to the contrastive nature of explanations, transitivity can be saved from other counterexamples in an utterly analogous fashion.

Dan Marshall (2015), on the other hand, tried to defend Humeanism about laws and to break the explanatory circle by denying that laws, considered as generalisations, are grounded in their instances. In his view, a law L does indeed (partly) scientifically explain its instance I, but what I (partly) grounds is not L itself, but the higher-level fact about L: the fact *that the generalisation L is a law*. Instances thus do not metaphysically explain laws, but rather the *lawhood* of laws.[10]

Lange (2018: 1351) answered Marshall by refining the prohibition on self-explanation:

(RPSE) The prohibition on self-explanation should be interpreted not only as prohibiting a fact q from helping to explain itself, but also as prohibiting q from helping to explain why (if q obtains) some other fact helps to explain q. Both of these are too circular to qualify as explanations.

According to Lange, Marshall's strategy for upholding Humeanism only seemingly avoids the circularity objection: if an instance I of a lawful generalisation L (partly) explains the fact that L is a law, which in turn (partly) explains why L (partly) explains I, then, by the principle of transitivity, I (partly) explains why L (partly) explains I, and this again violates (RPSE).

3. *Hicks's new proposal*

Recently, Hicks (2021) has proposed a new argument in defence of Humeanism about laws. He has argued that even if we granted to Lange the refined version of the principle of transitivity (RT) and the refined prohibition on self-explanation (RPSE), it would still not follow that the Humean account of laws leads to the explanatory circularity.

[10] Stefan Roski (2017) raised doubts as to whether this proposal for solving the circularity challenge was well motivated. He argues that any motivation we might have for claiming that the instances of a generalisation ground the meta-level fact that the generalisation is a law will *eo ipso* motivate the claim that they ground the generalisation itself.

Unlike Marshall, who claimed that instances did not ground laws that they are instances of, Hicks attempts to break the explanatory circle by denying its other part—i.e. he claims that laws do not scientifically explain their instances, but are instead meta-explanations of the first-order (typically causal) explanations.

In Hicks's view, if the fact that Fa is a cause of another fact Ga, then what explains the occurrence of Ga is not Fa together with the law that all Fs are G, but just the fact that Fa (2021: 535). Contrary to the well-known deductive-nomological model of explanation of Hempel and Oppenheim (1948), specific events need not be subsumed under a law (i.e. nomological generalisation) in order to be fully explained. The law explains further, on the meta-level, the explanatory connection between the first-level explanandum and explanans. Hence, the law does not explain its own instances, and the circularity is circumvented. Hicks here approvingly cites Skow (2016: 75),[11] who claims that

> the fact that the rock was dropped from one meter is offered as a reason why it hit the ground at 4.4 m/s, while the law that $s=\sqrt{2dg}$ is offered as a second level reason why, a reason why the drop height is a reason why the impact speed is 4.4 m/s. The law shows up in the answer to the second-level why question, not in the answer to the first level one.

As Hicks puts it, "laws are not themselves reasons why some event occurs, but instead are second-level reasons why the event's causes produce it" (2021: 540). If an event e is caused by another event c, then c explains e, and the law that c causes e (meta-) explains why c (first-order) explains e. One might object that c, by itself, is not enough for deriving e: it seems that it can do so only together with a law. According to Hicks (2021: 539), the law that c causes e does indeed feature in deriving e from c: however, not as a supressed premise at the same level with c, as it is assumed in the deductive-nomological model, but rather as an inference rule which justifies the transition from c to e. The last claim is labelled by Hicks as the inference rule requirement (IRR): the role of the law in an explanation is to enable deriving the explanandum from the explanans; the law itself is not part of the explanans and, hence, cannot be properly said to explain the explanandum; what it explains is the (second-level) fact that the explanans explains the explanandum.

This manoeuvre is sufficient to bypass the circularity issue as formulated with the original requirement of prohibition on self-explanation (PSE): although instances of a law (partly) ground the law, and thus explain it, the law, in turn, does not explain its own instances, but instead it explains (usually causal) connections between its instances and other events. However, it seems to fail (RPSE): instances help explain the law they are instances of, which again helps explain why some other facts explain the very instances in question.

In order to bypass this circularity, Hicks (2021) appealed to the contrastive nature of both explanations and meta-explanations. He claims

[11] Similar ideas can be found in Schnieder (2010), Ruben (1990) and Scriven (1962).

that the accusations of circularity can be supported only by what he labelled as the revised circularity argument (RCA), and then goes on to contest its soundness. The argument is reconstructed in the following way (Hicks 2021: 547):

(P1) An explanation is problematically circular if it uses e to help explain why (if e obtains) a given c can serve as part of the explanans in an explanation of e.

(P2) If the Inference Rule Requirement is true, then the laws explain why (if e obtains) a given c can serve as part of the explanans in an explanation of e.

(P3) If the laws are Humean, then e helps explain why the laws are what they are.

(IC) If the laws are Humean, and the Inference Rule Requirement is true, then e helps explain why (if e obtains) a given c can serve as part of the explanans in an explanation of e (from P2 and P3 via the transitivity of explanation).

(C) If the Inference Rule Requirement holds, and the laws are Humean, the explanation of e is problematically circular (from P1 and IC).

The premise (P1) in (RCA) is Hick's reformulation of Lange's refined prohibition on self-explanation (RPSE). Premises (P2) and (P3) are implications, with explanations in their consequents: while the explanation in (P3) is a first-level explanation, (P2) contains a meta-explanation as its consequent. The derivation of the claim of the explanatory circularity in the conclusion (C) of (RCA) proceeds in the following way: (P2) and (P3) imply, by the principle of transitivity, the intermediary conclusion (IC), which, together with (P1), gives (C). If we are correctly interpreting Hicks, he wants to claim that (RCA) is not a sound argument: in his view, (RCA) is either invalid or at least one of its premises is false. To defend his case, Hicks appealed to the contrastive nature of both explanations and meta-explanations: hence, the consequents of both (P2) and (P3) contain implicit contrasts. The principle of transitivity, by which (IC) should be derived from these two premises, can only be, accordingly, the refined principle of transitivity (RT) which takes contrasts into account. Now, Hicks maintains that, when the hidden contrasts in (P2) and (P3) are properly spelled out, it is either the case that the difference-maker in the explanandum in the consequent of (P3) does not coincide with the difference-maker in the explanans in the consequent of (P2)—invalidating thus the application of (RT) to those premises in deriving (IC)—or the premise (P3) is false. He believes that there is no way to specify the unstated contrasts in (P2) and (P3) so as to make them both true and connectable by the principle of (refined) transitivity. We will argue that he is wrong and that (RCA) is not only valid, but also sound. Hereinafter, we will proceed in the following manner: we will first outline Hicks's interpretation of (RCA) and the way he determines the hidden contrasts in premises (P2) and

(P3). Then we will argue that his proposed contrasts are neither the only feasible nor the most plausible ones. And finally, we will provide reasons for another reading of (RCA), which restores the argument's soundness and the circularity challenge for the Humean account of laws. However, before we turn to determining the relevant contrasts in (RCA), we would also like to point out two more general worries with Hicks's new defence of Humeanism about laws.

First, Hicks identifies first-level explanations of events with their causes. This is evident in (RCA) in the premise (P2), in which the laws are, according to the inference rule requirement, regarded as meta-explanations of the fact that the occurrence of an event e is, at the first level, explained by its cause c. In our view, conflating causes and explanations *prima facie* looks like mixing categories. Causes are usually events which cause other events; they bring them into existence, but do not explain them. Explanations, on the other hand, explain already existing events, but do not produce them. It seems that it is precisely the law that does the explaining; and if it is a causal law, in doing the explaining it will refer to the (kind of the) event's cause.[12] However, we are aware that to raise this concern means exactly to overturn some of the assumptions upon which Hicks rests his case for Humeanism.

The second concern is related to the status of laws in Hicks's account. A true generalisation is a law, according to BSA, only if it is derivable as a theorem in the best system (or, if it is not unique, in all the best systems). Hence, it is a theorem—a *proposition*. On the other hand, in order to respond to the circularity challenge, Hicks claimed that the laws were *inference rules*.[13] Thus, they would have to be both propositions and inference rules. But nothing can be both in a single context: propositions are truth-apt, while inference rules are not. Maybe the contexts in which the laws have the role of inference rules could be separated from those in which they function as propositions, but so far no such demarcation has been proposed by Hicks.

And now we turn to our main argument against Hicks. We claim that he does not succeed in avoiding circularity by appealing to the contrastive nature of explanations and meta-explanations in (RCA). Let us consider in more detail why he thinks that the contrasts contained in

[12] An anonymous reviewer suggested that a charitable reading of Hicks demands that we make room for a distinctive kind of causal explanation in which an event can both cause and explain some other event. We believe, however, that a cause would be able to explain its effect only if they are described in a certain way, and that a proper description would eventually include a lawful connection between these events. We cannot delve into details here, but we wish to emphasise that allowing for causes alone to explain their effects does not affect our main argument against Hicks, which is given below.

[13] Hicks seems here to subscribe to the best system account; see (Hicks 2021: 549). In an earlier article (Hicks 2018) he criticised BSA and suggested that it should be replaced by his Epistemic Role Account (ERA). However, our objection applies to ERA as well: in ERA, laws are theorems of the system which best balances strength and breadth, and hence propositions.

premises (P2) and (P3) either disable the application of transitivity to these premises or falsify the premise (P3). Hicks (2021: 548–549) himself gives an analysis of contrasts implicit in the meta-explanation in the consequent of (P2). (P2) says that the laws, which serve as inference rules according to the (IRR) contained in its antecedent, (at the second level) explain why c (at the first level) explains e. To be more precise, the fact that *if c then e* is an instance of a law, and not an accidental truth, enables the derivation of e from c. If the connection between the occurrence of c and the occurrence of e were merely accidental, c would not be able to explain e. It is now clear how the difference in the explanandum is related to the difference in the explanans in the consequent of (P2); hence, what (P2) claims, with contrasts spelled out, is the following:

(P2') If the Inference Rule Requirement is true, then the fact that *if c then e* is an instance of a law (rather than a mere accident) explains why c explains e (rather than not explaining e).

Hicks points out that both contrasts in (P2') presuppose that e occurs (and, for that matter, that c occurs as well). In the explanandum, it is presupposed that c and e are facts: the difference expressed by the contrast is that *between* there being an explanatory relation between those facts *and* there not being such a relation. The same holds for the explanans in (P2'): the relevant difference is that between the connection between c and e being lawful and it being accidental—but there would not have been any connection between c and e in the first place had they not both occurred. Hence, Hicks concludes that the difference between the occurrence and the non-occurrence of e is not relevant for the contrast in the explanans of (P2').

However, in order to deduce (IC) from (P2') and (P3) by the application of (RT), the difference-maker in the explanandum of (P3) has to coincide with the difference-maker in the explanans of (P2'), i.e. the appropriate contrasts have to match. Since, according to the considerations above, the difference between the occurrence and the non-occurrence of e does not affect the difference-maker in the explanans of (P2'), whether e occurs or not cannot be relevant for explaining the contrast in the explanandum of (P3) either. But Hicks seems to believe that the only possible ascription of contrast in the explanans of (P3) is exactly that between e's occurrence and its non-occurrence, i.e. he thinks that (P3), when the contrasts have been spelled out, expresses the following claim:

(P3') If the laws are Humean, then the occurrence of e (rather than its non-occurrence) helps explain why *if c then e* is an instance of a law (rather than a mere accident).

What the consequent of (P3') *says* is that the difference between e occurring and it not occurring *is* relevant for the difference-maker in the explanandum of (P3'): hence, (P3') is either false, or, if it is true, the difference-maker in the explanandum of (P3') cannot coincide with the

difference-maker in the explanans of (P2')—as the difference-maker in the explanans of (P2') is not affected by the difference between the occurrence and the non-occurrence of e. In the latter case, we would have an equivocation: behind their common expression, the explanans in (P2') would really not be the same as the explanandum in (P3'), which would be sufficient to block the application of the refined transitivity (RT) and to invalidate the argument (RCA).

The problem with Hicks's reasoning is that (P3'), as we announced earlier in the paper, is neither the only possible nor the most plausible interpretation of (P3). We believe that the contrasts which Hicks has set in the consequent of (P3) are not fitting, and we want to argue that, when the hidden contrasts in (P3) are properly determined, (P3) becomes both true and connectable by the principle of refined transitivity (RT) with (P2').[14] More precisely, a proper interpretation of (P3), in our view, will show that the mere occurrence of the fact e is equally irrelevant in the premise (P3) as it is in the premise (P2), thus making the two connectable by (RT). First we will analyse (P3) and offer another, more appropriate interpretation (P3"), in which the contrasts are determined by (P3)'s antecedent and which makes (P3) trivially true. Then we will argue that the interpretation (P3'), proposed by Hicks, amounts to a thesis unacceptable to Humeans.

The unspecified contrasts in the explanation contained in the consequent of (P3) are determined by that explanation's context, which, in turn, is dictated by (P3)'s antecedent. Unfortunately, in deciding which contrasts are left implicit in (P3)'s consequent, Hicks at no place appeals to its antecedent (which contains the claim *that the Humean account of laws is a true one*)—which is a different way of proceeding than in the treatment of (P2). When he spelled out the contrast in the meta-explanation contained in the consequent of (P2), Hicks paid due attention to the fact that the antecedent of (P2) is the inference rule requirement (IRR). According to (IRR), it is only the laws and not accidentally true generalisations that enable deriving a first-level explanandum e from its first-level explanans c. It is exactly the antecedent of (P2) that helped determine the supressed contrasts in its consequent. Now, following the same method, let us take a closer look at what is claimed in (P3)'s antecedent in order to arrive at its hidden contrasts.

The antecedent of (P3) is the thesis of Humeanism about laws. It claims that laws are grounded in the Mosaic: whether a certain generalisation is a law depends on what the Mosaic contains. According to

[14] It should be noted that Hicks (2021: 549–550) himself anticipated that some readers might be dissatisfied with his suggested contrasts in (P2') and could devise different contrasts instead: he considered several such competing proposals and found them all wanting and unable to support (RCA). However, none of the criticisms of his reading of (RCA) and rival proposals which he envisaged corresponds to what we wish to claim: in our view, Hicks's interpretation of (P2), as stating (P2'), is quite adequate—what we contest is his reading of (P3) and the contrasts he expressed in (P3').

the most influential version of Humeanism—that of David Lewis—the Mosaic determines the best system for our world, which in turn determines what laws are. In Lewis's view, a true regularity is a law if "it fits into some integrated system of truths that combines simplicity with strength in the best way possible" (1986: 122). Such integrated system is to be understood as a deductive systematisation, its strength being determined by the set of its consequences and its simplicity by the number and mutual similarity of its axioms.[15] The universal generalisations which appear as axioms in the best system are fundamental laws, while the universal generalisations which are deduced as theorems are derived laws. Predicates in the fundamental laws should refer to perfectly natural properties only, while predicates which appear in the derived laws can also designate properties which are natural to a sufficiently high degree (Lewis 1983: 368). Since the best system is the result of a trade-off between considerations of strength and considerations of simplicity, which pull in different directions, Lewis allowed that some of the system's strength could be sacrificed for an appropriate increase in the system's simplicity: the result is that the best system for our world need not be complete.[16] Consequently, if e is some particular fact, not only need it not be in the Mosaic, but it need not be derivable from it either. And although in most of the contemporary literature on Humeanism about laws it is tacitly assumed that the best system is deductively complete, this, as Kovacs (2021) points out, is neither the case in Lewis's original version of the best system account nor is it universally accepted within the Humean camp: thus, for example, in Braddon-Mitchell's (2001) version of Humeanism the best system is incomplete.

What Humeanism about laws therefore prohibits is that laws be determined by facts not in the Mosaic. Hence, the assumption that the laws are Humean, in the antecedent of (P3), naturally induces in its explanans the contrast between facts which are in the Mosaic and those which are not. Consequently, (P3) should be understood, *pace* Hicks, as stating the following claim:

(P3") If the laws are Humean, then e (rather than some fact not in the Mosaic) helps explain why *if c then e* is an instance of a law (rather than a mere accident).[17]

Now, (P3") is obviously true. If (P3) is read, as Hicks reads it, as abbreviating (P3') instead of (P3"), its truth will immediately become

[15] We are here roughly following the outline of Lewis's best system account as given in Kovacs (2021).

[16] Hicks (2018) seems to believe that strength always trumps simplicity. In his view, the best system is achieved by a trade-off between strength and breadth. However, such a system can also be incomplete, which is all that is required for our argument against his reconstruction of (RCA).

[17] Of course, if we make room for non-fundamental explanations, e need not be part of the Mosaic, but it still has to be grounded in the Mosaic. The relevant contrast in that case would be the one between e and some fact not grounded in the Mosaic.

suspect: in that case, (P3) would claim that a difference between e's occurring and its failing to occur would be responsible for a difference between laws being what they are and them being different. And if all that is assumed about e were that it occurred (i.e. that e were a fact), accepting (P3') would mean adopting the view according to which every change in facts produces a change in laws, which is highly contestable, to say the least. What Humeanism about laws, in its original formulation with supervenience, was designed to exclude is the claim that there are two possible worlds indistinguishable from one another with regard to facts they contain, but different with regard to laws which hold in them, that is, Humeans originally claimed that every change in laws implied a change in facts, and not that every change in facts implied a change in laws, which is what (P3') amounts to. While the former claim means that laws supervene on facts, the latter claim, contained in (P3'), is tantamount to saying that facts supervene on laws. Thus, if (P3') were accepted, together with the Humean thesis that laws supervene on facts, the supervenience which holds between facts and laws would become symmetric. And this should strike any advocate of the Humean account of laws as unacceptable: when the Humeans claim that the laws supervene on facts, what they have in mind is that *asymmetric* supervenience holds between them. Tolerating symmetric supervenience (or some other symmetric relation of ontological dependence) is hardly any better than admitting the initial charge of circularity, indeed it amounts to a form of circularity, only not of scientific but of *metaphysical* explanation: if there can be no difference at the subvenient level without a difference at the supervenient level, then such symmetric supervenience is bound to produce widespread cases of circular explanation.[18]

Now, conceding that the laws are not supposed to yield to every change in facts but are typically considered as being more resilient, nevertheless it may still be objected that whether an event e occurred or did not occur does affect the lawhood status of *if c then e*. Let us suppose that c occurred. If e failed to occur, then *if c then e* would not be true and, hence, would not be a law.[19] As much as this reasoning seems incontestable,[20] it is of no avail to Hicks. The occurrence of e rather than its non-occurrence does help explain why *if c then e* is a true rather than a false generalisation, but the latter contrast does not match the contrast in (P3')'s explanandum, for that contrast is between *if c then e* being a lawful generalisation and it being a merely accidentally *true* generalisation. In the explanandum in (P3'), it is already presupposed that *if c then e* is true; what needs explaining is why it is

[18] Kovacs (2021) makes similar points.

[19] We are grateful to an anonymous reviewer for raising this issue.

[20] Braddon-Mitchell (2001) in fact contested it: he believes that laws need not be true and allows for what he calls "lossy laws". We cannot consider his view in more details here.

moreover an instance of a law rather than a mere accident,[21] and for that purpose the difference between e occurring and it failing to occur is not relevant.

To sum up, (P3') is surely not the only possible reading of (P3), as Hicks seems to believe, since there is an alternative reading on the table, namely (P3"). Moreover, if our considerations above are correct, (P3') is not an admissible reading either: the choice of contrast in its explanans is neither motivated by the contrast in its explanandum nor by its antecedent; and what it claims seems highly implausible, especially to the Humeans. Contrary to (P3'), our suggested reading (P3") not only makes (P3) more plausible but also trivially true. However, since the explanandum in (P3") is the same difference-maker as the explanans in (P2'), the refined principle of transitivity (RT) can be applied to them. Together they give (IC), which with (P1) enables deriving the circularity challenge in the conclusion (C). (RCA) is thus both valid and sound. Somewhat imitating Lange's response to the counterexamples to the principle of transitivity (T), Hicks tried to demonstrate that the argument for the explanatory circularity of the Humean account of laws cannot be sound. We believe to have shown that his attempt failed and that appealing to the contrastive nature of both explanations and meta-explanations is not enough to save Humeanism about laws from the charge of circularity.

References

Armstrong, D. 1983. *What Is a Law of Nature?* Cambridge: Cambridge University Press.

Barnes, E. 1994. "Why P Rather than Q? The Curiosities of Fact and Foil." *Philosophical Studies* 73 (1): 35–53.

Beebee, H. 2000. "The Non-Governing Conception of Laws of Nature." *Philosophy and Phenomenological Research* 61 (3): 571–594.

Bhogal, H. and Perry, Z. R. 2017. "What the Humean Should Say About Entanglement." *Noûs* 51 (1): 74–94.

Bhogal, H. 2020. "Humeanism About Laws of Nature." *Philosophy Compass* 15 (8): 1–10.

Bird, A. 2007. *Nature's Metaphysics: Laws and Properties*. Oxford: Oxford University Press.

Braddon-Mitchell, D. 2001. "Lossy Laws." *Noûs* 35 (2): 260–277.

Carroll, J. 1990. "The Humean Tradition." *The Philosophical Review* 99 (2): 185–219.

Cohen, J. and Callender, C. 2009. "A Better Best System Account of Lawhood." *Philosophical Studies* 145 (1): 1–34.

[21] This point is readily acknowledged by Hicks himself. He writes: "Thus the question we're concerned about is not whether *if c then e* had not been a law, would it have been true. Rather, we are wondering whether had it been accidental, it would have been true. This is the question guided by the contrast in the explanandum. And the obvious answer is that yes, it would have been accidentally true" (Hicks 2021: 549).

Dorst, C. 2019. "Toward a Best Predictive System Account of Laws of Nature." *British Journal for the Philosophy of Science* 70 (3): 877–900.

Dretske, F. 1977. "Laws of Nature." *Philosophy of Science* 44 (2): 248–268.

Duguid, C. 2021. "Lawful Humean Explanations Are Not Circular." *Synthese* 199 (3–4): 6039–6059.

Emery, N. 2019. "Laws and Their Instances." *Philosophical Studies* 176 (6): 1535–1561.

Hempel, C. and Oppenheim, P. 1948. "Studies in the Logic of Explanation." *Philosophy of Science* 15 (2): 135–175.

Hicks, M. T. 2018. "Dynamic Humeanism." *British Journal for the Philosophy of Science* 69 (4): 983–1007.

Hicks, M. T. 2021. "Breaking the Explanatory Circle." *Philosophical Studies* 178 (2): 533–557.

Hicks, M. T. and van Elswyk, P. 2015. "Humean Laws and Circular Explanation." *Philosophical Studies* 172 (2): 433–443.

Hitchcock, C. R. 1996. "The Role of Contrast in Causal and Explanatory Claims." *Synthese* 107 (3): 395–419.

Kovacs, D. M. 2020. "The Oldest Solution to the Circularity Problem for Humeanism About the Laws of Nature." *Synthese* 198 (9): 1–21.

Lange, M. 2013. "Grounding, Scientific Explanation, and Humean Laws." *Philosophical Studies* 164 (1): 255–261.

Lange, M. 2018. "Transitivity, Self-Explanation, and the Explanatory Circularity Argument against Humean Accounts of Natural Law." *Synthese* 195 (3): 1337–1353.

Lewis, D. 1983. "New Work for a Theory of Universals." *Australasian Journal of Philosophy* 61 (4): 343–377.

Lewis, D. 1986. "A Subjectivist's Guide to Objective Chance." In *Philosophical Papers Vol. 2*. Oxford: Oxford University Press, 83–132.

Lewis, D. 1994. "Humean Supervenience Debugged." *Mind* 103 (412): 473–490.

Loewer, B. 2012. "Two Accounts of Laws and Time." *Philosophical Studies* 160 (1): 115–137.

Maudlin, T. 2007. *The Metaphysics Within Physics*. Oxford: Oxford University Press.

Marshall, D. 2015. "Humean Laws and Explanation." *Philosophical Studies* 172 (12): 3145–3165.

Miller, E. 2015. "Humean Scientific Explanation." *Philosophical Studies* 172 (5): 1311–1332.

Psillos, S. 2002. *Causation and Explanation*. Stocksfield: Acumen.

Roberts, J. T. 2008. *The Law-Governed Universe*. Oxford: Oxford University Press.

Roski, S. 2018. "Grounding and the Explanatory Role of Generalizations." *Philosophical Studies* 175 (8): 1985–2003.

Ruben, D-H. 1990. "Explanation in the Social Sciences: Singular Explanation and the Social Sciences." *Royal Institute of Philosophy Supplement* 27: 95–117.

Schaffer, J. 2005. "Contrastive Causation." *Philosophical Review* 114 (3): 297–328.

Schrenk, M. 2006. "A Theory for Special Science Laws." In H. Bohse and S. Walter (eds.). *Selected Papers Contributed to the Sections of GAP*.6. Paterbom: Mentis.

Scriven, M. 1962. "Explanations, Predictions, and Laws." In H. Feigl and G. Maxwell (eds.). *Minnesota Studies in the Philosophy of Science III*. Minneapolis: University of Minnesota Press, 170–230.

Schnieder, B. 2010. "A Puzzle About 'because'." *Logique et Analyse* 53 (211): 317–343.

Shumener, E. 2019. "Laws of Nature, Explanation, and Semantic Circularity." *British Journal for the Philosophy of Science* 70 (3): 787–815.

Skow, B. 2016. *Reasons Why*. Oxford: Oxford University Press.

van Fraassen, B. 1980. *The Scientific Image*. Oxford: Oxford University Press.

Croatian Journal of Philosophy
Vol. XXIII, No. 68, 2023
https://doi.org/10.52685/cjp.23.68.3
Received: December 15, 2022
Accepted: March 30, 2023

Bare Projectibilism and Natural Kinds: A Defense

IÑIGO VALERO
Independent researcher, Donostia, Spain

Projectibility has traditionally been given a prominent role in natural kind theories. However, where most of these theories take projectibility to be a necessary but insufficient feature of natural kinds, this paper defends an account of natural kinds according to which the naturalness of kinds is to be identified with their degree of projectibility only. This view follows thus the path opened by Häggqvist (2005), although it goes significantly further on two main respects. First, I develop and discuss two important dimensions of projectibility that are overlooked in Häggqvist's work. Second, I address two recent important objections (Magnus 2012 and Spencer 2015) against projectibility-based accounts.

Keywords: Natural kinds; projectibility; bare projectibilism; inductive power.

1. *Introduction*

The goal of this paper is to engage in the natural kind debate, and to put forward a projectibility-based account of natural kinds according to which the naturalness of kinds is to be identified with their degree of projectibility.

This view is congenial to a tradition of natural kind theories that has ascribed a central role to projectibility in the characterization of natural kinds. The current proposal, however, departs from other views in singling out no condition for naturalness other than projectibility itself. As such, where other theories have often taken projectibility to be necessary yet insufficient for naturalness, I propose, instead, to identify naturalness with projectibility alone.

This move does not constitute a complete novelty as it follows a path opened by Sören Häggqvist (2005) who, in his proposal "Radical Projectibilism", already argued in favor of this move. The current proposal, however, updates significantly Häggqvist's theory, I contend, by addressing important objections, as well as by emphasizing two important dimensions of projectibility that are not considered by Häggqvist: *graduality* and *abundance*.

Identifying naturalness with a gradual property such as projectibility, I argue, constitutes a significant departure from a tradition of natural kinds that has focused on drawing a demarcatory line between natural and non-natural kinds. Far from being a shortcoming of a projectibility-based account, I will show that understanding naturalness in a gradual way is the most appropriate way to counter the relevant notion of arbitrariness and, moreover, brings significant advantages over dichotomic approaches.

The paper is structured as follows. In section 2, I identify two desiderata that have constrained natural kind theories and which underpin Bare Projectibilism too. I will call these desiderata the *contrast desideratum* and the *science constraint*, respectively. The first of these states that a natural kind theory ought to explain the intuitive contrast between blatantly arbitrary categories (e.g. *discovered-on-a-Tuesday*) and those that seem to, following the classic metaphor, carve nature at its joints (e.g. *water*). The second desideratum states that a natural kind theory ought not to exclude scientifically legitimate categories. Having introduced these desiderata, in section 3, I give an overview of some of the most important natural kind theories and highlight that, while these theories have generally succeeded in meeting the contrast desideratum, all of them have, in some way or another, violated the science constraint. Then, in section 4, I introduce my updated version of Bare Projectibilism and focus on its two most distinctive features: *graduality* and *abundance*. I argue that the abundance of projectibility constitutes an advantage of Bare Projectibilism *vis à vis* alternative accounts of naturalness, insofar as it makes the theory extremely inclusive and thus, unlikely to violate the science constraint. Interestingly, though, the abundance of projectibility, which is so useful for meeting one of the desiderata, is the source of an important challenge for Bare Projectibilism. For the abundance of projectibility would seem to prevent Bare Projectibilism from meeting the contrast desideratum, as most categories can be said to be at least slightly projectible. I introduce this challenge in section 5, where I argue, not only that Bare Projectibilism meets the contrast desideratum, but more significantly that, by identifying naturalness with a gradual property, Bare Projectibilism meets this desideratum in a more appropriate way than its dichotomic rivals do. In section 6, I defend Bare Projectibilism from views that consider projectibility to be unnecessary for naturalness. More precisely, I discuss two counterexamples from Magnus (2012) and Spencer (2015) respectively, who argue that some scientifically legitimate categories are projectibly weak. In section 7, I conclude.

2. *Two desiderata for a natural kind theory*

One, if not *the* central motivation of natural kind theory is to explain the intuitive contrast that exists between categories that seem clearly arbitrary[1] and those that, following the classic metaphor, carve nature at its joints. Indeed, some groupings seem to correspond to specific anthropocentric concerns or perspectives (e.g. *pet*), while others have often been assumed to correspond to kinds that pre-date our classificatory practices or, at least, that are constrained by the way the world is, rather than by our particular interests. Trying for now not to make any strong commitment, we can identify the first desideratum that a satisfactory account of natural kinds should meet. Let us call this the *contrast desideratum*.

> *Contrast desideratum:* A natural kind theory should explain the intuitive contrast between kinds such as *discovered-on-a-Tuesday* and kinds such as *water*, *tiger*, and *electron*.

In trying to account for this contrast, natural kind theories have often taken *projectibility* to play a central role. Although this notion will be further fleshed out below, the basic and common idea is that alleged natural categories seem to be particularly projectible, meaning that they exhibit a distinctive capacity to support many inductive generalizations (Mill [1843] 1974). A kind such as *tiger*, for instance, which has often been considered a paradigmatic natural kind,[2] can figure in numerous generalizations regarding its behavior, morphology, lineage, etc. As such, upon observing a member of this kind we will be able to *project* onto it many as yet unobserved properties. We will be able to predict, for instance, that it will likely engage in predatory behavior, that it can run as fast as 65 km/h, or that it is a carnivore. Similarly, projections can also be made in the other direction. That is, from particulars to the kind. When zoologists, for instance, observe for the first time a morphological feature or behavior of a not very well-known species, they will often rightly assume that their discovery is not restricted to the observed organism but can, instead, be *projected* to all the members of its kind.

[1] I take the notion of "arbitrariness" to be the best candidate for appropriately contrasting with the philosophically relevant notion of naturalness. Other potential alternatives such as *social* or *artificial*, in contrast, do not seem to be apt. Indeed, the fact that certain entities are the result of human activities does not seem to mark a significant difference. What the notion of naturalness is supposed to capture, instead, is the fact that certain groupings seem to reflect objective differences in the world (social or otherwise). This, in turn, contrasts with those groupings that are the result of anthropocentric interests, or which are simply random collections.

[2] As it will be discussed below, the natural kind-status of biological species is no longer taken for granted. Additionally, some authors argue that species are individuals, not kinds (see Ghiselin 1974 and Hull 1978). This parallel debate, however, will not be addressed here as I am using these examples for expository purposes only, without intending to endorse any position on this specific matter.

In contrast, some categories do not seem to have that sort of inductive power. For instance, there is little we can predict or project by knowing that a particular is a member of the kind *discovered-on-a-Tuesday*: there do not seem to be many things unifying the members of this kind, beyond their membership of the kind itself.

On the face of it, it seems clear that some sort of *contrast* needs to be articulated and that the notion of projectibility can be a good starting point. As said above, this idea is far from original, as many have considered projectibility to be central to the characterization of natural kinds (Boyd 1999: 146; Magnus 2012: 10; Khalidi 2013: 18; Chakravartty 2023: 68).

While projectibility has tended to play a central role in the discussion of natural kinds, it has often been considered insufficient for characterizing naturalness. Indeed, most natural kind accounts do not identify naturalness simply with projectibility, but instead impose further conditions that kinds need to fulfil in order to count as natural. This attitude, often implicitly assumed, is explicitly endorsed by Khalidi (2018: 1381).

One of the reasons why natural kind theorists have considered projectibility to be insufficient for characterizing natural kinds is, I contend, the fear of being *overly inclusive*. For projectibility is arguably *abundant*, in the sense that most of the categories we employ, both within and outside of scientific discourse, exhibit a certain degree of projectibility. If you are not convinced about this abundance, notice that basic categories from ordinary language (e.g. *stone*), and even apparently arbitrary categories (e.g. *things heavier than my head*), allow for certain projections, useless as they might be.

As such, identifying natural kinds with projectible kinds could be considered to violate the contrast desideratum, as categories on both sides of it are at least minimally projectible. Magnus voices this concern when he suggests that one problem with identifying natural kinds with those kinds that support inductive inferences is that non-natural kinds such as *jade* also support many inductive generalizations.[3] He says:

> So it is typical to say that *jade* is not a natural kind. The problem is that there *are* general facts about jade. Both varieties are fairly hard minerals, which makes them *inedible* and *suitable for making stone tools*. These and many other predicates are projectible for jade *simpliciter*. (Magnus 2012: 12, original emphasis)

Although I will ultimately defend a projectibility-based approach, there is a sense in which this "over-inclusiveness fear" is well-founded. Indeed, as I will argue below, the abundance of projectibility is the source of an important challenge for a projectibility-based approach to natural kinds, as it is not immediately clear how such an account would meet the contrast desideratum.

[3] See Bird (2009: 502) for a similar point.

Before confronting the challenge for projectibility-based accounts, though, it is important that we emphasize another element that has played a key role in the development of natural kind theories. For a common assumption throughout the discussion of natural kinds has been that scientific inquiry is particularly well-suited to carve nature at it joints and that, in this sense, scientific categories are particularly good candidates for natural kinds. Although the more precise nature of the relation between natural kinds and scientific categories can take different forms, most authors within the literature have, implicitly or explicitly, endorsed views along these lines (Franklin-Hall 2015: 932; Khalidi 2013: xi-xii; Ereshefsky and Reydon 2015: 972–973).[4]

On the face of it, we can articulate the second desideratum for a natural kind theory as follows. Let us call this the *science constraint*:

> *Science constraint*: an account of natural kinds should not exclude legitimate scientific categories.[5]

As we shall see, the science constraint has played a decisive role throughout the development of natural kind theories. For, in attempting to articulate the contrast desideratum, most natural kind theories have been accused of violating this constraint in some way or other. That being so, this desideratum is responsible for a significant tendency towards inclusiveness that has characterized the development of natural kind theories. The following section considers some of the most important proposals and focuses on their difficulties in respecting the science constraint.

3. *Failures to preserve the science constraint: Towards inclusiveness*

3.1. *Natural kind essentialism*

The most significant, and likely the most discussed violation of the science constraint comes from Natural Kind Essentialism (NKE hereafter). Given that this case has been widely discussed in the literature and that NKE has become a minority position (see Ellis 2001, 2008; Wilkerson 1988), I will not delve far into these ideas here. It is im-

[4] Brian Ellis (2001) can be considered an exception to this attitude, as he is willing to concede that biological categories are not natural kinds. He says: "If evolution occurs in the gradual way that Darwin supposed, or if small changes in genetic constitution can be brought about artificially, then the distinctions between adjacent species—living, dead or yet to be created—must ultimately be arbitrary" (Ellis 2001: 169).

[5] Notice that a more radical version of this constraint might have it that scientific categories—at least in the ultimate stage of inquiry— *perfectly* correspond to natural kinds and, as such, that an appropriate natural kind theory should not only include all legitimate scientific categories, but also exclude non-scientific categories (e.g. folk categories). The alternative presented in this paper is more permissive and, in this sense, only requires natural kind theories *not to exclude* scientific categories, while allowing that some non-scientific categories might count as natural.

portant for our purposes, though, to emphasize that it is precisely its violation of the science constraint that has made NKE a marginal view among philosophers. Chakravartty puts this idea as follows:

> The most obvious and compelling sources of resistance to an exclusive commitment to kinds with essences are the sciences themselves. The kinds of objects investigated by the sciences are sometimes describable in terms of essences, but often resist this sort of description. The traditional view that kinds are ontologically distinguished by essences has a storied past, but many of the kinds one theorizes about and experiments on today simply do not have any such things. (Chakravartty 2007: 157)

As has often been pointed out, the most notorious failure of NKE comes from its incapacity to accommodate biological categories. Indeed, the standard view among philosophers of biology is that biological categories do not fit in a strict essentialist framework, as there is no single genotypic or phenotypic property that would serve to individuate species (Dupré 1981, 1993; Ereshefsky 2007; Khalidi 2013; Magnus 2012; Kitcher 1984).[6]

Given the status of biological species as paradigmatic natural kinds and legitimate scientific categories, essentialism's failure to accommodate them is likely the strongest instance of a violation of the science constraint that we can think of.

3.2. *Homeostatic Property Clusters*

The limitations of NKE in the biological domain constitute the main motivation for Boyd's (1991, 1999) account of natural kinds as Homeostatic Property Clusters (HPC). With this in mind, Boyd's proposal can be read as an attempt to provide a more flexible and inclusive framework that is able to accommodate biological categories, and thus able to preserve the science constraint.

According to the HPC view, natural kinds are clusters of properties whose stable co-occurrence is maintained by homeostatic mechanisms. That is, mechanisms responsible for preserving the properties of the cluster in a state of equilibrium. HPC theory thus departs from NKE in dropping many of its most controversial requirements, and by explaining the inductive potential of natural categories without positing essences.[7]

[6] Notice that although *intrinsic* biological essentialism has been for the most part abandoned (see Devitt 2008 for an exception), some authors have defended an alternative version of biological essentialism which individuates species in terms of extrinsic properties such as ecological or phylogenetic relations (see Griffiths 1999 or Okasha 2002). Still, it is crucial to keep in mind that contemporary biologists work with multiple species concepts, many of which do not individuate species in terms of intrinsic or extrinsic essences. As such, the claim that natural kind essentialism violates the science constraint still holds. I thank an anonymous reviewer for raising this point.

[7] Notice that certain authors consider HPC to be a more relaxed form of essentialism (see, for instance, Kornblith 1993) as the property clusters are playing

This flexibility makes HPC more inclusive than its essentialist predecessor, and arguably a better alternative for accommodating biological categories. Indeed, not only do species lack any good candidate to play the role of an essence, but also, the properties that biological kinds share are often the result of various mechanisms involving environmental pressures, interbreeding, developmental processes, and genetic descent, among other factors. HPC theory thus provides a compelling alternative account of the non-accidental co-occurrence of properties that ground our inductive practices involving biological categories.

Just like its predecessor, however, HPC has been accused of violating the science constraint and excluding legitimate scientific categories. Several authors (Ereshefsky and Reydon 2015; Khalidi 2013; Slater 2015) have argued in this direction and have suggested, with varying degrees of emphasis, that not all natural kinds constitute homeostatic property clusters. While some of these critics concede that HPC accommodates biological kinds, while failing to accommodate other kinds whose equilibrium does not seem to be sustained by homeostatic mechanisms (e.g. chemical elements, fundamental physical particles), some go as far as to insist that HPC does not even fit all biological species (Ereshefsky and Matthen 2005).

Be that as it may, many theorists agree on the idea that, in some way or another, HPC is still too restrictive, as it cannot accommodate the vast heterogeneity of scientifically legitimate categories.

Wary of the difficulties of providing a general theory of natural kinds that is able to encompass this heterogeneity, a recent trend in natural kind theory focuses on the epistemic utility characteristic of natural kinds and avoids making any metaphysical commitment as to what grounds this epistemic utility. More precisely, these views attempt to characterize the clustering of properties while remaining neutral about any specific metaphysical grounding for it. Let us consider these views, which, following Conix (2017), we may refer to as Bare Property Cluster accounts of natural kinds.

3.3. *Bare Property Clusters*

Bare Property Cluster (BPC) accounts of natural kinds constitute a significant departure from previous approaches to natural kinds insofar as they focus on the robust clustering of properties in virtue of which inductive inferences are reliable, without committing to any specific account of this clustering.

This departure is motivated by past failures on the part of previous natural kind theories, which, as we have seen, always seem to violate the science constraint in some way or another. Indeed, BPC defenders

the same epistemic role that essences are taken to play. For the purposes of this work, however, I will be using the label "essentialism" to refer exclusively to the view that identifies essences with necessary and sufficient conditions. I thank an anonymous reviewer for raising this point.

believe that no general grounding claim will be able to account for all natural kinds, and thus that the only way for a notion of natural kinds to be appropriately inclusive is for it to remain neutral regarding the metaphysical grounding of this robust clustering.

Matthew Slater (2015), for instance, claims that natural kind theories have focused too much on the "grounding claim" and should instead turn their attention to the epistemic usefulness of categories, without committing to any specific metaphysical story about essences or homeostatic mechanisms.

Slater argues directly against the HPC view and puts forwards an original proposal (*Stable Property Clusters*) that attempts to articulate more systematically how to understand the *stability* in virtue of which clusters of properties can support inductive generalizations and inferences.

To convey the relevant notion of *stability*, Slater presents the picture of a clique of friends with three members: Peg, Ralph, and Quinn. These individuals form a *stable* clique and like hanging out together. As such, spotting any of these three in the mall is generally a good indicator that the others will be there as well. This is, very roughly, the sense of stability that Slater wants to capture; the instantiation of a property of the cluster reliably indicates the presence of the other co-occurring properties of the cluster.

Similarly, Chakravartty (2007) suggests the metaphor of "sociability" to refer to all the ways in which properties enter into systematic relations and thus ground our inductive practices. As Chakravartty explains, the distribution of properties, or property instances, is not random in space-time. They have a tendency to group together in various ways, showing a degree of *sociability*. The strongest sociability is seen in essence kinds where certain sets of properties are always found together, whereas in other cases, the sociability is less strong and forms looser associations seen in cluster kinds. In this sense, the metaphor of sociability is intended to be neutral about, yet compatible with, more specific grounding accounts of these systematic patterns of sociability.

Interestingly, despite their attempts at inclusivity, it could be argued that even certain BPC accounts end up being too restrictive and violate the science constraint. In this line of thought, Manolo Martínez (2020) has suggested that Slater's SPC account could fail to include what he calls "synergic kinds". Let us flesh this out.

3.4. *Beyond Bare Property Clusters*

Martínez argues that some kinds (i.e. synergic kinds) ground inductive inference not, as in the case of co-occurring property clusters, because the instantiation of a property of the cluster is indicative of the instantiation of other (co-occurrent) properties of the cluster, but instead because "the joint instantiation of all or many of those properties […] plays the explanatory role for which the natural kind is recruited"

(2020: 1935). To illustrate the point and convey more vividly what is different about synergic kinds, Martínez makes use of Slater's clique example, as presented above, while incorporating some modifications.

Martínez suggests that we think of Peg, Ralph and Quinn not as a clique of friends that like each other's company, but instead as a rather tense love triangle. In this case, spotting only one of the three at the mall is not indicative of the presence of either of the other two, while spotting two of them together is a reliable indicator that the third one is *not* going to be there. The idea that this metaphor is meant to convey is that, when it comes to synergic clusters, the instantiation of properties, *individually*, is not a reliable indicator of the instantiation of other properties of the cluster. Instead, it is the joint instantiation of properties that allows for reliable inferences. This type of inference is *synergic,* Martínez argues, because the information gained from observing the instantiation of multiple properties is greater than the sum of the information obtained from each property's separate instantiation.

Martínez makes clear that this discussion is not otiose, as some scientific categories and inferences seem to have this synergic structure. More precisely, Martínez (2020: 1943–1944) suggests that this is the case with some categories involving *epistatic* interactions—the phenomenon in genetics where the effect of one gene on a phenotype is modified by one or more additional genes[8]—and categories from *brain connectomics*—a research program in neuroscience that seeks to uncover how neural connections ("connectomes") give rise to cognitive functions as well as how they are altered by various neurological and psychiatric disorders.[9]

What this discussion reveals, I wish to argue, is that even some BPC views such as Slater's, despite their attempted inclusiveness, seem susceptible to violating the science constraint. On the face of these successive failures, a more promising alternative, I suggest, is to focus exclusively on the inductive power of categories; that is, on projectibility. Indeed, if a recurrent problem of natural kind theories is that they fail to be appropriately inclusive, identifying projectibility with naturalness appears to be a good solution.[10] For not only is

[8] Martínez argues that the sort of non-linear relationships between genes and their effects on traits that characterize epistatic relations are often better described in terms of synergic kinds rather than HPC kinds. For instance, he suggests that fruit-fly wings, whose shape is known to be underwritten by epistatic effects, form a synergic kind (i.e. *fruit-fly wing*) and not a traditional HPC kind (2020: 1942).

[9] Martínez argues that current knowledge about the human connectome highly suggests that an accurate description of the human brain will require more than a characterization in terms of mere aggregation of co-occurrent properties in a cluster. As such, he claims that *human brain* can be considered a highly informationally synergic natural kind (2020: 1943).

[10] It could be argued that identifying naturalness with projectibility runs the risk of conditionalizing the existence of natural kinds to the presence of cognitive agents such as humans capable of drawing such projections. This worry, however, is misplaced as a kind being projectible or not does not depend on an agent drawing

projectibility, as mentioned above, abundant among categories; it is also neutral with regard to specific metaphysical grounding claims. A projectibility-based account, then, will stand out from the rest because of its inclusivity and, as such, will have no trouble in subsuming both Slater's Stable Property Clusters and Martínez's synergic kinds (along with kinds with essences, HPCs, etc.).[11]

As we will see, however, this inclusivity will be the source of a potential problem for this approach that will need to be properly dealt with. In the following section I present the approach. Then I introduce the challenge.

4. *Bare Projectibilism: an update*

As we have seen, most natural kind accounts, while taking projectibility to be *necessary*, rarely deem it to be *sufficient* for naturalness. A notable exception to this tendency, however, is provided by Sören Häggqvist (2005), who has argued for a projectibility-based approach to natural kinds.[12]

My proposal, then, follows the path opened by Häggqvist, but takes two significant steps further, as follows. First, I elaborate the account in response to a serious challenge that is overlooked by Häggqvist. Then I develop some implications that follow from identifying naturalness with projectibility, and which make the proposal depart radically from most traditional approaches to natural kinds. Let us consider each of these ideas in turn.

Häggqvist rightly assumes that a significant benefit of a projectibly-based account of natural kinds is its inclusiveness. As mentioned above,

projections with it but, instead, on whether the properties of the kind tend to co-occur together or not. I thank an anonymous reviewer for raising this worry.

[11] More particularly, when it comes to Martínez's synergic kinds, notice that despite exhibiting a different inferential structure than HPC kinds do, they also ground inductive inference. Indeed, Martínez's main goal when identifying synergic kinds is precisely to expose the limitation of HPC theories to account for the success of our inductive practices. In this sense, an account such as Bare Projectibilism, which identifies naturalness with projectibility, will have no trouble in subsuming synergic kinds too. This discussion is indebted to an anonymous reviewer.

[12] Although Häggvist's view is sometimes included among Bare Property Cluster accounts (see Lemeire 2021; Conix 2017), it is important to highlight an important difference that might set it apart from these views. For, although Häggqvist's view fits in among BPC accounts regarding its neutrality *vis-à-vis* any specific metaphysical grounding for the robust clustering of properties, it departs from these views in incorporating the possibility of certain robust clusters being *brute*. That is, having *no ground*. More precisely, Häggqvist (2005: 82) claims that there is no principled reason to assume that there will always be a causal explanation (be it in terms of essences or more loose causal mechanisms) to account for the clustering of properties. Some of these robust clusters, he claims, might simply be a brute matter of fact. He suggests that this could be the case with certain kinds from fundamental physics, when there does not seem to be any causal explanation for the perfect clustering of properties (2005: 81).

the abundance and neutrality of projectibility makes it difficult for it to exclude any potential natural kind candidates. Häggqvist, however, does not seem to notice that this abundance is a double-edged sword, as it might make the account *overly* inclusive. More precisely, identifying naturalness with projectibility threatens to violate the contrast desideratum and fail to account for the intuitive difference between categories such as *discovered-on-a-Tuesday* and *water*, insofar as categories on both sides of the contrast seem to be, at least, minimally projectible. The challenge for a projectibility-based approach to natural kinds, then, is not to preserve the science constraint—which seems easily satisfied[13]—but to meet the contrast desideratum.

The other aspect that differentiates this proposal from Häggqvist's is its emphasis on an aspect of projectibility that Häggqvist does not consider: *graduality*. For, crucially, projectibility is not an on-off feature of kinds, but, quite to the contrary, a *gradual* property that can be instantiated in varying degrees. Although not often fully exploited, the idea that projectibility is gradual is not an original one (see Dorr 2019: 42; Griffiths 1999; Khalidi 2018; Magnus 2012: 12; Millikan 2000). Furthermore, it is often acknowledged that kinds can be projectible along two different gradual dimensions (Griffiths 1999: 217; Khalidi 2018: 1383; Millikan 2000: 26): on the one hand, the projections or generalizations in which a kind enters can be more or less *robust*. On the other hand, kinds can be more or less projectible depending on the number or *variety* of generalizations they allow for. Let us flesh this out.

Following Khalidi (2018), we can roughly characterize the *robustness* of a generalization by the number of exceptions it has. While some generalizations are universally true and hold under all circumstances, others, although not universal, hold under an exceptionally large range of circumstances, while others hold only under rather specific circumstances and require significant ceteris paribus clauses (Khalidi 2018: 1382).[14]

The *variety* dimension, instead, corresponds to the number of generalizations in which kinds can enter. Although it is generally expected that paradigmatic natural kinds can figure in numerous generalizations—Mill went as far as to hold that "Real Kinds" could enter into *infinitely many* generalizations ([1843] 1974: I vii §4)—Khalidi suggests that some paradigmatic natural kinds might actually figure in very few (e.g. *electron*). Khalidi quickly adds, though, that the latter's poor performance in the variety dimension is compensated by the great (or even universal) *robustness* of the generalizations into which they enter. As we shall see in section 6.1, distinguishing these two dimensions of projectibility will be useful for defending the strong projectibility of certain scientific categories against accusations to the contrary (Spencer 2015; Magnus 2012).

[13] I address objections challenging the *necessity* of projectibility in section 6.

[14] See Woodward (2000) for a more detailed discussion and characterization of non-accidental generalizations.

Now, identifying naturalness with a seemingly abundant and gradual property constitutes a significant departure from traditional natural kind theories that have generally focused on drawing a demarcatory line between natural and non-natural kinds. The view defended here, instead, takes naturalness to be a gradual property and, as such, presents a novel framework where the relevant question is not whether a kind is natural or not (given that most kinds, as we have seen, are at least minimally natural), but instead, its *degree of naturalness*.

As I will argue next, it is precisely the emphasis on the graduality of projectibility, and hence the graduality of naturalness, that will help Bare Projectibilism to address the challenge introduced above and meet the contrast desideratum.

5. *The challenge of projectibility-based accounts of naturalness: Meeting the contrast desideratum*

According to the reconstruction provided above, two main theoretical constraints have driven natural kind research. On the one hand, I have emphasized that the main goal or desideratum of natural kind theories has been to articulate an intuitive contrast between arbitrary categories and those that, following the traditional metaphor, carve nature at its joints (i.e. the contrast desideratum). On the other hand, I have shown how the main attempts to account for this contrast have successively violated the science constraint by excluding scientifically legitimate categories. We have also seen that even arguably very inclusive accounts such as Slater's Stable Property Cluster theory might exhibit this problem.

On the face of it, one sensible alternative, I suggested, is to follow Häggqvist and focus exclusively on the presumably abundant and neutral notion of projectibility. As discussed above, however, the main benefit of projectibility can also constitute a potential weakness, as it is not immediately obvious, given this abundance, how a projectibility-based account would meet the contrast desideratum.

To address this challenge, I suggest that we focus on the *graduality* of projectibility. For, although the *abundance* of projectibility threatens to blur the contrast, its graduality allows us to highlight that *not every kind is projectible to the same degree* and hence, not natural to the same degree. According to Bare Projectibilism, then, the intuitive contrast between categories such as *discovered-on-a-Tuesday* and *water*, is just the contrast between the two extremes of a spectrum. Notice that by identifying naturalness with a gradual property, Bare Projectibilism departs from the tradition of drawing a sharp demarcatory line between natural and non-natural kinds. In what follows, I argue that, far from being a shortcoming of this view, understanding naturalness as a gradual property is the appropriate way to counter the relevant notion of arbitrariness.

To elaborate, notice that if we want our notion of naturalness to stand in appropriate contrast with the notion of arbitrariness, we need a characterization that captures nuanced differences and not only extreme ones. Consider for instance, the kind *pet*.[15] Although this kind seems *more arbitrary* than the kind *tiger*, it does seem *less arbitrary* than the kind *animals-belonging-to-the-emperor*.[16] Similarly, although everything seems to suggest that the kind *tiger* is not arbitrary, we also have reasons to think that it is *more arbitrary* than the kind *gold*. The more examples we consider, the clearer it will be that it does not seem possible to separate all kinds into two perfectly discrete boxes, as the contrast desideratum would have us believe. What this suggests, instead, is that the difference considered in the contrast desideratum is but one particular *extreme* instance of a more general and ubiquitous relation: *more natural than*.

As soon as we appreciate this, we can see that Bare Projectibilism is in a better position than alternative dichotomic accounts to articulate this more general relation. For dichotomic accounts, insofar as they posit a single sharp demarcatory line, are only able to capture the particular extreme case, and not the more specific ones. In this sense, they are unable to account for the more general relation *more natural than* of which the contrast desideratum is but one (extreme) instance.[17]

That being so, a projectibility-based account which characterizes naturalness in terms of a gradual property seems particularly well suited to counter the relevant notion of arbitrariness satisfactorily, and to accommodate both extreme and nuanced contrasts. This is the sense in which I contend that Bare Projectibilism not only meets the contrast desideratum, but does so in a more appropriate way, as it also accommodates the more general cases that dichotomic accounts do not accommodate.

As an additional illustration of the potential limitations of dichotomic accounts of naturalness, consider the much discussed revision of the concept FISH which, roughly, went from tracking the kind *aquatic animal*—which included whales and certain invertebrates such as clams, starfish, etc.—to tracking the kind *cold-blooded vertebrate with gills*. Let us call the former $fish_{AQUATIC}$ and the latter $fish_{GILLS}$. Although it is uncontroversial that the current English term 'fish' refers to $fish_{GILLS}$, philosophers disagree on the "natural kind" status of these two kinds. According to John Dupré's (1993) Promiscuous Realism, insofar as both kinds stress important sameness relations and serve legitimate purposes, they should both count as natural (1981: 92). On Khalidi's

[15] Interestingly, Khalidi (2018) uses this kind as an example of a paradigmatic *non*-natural kind.

[16] See Borges's (1942) essay "The Analytical Language of John Wilkins" and the curious taxonomy of animals suggested there.

[17] See Lewis (1983) for a different view on the graduality of naturalness.

more restrictive view, in contrast, only the alleged[18] scientific category $fish_{GILLS}$ counts as a natural kind (2013: 62).

I want to use this case to illustrate that, independently of whether or not one counts $fish_{AQUATIC}$ as a natural kind, a dichotomic approach will face some significant limitations and will lead to some counterintuitive results. As such, I contend that the problem of these views does not stem from *where* they draw the natural/non-natural demarcatory line but, rather, from drawing such a line at all. Let us consider this case in more detail.

Dupré's Promiscuous Realism tells us that, provided we do not associate the notion of a natural kind with essentialist views, we have good reasons to think of $fish_{AQUATIC}$ and $fish_{GILLS}$ as equally natural. More particularly, Dupré believes that scientific categories are not fundamentally different from folk ones (1999: 462) and, as such, does not see any reason to dismiss the folk category $fish_{AQUATIC}$ as non-natural. Although Dupré is certainly right to emphasize the utility of this kind and the fact that there does not seem to be any fundamental difference between $fish_{AQUATIC}$ and $fish_{GILLS}$, his account does not tell us anything about the intuitively plausible *improvement* that has occurred in the conceptual transition from $fish_{AQUATIC}$ to $fish_{GILLS}$. While I agree with Dupré in not thinking that there is any fundamental or metaphysical difference between these two kinds, I believe, however, that it makes sense to think of $fish_{GILLS}$ as being *more natural* than $fish_{AQUATIC}$. For one thing, the kind $fish_{GILLS}$ groups particulars in a way that seems to allow for more interesting generalizations than the kind $fish_{AQUATIC}$ and, additionally, seems to provide a deeper understanding of the aspect of reality it represents. Notice that Khalidi also emphasizes this apparent contrast, and after insisting that the category $fish_{AQUATIC}$ has no inductive value (2013: 62), he suggests that the category $fish_{GILLS}$, in turn, is scientifically useful. He says:

> It is instructive to contrast this inclusive use of the term 'fish' [$fish_{AQUATIC}$] with the 'scientific one'[$fish_{GILLS}$]. […] Despite the fact that it is not a unitary taxon from the evolutionary or phylogenetic point of view, the category *fish* [$fish_{GILLS}$] has undisputed value as an epistemic kind. There are a number of branches of science, such as ichthyology and marine biology, which use this category to explain and predict natural phenomena. (Khalidi 2013: 62–63)

Although I will ultimately suggest that Khalidi goes too far in positing a fundamental difference between these two categories, a permissive dichotomic account such as Dupré's, which locates both $fish_{AQUATIC}$ and $fish_{GILLS}$ on the "natural side" of the divide, is not satisfactory either, as

[18] Dupré (1999) casts serious doubt on the status of $fish_{GILLS}$ as a scientific category. Indeed, notice that $fish_{AQUATIC}$ is not a monophyletic kind and thus, according to authors influenced by cladism, not an objective scientific category (see Boucher 2022).

it is unable to articulate this intuitive difference in terms of naturalness. To be clear, my contention against Dupré's account does not target its promiscuity or permissiveness. I am actually very sympathetic to this attitude. My complaint is, rather, that we need to complement this permissive picture with a gradual account in order to emphasize significant differences that will otherwise remain overlooked. Let us turn to consider the other side of the picture: Khalidi's more restrictive approach to the case.

Khalidi believes that "not all purposes are created equal" (2013: 62) and that kinds introduced for epistemic purposes have to be prioritized over those that serve other non-epistemic or pragmatic purposes. As such, he argues that scientific categories will tend to correspond to natural kinds, whereas folk ones will not. Unsurprisingly, then, Khalidi dismisses $fish_{AQUATIC}$ as non-natural, while insisting that $fish_{GILLS}$ is a natural kind. The problem, again, is that a single sharp demarcatory line is not enough to capture the nuanced differences that kinds may exhibit. For, while it is plausible to think that there is a contrast between $fish_{AQUATIC}$ and $fish_{GILLS}$ in terms of naturalness (as Khalidi duly emphasizes), it is equally plausible to think that a similar contrast arises when we compare the allegedly non-natural $fish_{AQUATIC}$ with a random category such as *wet creature*; a contrast which, I contend, an account of naturalness ought to capture. Khalidi, however, is unable to account for such differences in terms of naturalness given that, on his view, both $fish_{AQUATIC}$ and *wet creature* are equally non-natural.

The limitation of having only two discrete options (either natural or non-natural) also explains why Khalidi seems forced to overstress the difference between $fish_{AQUATIC}$ and $fish_{GILLS}$, and refer to the former as if it were inductively worse than it actually is. He says: "When the category fish includes aquatic animals, such as crayfish, jellyfish, starfish, and some mollusks, as well as whales and dolphins, it ceases to have value as an inductive category" (Khalidi 2013: 62). I believe, however, that Khalidi is too quick in making this assessment. Indeed, as we noted above, the abundance of projectibility guarantees that most categories, including $fish_{AQUATIC}$, will exhibit some degree of projectibility. In this particular case, a category such as $fish_{AQUATIC}$, although inductively weaker than $fish_{GILLS}$, can still have a significant inductive value. Notice, for instance, that knowing that x is a member of the kind $fish_{AQUATIC}$ allows us to know, among other things, that x lives in the water for all or most of its life, that x requires water to survive, that x has adapted to move efficiently through water, etc.

Hopefully, this discussion has served to illustrate that approaching these cases equipped with only two discrete boxes constitutes a serious limitation for dichotomic accounts of naturalness. A gradual account such as Bare Projectibilism, in contrast, seems better able to accommodate both the extreme contrasts and the more nuanced ones.

6. *Bare Projectibilism: a defense*

As suggested above, one of the reasons why many authors have resisted projectibility-based accounts is a fear of being overly inclusive (recall Magnus's (2012: 12) resistance to counting *jade* as a natural kind). As such, many natural kind theorists have taken projectibility to be necessary but insufficient for naturalness and have thus come up with further conditions for demarcating natural from non-natural kinds.

Interestingly, some authors have voiced concerns with projectibility-based accounts that take the opposite direction, as it has also been argued that projectibility may not, after all, be *necessary* for naturalness. More precisely, some authors have argued that some scientifically legitimate categories are not very projectible and, as such, that a projectibility-based account will fail to be appropriately inclusive. This worry is to be taken seriously; for, if these considerations were right, Bare Projectibilism would, in its own way, also violate the science constraint. I will consider two such arguments. First, I will present Quayshawn Spencer's argument regarding the poor inductive power of superheavy elements. Then, I will turn to considering a similar contention from Magnus involving polymorphic species. My strategy for resisting these potential counterexamples will consist in arguing that neither Spencer nor Magnus succeed in making the case for the poor projectibility of their respective examples. I will thus argue that both superheavy elements and polymorphic species are significantly projectible categories and, thus, (non-trivially) satisfy the science constraint.

Having anticipated this, let us consider Spencer and Magnus's potential counterexamples in more detail.

6.1. *Superheavy Element 117*

Spencer argues that a natural kind theory that focuses exclusively on the inductive power of kinds (i.e. projectibility) will fail to include certain paradigmatic natural kinds which, he claims, are "notoriously inductively weak" (2016: 162). To substantiate this idea Spencer presents the case of superheavy elements, and focuses in particular on element 117, also known as "tennessine". Indeed, given the seemingly indisputable status of chemical elements as paradigmatic natural kinds, it would be problematic for any theory of naturalness to exclude such paradigmatic exemplars or, in the case of a gradual account, to ascribe them the same degree of arbitrariness as categories such as *discovered-on-a-Tuesday* and the like.

I will suggest, however, that Spencer does not succeed in making the case for the weak projectibility of element 117. More precisely, I will argue that this element supports relevant inductive generalizations and that Spencer's incorrect assessment derives from conflating projectibility with other notions in the vicinity, such as our *capacity* to draw inductive inferences, or the inductive *method*. Let's consider Spencer's view in more detail. Concerning chemical element 117, he says:

> Since only six atoms of element 117 have ever been synthesized, and since the atoms that have been synthesized have existed for less than a second, nuclear chemists have not been able to get "many inductive generalizations" out of 117. Furthermore, the latter is not a temporary setback. Due to the nuclear instability of 117, it is not the sort of kind that we *can* generate many inductive generalizations with it. Thus, unlike other elements, we know nothing about 117's properties at standard temperature and pressure—such as its phase, its density, its melting point, its boiling point, its ionization energies, or its atomic radius. […] So, natural kind theories that require natural kinds to be inductively powerful fail to predict the existence of inductively weak paradigm natural kinds, such as superheavy elements. (Spencer 2016: 162)

The first thing to notice is that Spencer's claim regarding the weak projectibility of *tennessine* should not be understood merely as stating that, given its nuclear instability, we lack the *capacity* to learn as many things about it as we can about other, more stable elements. For this limitation would simply amount to us *knowing* comparatively fewer projections supported by this category, but would not be indicative of the category being projectibly poor.[19] Rather, Spencer's claim must not only be that we cannot *learn* tennessine's properties but, more radically, that tennessine *lacks* the relevant properties typical of other non-superheavy elements (e.g. melting point, density, etc.) and, as such, that there are few projections we can make about it.

With this clarification in mind, in what follows I put forward various considerations that cast serious doubt on this view. As such, I argue that we have no compelling reasons to believe that *tennessine* (along with other superheavy elements) is significantly less projectible than other chemical elements.

First of all, notice that the intrinsic instability of tennessine already constitutes a very *robust* general fact about this element; one on which the experiments to synthesize it heavily relied.[20]

Although I will also contend that tennessine has many other projectible properties, notice that having a very robust one (i.e. instability) is already a good indicator that this category is inductively powerful. For, we may recall, the projectibility of a category depends not only on the *variety* of projections that it supports, but also on the *robustness* of those projections. As Khalidi suggests, the fact that projectibility ranges over two dimensions allows some very projectible categories to be so, not in virtue of supporting *many* inductive generalizations, but instead in virtue of the (few) generalizations they support being very robust

[19] To see this through an example, consider the case of *Phobaeticus chani*, a stick insect with outstanding camouflage skills. We know very little about this insect, partly because only a few specimens have been observed to date. It seems clear, though, that it cannot be deduced from this epistemic limitation and our corresponding lack of knowledge about this insect that this category is inductively weaker than any other species category that is more easily observed and studied.

[20] Slater (2013: 147) makes a similar point concerning the "stable instability" of uranium.

(or the other way around). This could be the case, Khalidi suggests, with some kinds from fundamental physics, such as *electron*, which although generally characterized only in terms of three properties (spin, charge, and weight), is a very projectible category due to the fact that these properties are perfectly clustered. More generally, Khalidi (2018: 1383) suggests that when it comes to the utility of kinds for scientific inquiry, low performance in one of the two dimensions can be compensated by a high score in the other.

Now, apart from the robust instability of tennessine, notice that, although it is certainly the case that we cannot observe and measure the behavior and properties of this element by conventional means, we can nonetheless build models to *predict* many of its properties. This is crucial, as it suggests that, even when it comes to the *variety* dimension of projectibility, *tennessine* performs significantly better than what Spencer would have us believe. More specifically, some of these models have predicted that tennessine's melting point will be somewhere in the range of 350–550 ºC (Hoffman, Lee and Pershina 2010: 1728), that its boiling point is 610 ºC (Takahashi 2002), that it has a density between 7.1 and 7.3 g/cm^3 (Bonchev and Kamenska 1981),[21] and that it is *solid* at standard temperature and pressure (Bonchev and Kamenska 1981). Additionally, values for its ionization energies (Chang, Li, and Dong 2010) and atomic radius (Bonchev and Kamenska 1981) have also been predicted.

Unfortunately for Spencer, these predictions are clearly at odds with the view that tennessine *lacks* the relevant properties and, in this sense, in stark tension with his assessment regarding its poor projectibility. They suggest not that tennessine is "notoriously inductively weak" but, quite to the contrary, that it can support a significant number of inductive generalizations.

Now, while I take these considerations to be sufficient to make the case that *element 117* is significantly projectible, there is another idea that might serve to strengthen the case, and which is thus worth presenting. For, according to the standard view of quantum physics, radioactive decay—the phenomenon in virtue of which unstable elements are short-lived—is probabilistic. This is important, as it entails that there is always an infinitesimal chance of a sample of tennessine lasting long enough to be tested, manipulated, measured, etc. Although extremely improbable, the fact that this constitutes a possibility gives us further reason to believe that the nuclear instability of tennessine, although an important epistemic limitation, does not affect its metaphysical status and, as such, does not constitute a reason to doubt that this element is *as projectible as* any other chemical element.

[21] Notice that these results being presented in terms of intervals is again, due to an epistemic limitation. The idea is not that tennessine has no precise melting point or density, but rather that our current means of prediction do not allow us to go beyond predicting ranges.

Finally, one could object that the fact that tennessine's properties have been discovered through an abductive rather than an inductive method suggests that this category is not very projectible. Indeed, Spencer seems to have something like this in mind when he suggests, later on: "117's lack of inductive power does not undercut its epistemic utility in nuclear chemistry. It's just that its epistemic utility is different. It's abductive, not inductive" (2016: 162). I contend, however, that deeming element 117 weakly projectible for such reasons would amount to confusing projectibility with the inductive *method*. In conflating these two notions, one would fail to notice that the reason that projectibility is often considered distinctive of natural categories is not connected with the *method* through which we learn generalizations about them, but rather with the very fact that they support such generalizations. The distinctive feature of natural kinds—and the reason for which projectibility has generally been taken to be characteristic of them—is not that we learn things about them through any particular method (e.g. observation of particular members, followed by inductive generalization to the whole kind), but rather, that what we know and learn about them is projectible to all the members of the kind. In this sense, I conclude, *contra* Spencer, that *element 117* is significantly projectible and, accordingly, does not constitute a successful counterexample to a projectibility-based account of naturalness.

6.2. *Magnus's polymorphism*

A similar case against the *necessity* of projectibility for naturalness is put forward by P. D. Magnus. He contends that focusing only on the inductive power of categories to determine natural kindhood risks overlooking certain legitimate scientific categories which do not appear to be very projectible.

More to the point, Magnus suggests that focusing on projectibility ultimately leads to focusing on—and eventually overemphasizing—*similarity*. This is so, he insists, because the projectibility of a category is grounded in its members' sharing many relevant properties. He says:

> Coming at natural kinds in this way [by focusing on projectibility] leads us to suppose that members of a natural kind are connected by similarity. The reason that this *A* can be used as a proxy for other *A*s is that they all resemble one another in many respects. (Magnus 2012: 11)

With this connection in mind, Magnus goes on to complain that certain natural kind theories such as HPC have focused too much on similarity, and have therefore failed to see that scientific taxonomy does not always seek to individuate categories by stressing similarities. He refers to this alleged tendency of overemphasizing similarity and projectibility as *similarity fetishism*. He says:

> Quine is part of a tradition, going back to Mill, which assumes that membership in the same kind is a matter of having a large number of properties in common. Call this similarity fetishism. The yoke of similarity fetishism

> makes the induction assumption unable to accommodate kinds which are not joined by similarity and thus makes it insufficient to serve as a definition of 'natural kind'. (Magnus 2012: 12)

To illustrate his point, Magnus focusses on the case of polymorphic species. That is, species whose members can be grouped in different subcategories according to significant recurring differences. Although many species are polymorphic (one of the most common examples being sexual dimorphism in mammals), some species stand out from the rest by exhibiting remarkably extreme differences (in morphology, behavior, etc.). According to Magnus, a projectibility-based approach to naturalness that "fetishizes" similarity among the instances of a kind would thus have no reason to group these extremely divergent morphs under the same category. To make his case more vivid, Magnus offers the example of the highly sexually dimorphic seadevil.[22] He says:

> Take a specific seadevil species, such as *Linophryne arborifera* […]. Females and males are so dissimilar that there are few inductions one can make about the species in general from a single sample. If one were simply looking for projectible predicates, then the species would not be a relevant kind at all. (Magnus 2012: 160)

In what follows I will try to make the case, *contra* Magnus, that highly polymorphic species such as *Linophryne arborifera* are significantly projectible, or at least substantially more so than what he suggests. More precisely, I will argue that polymorphic species, despite diverging significantly in morphological and behavioral features, still share some very important diachronic properties (e.g. shared ancestry), in a way that supports many relevant inductive generalizations. Additionally, I will also suggest that polymorphic species share many relevant synchronic properties related to their impact on ecosystems, their habitat, and even their morphology.

Before getting into the details of Magnus's case, though, notice that Ereshefsky and Reydon (2015, 2021) raise a similar worry against projectibility-based views. Their contention is that biological taxonomy often focuses on highlighting *history* or *ancestry*, which, they suggest, does not always overlap with similarity. They say:

> The challenge for those that assert that natural kinds are groups of entities with numerous similarities is that classifying by similarity and classifying by history can conflict. And when they do conflict, the view that natural kinds are inductive kinds fails to capture the classificatory practices of those biologists that classify by history. (Ereshefsky and Reydon Forthcoming)

Magnus also seems to draw this contrast between *history* and *similarity* when he suggests that what unifies the members of a (dimorphic)

[22] To get an idea of how significant the differences between the female and male morphs of this species are, notice that the males, which are five times smaller than the females, were for a long time thought to be parasites attached to the females' bodies, until it was later discovered that they were essential for reproduction (these cases are known as "sexual parasites").

species is not similarity but, rather, "a common causal history over evolutionary time" (2012: 162).

I will argue, however, that Ereshefsky and Reydon (2015, 2021), and Magnus (2012), are too quick to assume that similarity amounts to "superficial similarity" or, more precisely, to *intrinsic similarity*. For there does not seem to be any principled reason not to count "shared history" or "shared ancestry" among the relevant properties that members of a (dimorphic) species *share*, and thus among the properties in virtue of which they can be considered to be significantly *similar*. Moreover, not only are these *extrinsic similarities* relevant from the point of view of evolutionary biology, but crucially for our purposes, they ground many important inductive generalizations. This point is vividly made by Chakravartty (Forthcoming: 6) who, against Ereshefsky and Reydon, insists that the focus of biological taxonomy on ancestry is not in tension with highlighting inductively powerful categories. Quite to the contrary, the aim is still to make inductive inferences.

Khalidi (2021) too, in investigating the aptness of etiological kinds as natural kind candidates, also suggests that these kinds, characterized by sharing diachronic properties—a subtype of extrinsic properties—support *retrodictions* (i.e. predictions of the past), which are a particular form of projection. He says: "For instance, if we identify a rock as a meteorite based on its fusion crust, we can infer that it had an extra-terrestrial origin and a certain causal trajectory through the earth's atmosphere" (2021: 14). Similarly, if we identify an organism as a *Linophryne arborifera,* we can infer, for instance, how closely related it is to another given organism. Faced with these considerations, I argue that we have reasons to think that *Linophryne arborifera*, as well as other polymorphic species, will support many important retrodictions involving their evolutionary history (e.g. evolutionary closeness to other species, developmental pathway, etc.).

Additionally, as if acknowledging these shared diachronic similarities were not enough to defend *Linophryne arborifera*'s status as a significantly projectible category, notice that members of this species also share relevant synchronic properties related to their impact on ecosystems, their habitat, and even their morphology. Interestingly, even Magnus acknowledges that members of this species category share important morphological traits (Magnus references Pietsch (2009: 24–30) as providing an extended account of morphological traits shared by both morphs). Somewhat surprisingly, though, Magnus does not seem to take these morphological similarities into consideration when it comes to assessing the projectibility of the category. The reason for this, he suggests, is that the "properties of males are insufficient to diagnose species" (Magnus 2012: 161). This consideration, however, even if true, does not jeopardize the projectibility of the category as a whole. For, independently of whether the morphological traits of males are enough to individuate the species or not, inasmuch as both morphs share morphological properties that are relevant from a biological

standpoint, these shared similarities contribute towards making the species category more projectible in the relevant sense.[23]

Overall, these considerations suggest that Magnus overestimates the impact of the divergent female and male morphologies on the projectibility of polymorphic species categories such as *L. arborifera*. As such, I conclude, *contra* Magnus, that it is not true that "if one were simply looking for projectible predicates, then the species would not be relevant at all" (2012: 160), and that *L. arborifera* is significantly projectible (in the specific non-trivial sense specified above).

Finally, it could perhaps be argued that Magnus's case against projectibility-based accounts is not only based on the idea that this category is weakly projectible—which, as we just saw, seems doubtful—but, additionally, on the claim that projectibility is not the relevant feature in virtue of which different domains of biology favor this category. More precisely, Magnus suggests that the rationale for grouping together the members of a (dimorphic) species is not similarity or projectibility but, instead, *explanatory* considerations. He says: "Explanatory considerations identify *L. arborifera* as a legitimate taxon, even if it is not an inductively robust category" (2012: 162). The picture Magnus presents, then, is one where explanatory considerations are in tension with, and (sometimes) prioritized over, similarity and projectibility.

This view, however, is not without controversy. As Miles MacLeod (2014) suggests in his review of Magnus's monograph, not only does Magnus provide no account of what makes a kind explanatory *qua* kind but, moreover, "it is also arguable that what grounds a kind as explanatory is similarity among its members in the first place" (MacLeod 2014: 337). Importantly for our purposes, if something along the lines of MacLeod's view is correct, then, by emphasizing the explanatory value of *L. arborifera*, Magnus would not thereby discount its projectibility but, quite to the contrary, provide further reasons in favor of this category being projectible in the relevant sense.

Notice that Magnus's own example seems to point in this direction. Indeed, we have seen that he identifies as explanatorily relevant the fact that members of *Linophryne arborifera* have a *common causal history*. But, if the above considerations are on the right track, focusing on a common causal history amounts to focusing on *similar* extrinsic properties. In this sense, his example does not involve any tension between explanatoriness and projectibility but, instead, a case where both dimensions are grounded in the extrinsic similarities of the category.

Accordingly, I conclude that Magnus's case does not succeed as a counterexample to projectibility-based accounts of naturalness. I have

[23] To draw a simple comparison, consider two important shared morphological similarities of tigers: *having stripes*, and *having four legs*. It is clear that these two morphological properties by themselves are not enough to individuate the species (i.e. *Panthera tigris*). Still, these similarities contribute towards making the tiger category *more projectible*. The same goes, I suggest, for more extreme cases of dimorphic species such as *Linophryne arborifera*.

argued that, in virtue of their intrinsic and extrinsic similarities, the categories that correspond to polymorphic species support many relevant inductive generalizations. Moreover, I have shown that even if Magnus is right in his claim that the rationale for grouping polymorphic organisms under a single species category is the explanatory potential of the resulting category, this does not pose a challenge to its projectibility but, quite to the contrary, provides further reason not to doubt it.

7. *Conclusion*

In this paper I have put forward an original view according to which the naturalness of a kind is to be identified with its degree of projectibility. Although projectibility has traditionally been given a prominent role in natural kind theories, the current proposal departs from other theories in singling out no other additional condition for naturalness. As such, a distinctive characteristic of Bare Projectibilism is that, by identifying naturalness with a gradual property such as projectibility, the notion of naturalness itself becomes one of degree. Rather than constituting a shortcoming of the view, I have argued that understanding naturalness in a gradual way not only appropriately counters the relevant notion of arbitrariness but, moreover, brings important advantages over dichotomic alternatives. Finally, I have addressed objections involving potential counterexamples to a projectibility-based account.

References

Bird, A. 2009. "Essences and Natural Kinds." In R. Le Poidevin, P. Simons, A. McGonigal and R. P. Cameron (eds.). *The Routledge Companion to Metaphysics*. New York: Routledge, 497–506.

Bonchev, D. and Kamenska, V. 1981. "Predicting the Properties of the 113–120 Transactinide Elements." *Journal of Physical Chemistry* 85 (9): 1177–1186.

Borges, J. L. [1942] 1964. "The Analytical Language of John Wilkins." In: *Other Inquisitions, 1937-1952*. Austin: Texas University Press, 101–105.

Boucher, S. C. 2022. "Cladism, Monophyly and Natural Kinds." *Croatian Journal of Philosophy* 22 (64): 39–68.

Boyd, R. 1991. "Realism, Anti-Foundationalism, and the Enthusiasm for Natural Kinds." *Philosophical Studies* 61: 127–48.

_____. 1999. "Homeostasis, Species, and Higher Taxa." In . R. A. Wilson (ed.). *Species: New Interdisciplinary Essays*. Cambridge: MIT Press, 141–185.

Chakravartty, A. 2007. *A Metaphysics for Scientific Realism: Knowing the Unobservable*. Cambridge: Cambridge University Press.

_____. 2023. "Last Chance Saloons for Natural Kind Realism." *American Philosophical Quarterly,* 60 (1): 63-81.

Chang, Z., Li, J. and Dong, C. 2010. "Ionization Potentials, Electron Affinities, Resonance Excitation Energies, Oscillator Strengths, and Ionic Radii of Element Uus (Z = 117) and Astatine." *Journal of Physical Chemistry* 114 (51): 13388–13394.

Conix, S. 2017. *Radical Pluralism, Ontological Underdetermination, and the Role of Values in Species Classification*. PhD thesis. https://api.repository.cam.ac.uk/server/api/core/bitstreams/33c9e973-4478-4dc4-b6d0-c869a6e84761/content

Devitt, M. 2008. "Resurrecting Biological Essentialism." *Philosophy of Science* 75 (3): 344–382.

Dorr, C. 2019. "Natural Properties." *The Stanford Encyclopedia of Philosophy* (Fall 2009 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/fall2019/entries/natural-properties

Dupré, J. 1981. "Natural Kinds and Biological Taxa." *Philosophical Review* 90 (1): 66–90.

_____. 1993. *The Disorder of Things: Metaphysical Foundations of the Disunity of Science*. Cambridge: Harvard University Press.

_____. 1999. "Are Whales Fish?" In D. L. Medin and S. Atran (eds.). *Folkbiology*. Cambridge: MIT Press, 461–476.

Ellis, B. 2001. *Scientific Essentialism*. Cambridge: Cambridge University Press.

_____. 2008. "Essentialism and Natural Kinds." In S. Philos and M. Curd (eds.). *The Routledge Companion to Philosophy of Science*. New York: Routledge, 139–149.

Ereshefsky, M. 2001. *The Poverty of the Linnaean Hierarchy*. Cambridge: Cambridge University Press.

_____. 2007. "Species, Taxonomy, and Systematics." In M. Matten and C. Stephens (eds.). *Handbook of the Philosophy of Science: Philosophy of Biology*. Amsterdam: North-Holland, 403–427.

Ereshefsky, M. and Matthen, M. 2005. "Taxonomy, Polymorphism and History: An Introduction to Population Structure Theory." *Philosophy of Science* 72 (1): 1–21.

Ereshefsky, M. and Reydon, T. 2015. "Scientific Kinds." *Philosophical Studies* 172: 969–986.

_____. Forthcoming. "The Grounded Functionality Account of Natural Kinds." In W. Bausman, J. Baxter, O. Lean, A. Love, C. K. Waters (eds.). *From Biological Practice to Scientific Metaphysics: Minnesota Studies in the Philosophy of Science*. Minneapolis: University of Minnesota Press.

Franklin-Hall, L. R. 2015. "Natural Kinds as Categorical Bottlenecks." *Philosophical Studies* 172: 925–948.

Ghiselin, M. 1974. "A Radical Solution to the Species Problem." *Systematic Zoology* 23: 536–544.

Goodman, N. 1973. *Fact, Fiction, and Forecast: Third Edition*. Indianapolis: Bobbs Merrill.

Griffiths, P. E. 1999. "Squaring the Circle: Natural Kinds with Historical Essences." In R. A. Wilson (ed.). *Species: New Interdisciplinary Essays*. Cambridge: MIT Press, 209–228.

Häggqvist, S. 2005. "Kinds, Projectibility and Explanation." *Croatian Journal of Philosophy* 13: 71–87.

Hoffman, D. C., Lee, D. M. and Pershina, V. 2008. "Transactinide Elements and Future Elements." In L. R. Morss, N. M. Edelstein and J. Fuger (eds.). *The Chemistry of the Actinide and Transactinide Elements*. Dordrecht: Springer, 1652–1752.

Hull, D. 1978. "A Matter of Individuality." *Philosophy of Science* 45: 335–360.

Khalidi, M. 2013. *Natural Categories and Human Kinds: Classification in the Natural and Social Sciences*. Cambridge: Cambridge University Press.

_____. 2018. "Natural Kinds as Nodes in Causal Networks." *Synthese* 195: 1379–1396.

_____. 2021. "Etiological Kinds." *Philosophy of Science* 88 (1): 1–21

Kitcher, P. 1984. "Species." *Philosophy of Science* 51: 308–333.

Kornblith, H. 1993. *Inductive Inference and Its Natural Ground*. Cambridge: MIT Press.

Lemeire, O. 2021. "The Causal Structure of Natural Kinds." *Studies in History and Philosophy of Science* 85: 200–207.

Lewis, D. 1983. "New Work for a Theory of Universals." *Australasian Journal of Philosophy* 63: 343–377.

MacLeod, M. 2014. "Following Through on Naturalistic Approaches to Natural Kinds." *Metascience* 23: 335–338.

Magnus, P. D. 2012. *Scientific Enquiry and Natural Kinds: From Planets to Mallards*. New York: Palgrave Macmillan.

Martínez, M. 2020. "Synergic Kinds." *Synthese* 197 (5): 1931–1946.

Mill, J. S. [1843] 1974. "A System of Logic." In J. M. Robson (ed.). *The Collected Works of John Stuart Mill*. Volume VII. Toronto: University of Toronto Press.

Okasha, S. 2002. "Darwinian Metaphysics: Species and the Question of Essentialism." *Synthese* 131 (2): 191–213.

Pietsch, T. W. 2009. *Oceanic Anglerfishes: Extraordinary Biodiversity in the Deep Sea*. Berkeley: University of California Press.

Slater, M. 2013. *Are Species Real? An Essay on the Metaphysics of Species*. New York: Palgrave Macmillan.

_____. 2015. "Natural Kindness." *British Journal of Philosophy of Science* 66: 315–411.

Spencer, Q. 2016. "Genuine Kinds and Scientific Reality." In C. Kendig (ed.). *Natural Kinds and Classification in Scientific Practice*. New York: Routledge, 157–172.

Takahashi, N. 2002. "Boiling Points of the Superheavy Elements 117 and 118." *Journal of Radioanalytical and Nuclear Chemistry* 251 (2): 299–301.

Quine, W. V. 1969. "Natural Kinds." In W. V. Quine. *Ontological Relativity and Other Essays*. New York: Columbia University Press, 114–138.

Wilkerson, T. E. 1988. "Natural Kinds." *Philosophy* 63 (243): 29–42.

Woodward, J. 2000. "Explanation and Invariance in the Special Sciences." *British Journal for the Philosophy of Science* 51: 197–254.

Croatian Journal of Philosophy
Vol. XXIII, No. 68, 2023
https://doi.org/10.52685/cjp.23.68.4
Received: November 23, 2022
Accepted: April 3, 2023

A Tension in Some Non-Naturalistic Explanations of Moral Truths

MAARTEN VAN DOORN
Radboud University, Nijmegen, Netherlands

Recently, there has been some excitement about the potential explanatory payoffs the newish metaphysical notion of grounding seems to have for metaethical non-naturalism. There has also been a recent upsurge in the debate about whether non-naturalism is implausibly committed to some acts being wrong because of some sui generis piece of ontology. It has, in response, been claimed that once we have a clear enough picture of the grounding role of moral laws on non-naturalism, this is not (objectionably) so. This move, I argue, is inconsistent with certain constraints on what non-naturalist-friendly moral laws must be for them to do the explanatory work non-naturalism requires of them elsewhere. In other words, there is tension between the grounding reply to the supervenience objection and the grounding structure implied by some responses to the normative objection.

Keywords: Non-naturalism; meta-ethics; grounding; moral justification; moral explanation.

1. *Introduction*

According to metaethical non-naturalism, there are moral properties and facts that are objective (mind-independent) and metaphysically robust (the non-naturalist's notion of moral properties and facts carries ontological commitment).[1] The nature of these robust properties (including relations) is aptly characterized in terms of inherent, authoritative guidance. That makes them *sui generis*, non-natural and (in

[1] Hence the difference with so-called 'quietist' or 'non-realist' versions of moral non-naturalism: these views (seek to) avoid this ontological commitment (Parfit 2011, 2017).

some sense) isolated from the causally efficacious properties that shape the content of our beliefs about the empirical world.

Put like this, the claim that normative properties are non-natural, serves as a theoretical claim about the metaethical status they have to possess if—as non-naturalism sees things—our theories are to capture a robust sort of ethical objectivity and normativity. They will have to be non-natural facts and properties because they must be *irreducibly evaluative* (cf. Fitzpatrick 2018: 554).

Stephany Leary (2016: 8), for instance, writes that "[Non-naturalism] takes the very nature of these properties to involve something like *to be promoted-ness* or *to be considered-ness* (or *to be doneness*, as Mackie (1977) says), so that they objectively 'call out' for certain responses in us." For example, the non-naturalist may take *being right* to be a *sui generis* normative property and stipulate that the essence of *being a happiness-maximizing act* involves *being right*. In that case, since the essence of *being a happiness-maximizing act* involves a *sui generis* property, it is itself a normative property.

Generalizing, the view is that some acts and states of affairs have a primitive feature of normativity; and it is this primitive feature that privileges them from the point of view of reality.

Now, the worry goes that this, as David Enoch (2021: 1691) writes, commits non-naturalism to conditionals like "if human pain and dog pain have no non-natural property in common (seeing that human pain is intrinsically bad, and that intrinsic badness is on my view a non-natural property), dog pain is not intrinsically bad."

Among others, Melis Erdur (2016), Matt Bedke (2020), Max Hayward (2019) and Shamik Dasgupta (2017) have recently emphasized that the reasons explaining, say, the wrongness of genocide have to do with pain and suffering and not non-natural properties. So, they argue, if non-naturalism is committed to the thought that the wrongness of genocide is ultimately the distribution of some causally inefficacious non-natural properties, that does not seem good for non-naturalism.

Proponents of the view recognize this, and have developed various responses. Those often revolve around various roles *grounding laws* play in the non-naturalist's framework. At this juncture, there is, I believe, an interesting and unnoticed connection between this newer complaint about non-naturalisms first-order moral implications on the one hand and metaphysical worries that have traditionally surrounded the view on the other. For it has also been tried, recently, to meet that second set of concerns by marshalling grounding laws to do certain work. There is reason to think however, as I argue in this paper, that the grounding reply to supervenience is inconsistent with some replies to the normative objection.

I proceed as follows. In section 2, I present the normative objection in more detail. Here I'll also introduce the family of responses to it that I think are incompatible with the grounding reply to the su-

pervenience objection. I take up that traditional metaphysical quibble about supervenience in section 3, where I also outline the grounding response and why it seems very natural one for the non-naturalist to give. But, I show in section 4, because it requires that moral laws play an explanatory role *vis-à-vis* the distribution of moral properties, it imposes constraints on what they can be. In section 5, I argue that these constraints are inconsistent with non-natural ontology playing no role in moral justification on non-naturalism, and thus with some responses to the normative objection.

2. *The normative objection and No Partial Grounds*

Non-naturalists are typically reluctant to accept the metaphorical charge that their *sui generis* properties float around in the ether. This is strongly suggested by their denial that these properties are supernatural properties, though the line between non-natural properties and supernatural properties is notoriously difficult to draw (Väyrynen 2018). Nevertheless, facts about these non-natural properties are the truthmakers of normative beliefs, like astronomical facts are the truthmakers of beliefs about celestial objects. So understood, then, the non-naturalist's claim is that there are correct answers to ethical questions insofar as there are ways of living that are objectively favored by the patterning of these non-natural properties. Indeed, many non-naturalists have recently adopted metaphysicians' talk of being *joint-carving*, or *elite*, and interpret the question of which normative concepts are the *right* ones to use as one of which normative concepts are *joint-carving* (Eklund 2019: 3).

This is thought to give the view certain advantages in accounting for strong moral objectivity in cases of (hypothetical) normative disagreement. Consider how Enoch and McPherson (Enoch and McPherson 2017: §6; Enoch 2011: §5.3; McPherson 2011) put it in terms of reasons and 'schmeasons'. They ask us to consider two linguistic communities: the 'reasoners' and the 'schmeasoners,' both of which have a certain term ('reason' and 'schmeason', respectively) that they take to be central to their normative practices. And in each community, the thought experiment continues, there are sophisticated practices of criticism and evaluation that use the relevant term. The reasoners and schmeasoners, however, have reached quite deviating substantive views in their respective best overall accounts of their common-sense judgments and intuitions. And if we suppose that these practices are coherent, and constitute their own domains, then both communities might be functioning quite well relative to their respective domains. Unfortunately for the schmeasoners, it is bad that they are sensitive to schmeasons rather than reasons. This, unfortunately for the reasoners, seems to be an objection that can be raised perfectly symmetrically from within each of the two domains. For the schmeasoners can urge it is 'schbad' that we respond to reasons rather than schmeasons.

Here non-naturalists suggest that their metaphysically committed realism is the only way to capture what we intuitively want to say, for only the non-naturalist can say that only the reasoners track the normative structure of reality. After all, if there are no mind-independent moral facts, it's not possible to be wrong about these facts either. And then there might be nothing we could tell the schmeasoners about why their ideas about what reasons she has are mistaken. This means that the disagreement has a worrying symmetry. However, this violates the way we normally think about moral disagreement as being asymmetrical. When two people make conflicting normative judgements, at most one of these judgements is correct.

Because of such considerations, on the flip side, it has seemed to many that a very natural reading of non-naturalism is that non-natural facts and properties are higher-level reasons why (cf. Väyrynen 2021) and as such figure in moral justification. Along these lines, for example, Erdur (2016) has argued that metaethical views terminate chains of substantive moral why-questions, and as such must be substantively moral themselves. Once a question is asked about an abstract normative theory, the appropriate next step, according to Erdur, is to ascend to the level of metaethics. Metaethical theories, therefore, may naturally be heard as very general substantive moral claims about why (in the end) right things are right and wrong things are wrong.

Like Erdur, many have interpreted the way non-naturalism locates the source of normativity in a realm of non-natural facts as a commitment to the thought that what ultimately accounts for the wrongness of, say, genocide is some non-natural part of the universe. Because according to that line of thinking, conformity to the facts about the distribution of certain inherently normative non-natural properties constitutes the moral bottom line. So the wrongness of anything is conditional on the distribution of these properties (Erdur 2016: 597).

Hayward (2019) has relatedly argued that non-naturalism in his version of the normative objection makes us counterfactually conditionalize our world-directed moral beliefs on the existence and pattern of non-natural facts. But, says Hayward, it seems misguided to accept this conditional—to accept the moral judgment that you ought to change your moral judgments to match how certain non-natural properties pattern (rather than to match what causes happiness, avoids suffering, etc.). One should not, for instance, change one's mind that pleasure is good and pain bad simply because there is no non-natural property that one has and the other lacks (and vice versa). Indeed, this engenders a strange skepticism on which for all we know, our moral system does everything we want of it—it promotes happiness and minimizes suffering, and so forth—but actually was really false. If a failure to correspond with non-natural moral reality falsifies the moral views of alien ethical cultures, every positive moral view, however central to our culture, would be falsified by the complete and total absence of non-

natural facts. The consequential Parfitian claim that pain, happiness, suffering, and the like, lack value if naturalism is true seems wrong-headed. Our norms of moral evidence legislate that such metaphysical considerations about a non-natural realm could not in principle be relevant to the question of whether I ought to comfort my suffering partner, or whether anything matters.

This objection will seem incoherent to some non-naturalists. They might say: "The non-naturalist's view is that non-natural property NN_1 *is* the property goodness, and that information about non-natural properties $NN_1 \ldots NN_n$ *is* information about morality. It's *obvious* that we should promote goodness and be moral. Hence it's *obvious* that we should promote NN_1 and act according to $NN_1 \ldots NN_n$." Indeed, the non-naturalist will object I'm begging the question—it's only by treating non-naturalist claim as false that the outlined objection is coherent. NN_1, NN_2, and so forth, are *ex hypothesi* normative properties and facts. So, information about their patterning cannot be non-normative information. Rather, NN_1 is, for example, information about an act's to-be-doneness. And it's incoherent to claim that having the non-natural property of, e.g., to-be-doneness settles nothing about an act's to-be-doneness. The non-naturalist's view is that one can't disentangle reasons and non-natural properties like that. When we get information about the latter, we get information about reasons and requirements—not about some kind of *stuff*.

I won't disagree that it's obvious that we should promote goodness. But as Dasgupta (2017: 301) has pointed out, this puts a constraint on what goodness is. Whatever it is, it had better be something we should promote. Consider, by way of analogy, the following toy theory of oxygen: that oxygen is a colorless, odorless, and tasteless gas of which an adult human at rest inhales about two grams per minute. This then puts a constraint on a chemical theory of oxygen: whatever chemical substructure constitutes oxygen, it had better behave as the thingy that living organisms breathe. If someone claimed that oxygen is the element zinc (Zn), we can object that bodies of zinc are, most pressingly, not the thingy that organisms breathe (nor are they colorless gasses). Posit any chemical substructure you like, but don't call any of them 'oxygen' unless you've already shown it's the thing that living organisms breathe. That would not, Dasgupta claims, be playing fair.

Similarly, to play fair, the non-naturalist must *first* establish that we *should promote* any *sui generis* non-natural property before it's fair to call this property 'goodness'. She should not call any alleged feature of reality 'goodness' until she has already shown that she has something you should promote or upon which we should conditionalize moral commitment. She should not simply *assume* that the non-natural properties she claims exist are the ones that we are talking about when we ask the relevant normative questions. It must first be shown that certain non-natural properties are obedience-worthy before they them-

selves are worthy of the name 'morality'. It must first be shown that any non-natural properties bear on what we have reason to do before proposing that truths about the patterning of these properties deserve the title 'normative truths'.

This gives the normative objection its bite: non-naturalism posits natural facts and laws about the distribution of certain non-natural properties as joint grounds of particular moral facts. But there's a strong intuition that moral facts should not be grounded in thus dependent on the patterning of causally inefficacious non-natural properties. We should not leave our "first-order views hostage to a non-natural realm", as Bedke (2022: 13) puts it. There have been many responses to this normative argument against non-naturalism (Blanchard 2020; Horn 2020; Enoch 2021). In this paper, I focus on one of them. I call it: No Partial Grounds.

According to this response, we should *not* see non-natural properties and laws about their patterning as doing any morally justificatory work (see e.g., Chappell 2019). Non-natural properties, it is claimed, are not the entities that *make* acts wrong, nor are they the ultimate explanation of, e.g., the wrongness of genocide.[2] So *ipso facto* general facts about their patterning do not enter into a grounding relationship with particular moral facts. Rather, the view is that facts like 'pain and suffering make genocide wrong' *constitute* the moral reality realists accept: "If there are facts about which actions are right and wrong, and facts about what *makes* those actions right or wrong, and these facts do not constitutively depend on the endorsement of any actual or hypothetical agent, it is *plausibly these facts themselves* which (at least partially) constitute moral reality" (Horn 2020: 347). But these facts about what makes acts wrong don't depend on principles about the patterning of a non-natural realm, as the normative objection has it.

Consider, for illustration, the contrast with a Divine Command Theory. Suppose someone offers a theory according to which the expressions of some creature are obedience worthy. Suppose she further says there's no explanation for why this creature and those utterances of her have that normative role. We would say her account is crucially incomplete, and insist on an explanation. According to No Partial Grounds, the air of incompleteness that surrounds this toy DCT derives from how it makes facts about action-guidingness not metaphysically fundamental, but grounds them in non-normative facts about some creature's will. If that's the structure of your normative theory, you owe people an explanation of why they ought to listen to that particular creature. Non-naturalism, by contrast, has a different structure because it conceives of facts about action-guidingness as metaphysically

[2] Even though, as Horn (2020: 349) admits while defending non-naturalism, "In fairness … [non-naturalists] have sometimes characterized their own views in ways that sound like they are making substantive commitments about what *makes* actions wrong." See, for example, Erdur's (2016: 600) discussion of Shafer-Landau and Enoch.

fundamental, not grounded in anything. Hence there's no explanatory gap to be filled.

The normative objection, the idea is, only takes off because we incorrectly assume that non-natural properties and principles about their distribution are Partial Grounds of moral facts. But this is not so, since the normativity of something like pain *is itself* a non-natural fact but is *fully grounded* in the natural pain-facts. So there's no implausible dependence on the patterning of properties in a non-natural realm.

However, Partial Grounds is exactly what the grounding reply to the supervenience objection presupposes. This makes denying it costly for the non-naturalist.

3. *The grounding role of moral laws*

Assume we ought to give more to combat drought. Why is this so? Well, because of (the natural facts about) the suffering of all those starving to death and their loved ones, and the (natural) fact that giving more will alleviate it, presumably by increasing reliable access to food. Are these natural facts enough for grounding the duty? Well, if natural facts are moral facts' *full* ground, then it seems counterintuitive to say that moral facts are, at the same time, *sui generis*, very different from natural facts. How can they both be fully grounded in natural facts and also be discontinuous with them?

Indeed, it is standard that non-naturalism seems committed to the claim that at least one moral fact is not fully grounded in non-normative, natural facts. Where, intuitively a full ground is enough on its own to ground what it grounds, and a mere partial ground isn't enough on its own to ground what it grounds. Non-naturalists of course agree that atomic normative facts are always somehow grounded in the natural facts, but insist that this connection does not amount to a full metaphysical ground. The challenge for the non-naturalist is to give some positive account of this connection.

A natural idea is that *general laws* play a role in metaphysically grounding particular moral facts. On this view, particular normative facts are metaphysically grounded in the relevant natural facts together with general normative principles connecting the two. What makes it the case that we ought to give humanitarian aid, it is very natural to say, is suffering, *and* that we *ought* to alleviate suffering when we can.

Nothing blocks non-naturalists from holding that particular things' non-normative properties partially explain their normative properties. But for the non-naturalist, such cases must, on this proposal, involve some further moral law that is part of the ultimate explanation in these cases. For example, if Donald is bad because he's a liar, it seems Donald's being a liar explains (in the immediate sense) his being bad. But this is true, for the non-naturalist, only because (say) it is an independent normative fact that being a liar makes one bad. Ultimately, Donald's badness depends not just on his being a liar, but also on that

normative fact.

Gideon Rosen (2017b: 138; cf. Maguire 2015: 194) calls the resulting view *Bridge-Law Non-Naturalism*: "Whenever a particular action *A* possesses a normative property *F*, this fact is grounded in the fact that *A* satisfies some non-normative condition *φ*, together with a general law to the effect that whatever *φ*s is *F*." Particular ethical facts obtain in virtue of more general ethical facts together with pertinent non-ethical facts. For example, the full explanation of why an action was wrong involves two kinds of facts: (i) a particular natural fact—you lied—and (ii) a general connecting grounding fact—for all act acts, if it was a lie, it was wrong in virtue of being a lie. Fundamental normative principles are metaphysically prior to particular normative facts and help ground them.

In this way, grounding explanations have been said to resemble covering-law explanations (Rosen 2017a: 285). This gives us a tripartite, law-based view of grounding explanations as model for moral explanations:

Grounds: particular natural fact(s).
Law: general explanatory grounding law about what grounds what.
Explanandum: particular normative fact.

Several reasons have been noted why non-naturalists should accept a picture like this. Indeed, David Enoch acknowledges there are "theoretical reasons to think that Robust, non-naturalist, Realism needs moral principles to do serious grounding work." And Selim Berker (2019: 913) even contends that "the very tenability of [the non-naturalist's] metanormative view depends on something like [Partial Grounds] being true."

In particular, one important motivation for non-naturalist's ascending to this picture has been that it offers a swift reply to traditional supervenience worries. Moral facts, according to this response to the supervenience challenge, supervene on non-moral facts because moral facts are *made the case* by non-moral facts. The supervenience of the moral on the natural is explained by an 'underlying' grounding relationship in which the natural properties non-causally make some entity have some normative property. Grounding is supposed to deliver exactly the deeper metaphysical explanation that the supervenience challenge asked for.[3] The supervenience of the moral properties on the base properties is explained by the fact that the base properties ground the moral properties. As Ralf Bader (2017: 116) puts it:

> [Positing a grounding relation ensures] that there is dependent-variation of the grounded properties on their grounds. A grounding relation explains

[3] Wielenberg (2014: 33) is one example of a non-naturalist giving this reply. In replying to Railton (2017: §7), Parfit (2017: 106) does it too, although he seems to deliberately avoid the word 'grounding'. See also Bader (2017: §4), Berker (2018: §2), Enoch (2019: 4), Leary (2016), Roberts (2018: §4), and Rosen (2017b, 2020) on the role of grounding in replying to the supervenience objection.

> why that which is dependent, namely the normative, varies with that on which it depends, namely the non-normative. The grounding of normative in non-normative properties implies the supervenience of the former on the latter, thereby allowing us to discharge the explanatory burden that is incurred when positing the supervenience of the normative positing a grounding relation ensures that there is dependent-variation of the grounded properties on their grounds.

One might now ask what, in turn, grounds the laws. Typically, grounding explanations are mediated by essences. That is, in the paradigm cases, whenever *A* grounds *B*, there exists an item (or items) whose nature ensures that every *A*-like fact grounds a corresponding *B*-like fact (cf. Litland 2015).

However, the non-naturalist's key thought is that the essences of the normative properties do not in general fix the true general principles on which they figure, some of which are thus genuine synthetic laws about which metaphysicians who know the essences of moral properties can disagree (Rosen 2017b: 146). In other words, non-naturalism holds that the essences of normative properties do not in general fix non-normative necessary and sufficient conditions for their instantiation. That would entail that the natures of the normative properties and relations, collectively or taken one at a time determine naturalistic necessary and sufficient conditions for their application. But that would make them natural properties (Rosen 2017a: 291). To claim it is in the nature of the normative that some non-normative facts ground some normative facts is a distinctly naturalistic claim (Rosen 2017b: 291). Ethical non-naturalism, by contrast, is the view that in at least one case, the essences of the normative properties fail to determine naturalistic necessary and sufficient conditions for their application. This is why, as Rosen (Forthcoming: 12, my emphasis) writes, non-naturalism needs its principles:

> The naturalist's key thought, it seems to me, is not that each normative property is separately definable in non-normative terms. It is rather that the normative facts are fixed by the wholly non-normative facts (e.g., facts of physics and psychology) together with the natures of the normative properties and relations. On this sort of view, anyone who knows the non-normative facts is in a position to derive the ethical facts provided she also knows what it is for an act to be right, good, rational, etc. The non-naturalist's distinctive commitment is that someone who knew the natural facts and the essences might still be in the dark about the *synthetic principles* that connect the normative facts to their non-normative grounds.

In other words, non-naturalism holds that the essences of normative properties *do not* in general fix non-normative necessary and sufficient conditions for their instantiation.

On this picture, particular ethical facts obtain in virtue of more general ethical facts together with pertinent non-ethical facts. And as we ask what grounds those 'synthetic principles', general normative laws will figure at every step. The regress could conceivably be infinite. But

more likely is that it will terminate in fundamental laws: the supreme principles of normativity. On the non-naturalist picture, there is thus an elite set of metaphysically necessary true moral laws that are the ungrounded normative facts upon which all the other normative facts rest.

Appealing to grounding in order to explain the distribution of moral properties, then, is incomplete without an account of the "synthetic principles" that conspire with the underlying natural facts to ground the particular moral facts and explain moral supervenience.

Specifically, the non-naturalist needs to *show* that laws are able to play the metaphysical grounding role given to them. On her account, moral principles are themselves part of what explains why individual actions have the moral properties that they do. But, as I will explain shortly, not everything that they could be like would be able to do this. This means that the viability of the grounding response to explanatory worries surrounding non-naturalism depends on an account in which moral laws *can* and *do* play a determining role *vis-à-vis* the distribution of moral properties. Which is inconsistent with the No Partial Grounds reply to the normative objection.

4. *What non-natural moral laws must be like to ground*

How can Principles as Partial Grounds have the far-reaching consequences Berker talks about? As we saw, a central commitment of non-naturalism is that there are "synthetic principles" connecting the natural to the moral: there are true normative principles as ungrounded normative facts, upon which all other normative facts rest. Such (fundamental) normative principles are metaphysically prior to particular normative facts, which they help to ground. Yielding a picture on which these (non-natural) laws don't play a role in making moral facts the case, but on which moral facts are fully grounded in natural facts instead, seems at odds with non-naturalism. This is because if the full ground of moral facts includes only natural facts, moral facts no longer seem to have their own radically different, *sui generis*, non-natural metaphysical category. So if moral laws are not partial grounds, there seems to be no reason to believe that moral facts are not natural facts.

To support my claim that not everything they can be like allows them to play this role, I start by giving two examples to show why not everything that moral laws could be like would allow them to play the role the non-naturalist needs them to play.

Since this required role is *explanatory*, the principles cannot, firstly, be mere regularities. The mere fact that all *A*s are *B*s cannot *explain* the fact that a given *A* is *B*. Rather, they have to be proper *laws*: general facts that account for their instances and are not explained by them. For a general connecting principle—between, e.g., suffering-facts and duty-facts—to figure in the grounds is for it to *govern its instances*. And for a principle to govern its instances is to be part of what *makes*

any instance obtain. And the only ones that can on pain of circularity govern their instances are ones that are not plausibly grounded in their instances.

To see the point about circularity, suppose the general fact that if something A-like obtains, so too does something B-like is made true, at least in part, by its instances—by that A-like thing and B-like thing, and that one, and so forth. But then if the general principle is included in the grounds of ordinary grounded facts, each instance is also partly grounded in the generalization. Each instance is partly grounded in the generalization, *and* the generalization is partly grounded in each instance. This violates the asymmetry of grounding.

So, moral laws cannot be mere regularities because they have to be prior to their instances in the metaphysical grounding order. A second thing they cannot be is mere, as it were, epistemic scaffoldings. On this view, it is not moral truth itself, but our epistemic capacities and limitations that necessitate postulating moral laws. Sean McKeever and Michael Ridge, for example, can be understood as having a view along these lines. They defend moral generalism as a prescriptive thesis, arguing that principles are guides in moral thought and discourse, and that the prominent role these guides play in our practices is what necessitates our commitment to them (McKeever and Ridge 2006: 177–8). This point about the epistemic or practical *need* for general principles, however, is not enough for the non-naturalist purposes. She must also show that the principles *actually determine* the moral facts. What is required is an account of how laws manage to *have* this explanatory power. Arguing that moral laws are required for an *epistemically satisfying* story about why, for example, suffering-facts ground duty-facts, does not suffice for defending their role in a *metaphysically complete* story about this grounding relationship.[4]

Why not? The problem is that, for Principles as Partial Grounds to be true, moral laws can't be mere descriptions of metaphysical dependence relations. On such a view, true moral principles track the natural-moral metaphysical dependence relations obtaining 'out there'. For example, the statement "I promised to F" explains the statement "I am obligated to F" in virtue of a metaphysical dependence relation that exist between obligations and promises. This view of moral laws, however, would have laws be describers of metaphysical explanation, rath-

[4] A possible reply is that, maybe particular ethical facts can be fundamental in the metaphysical grounding order. For example, 'the pain and suffering of *this* genocide makes it wrong' would be an example of a fundamental non-natural fact that constitutes moral reality. An argument against that view is that it's at odds with a central feature of ethical practice: we normally think that moral explanation presuppose general principles. We can refute a moral explanation of the form 'it's wrong to push the fat man because doing so is φ,' by citing a merely possible counterexample to the implied general law: whatever φs is wrong. This shows that the moral law implicit in the explanation is not a mere regularity, but rather a modalized generalization of some sort.

er than themselves explanatory (in the right kind of way, as I outline below). This is to play an epistemic role only, namely to direct towards the underlying metaphysics, without being part of the metaphysics—without helping to make the connection between promises and obligations obtain (Kim 1994: 67–8). Each instance of, e.g., wrongness is then fully explained by a particular natural fact. This amounts to saying that the full ground of moral facts are natural facts. This amounts to saying that the full ground of moral facts are natural facts—which is not non-naturalism.

On that view, the law needs to be an additional, more fundamental entity in the explanation that explains—is responsible for—any emergent regularity between, e.g., promises and obligations. This corresponds to the non-naturalist thought that there exists a metaphysically robust moral realm, conformity to which is the ultimate standard for right and wrong (cf. Erdur 2016: 598).

To recapitulate, moral laws can't be grounded in the particular moral facts they subsume, since to the contrary, those particular facts are partly grounded in the laws. Moreover, we aren't after *epistemic justification* for a belief that a particular moral fact obtains given that a particular natural fact obtains. On that role, moral generalizations only license certain natural-moral inferences, but their explanatory power is derivative from the metaphysical dependence relations they *depict* rather than *make the case*.

Instead, we want to know what *underwrites* such inferences licensed by the generalizations. For this, we need the sort of explanation that gives an account for why things are the way they are.

Now, what must moral principles be like for them to do this work? What must moral laws be like such that the grounding role it assigns to moral principles as an additional entity in the explanation of particular moral facts can be vindicated?

First desideratum: in order for these laws to play a role that's metaphysically explanatory, they must *play a* (non-causal) *determining role* regarding the distribution of moral properties (the analogy would be non-Humeanism about the laws of nature where they play a determining role in making events come about). If moral laws are to do grounding *work*, they need to be partly *responsible for* the moral facts they help to ground. That's just *what it is* to have a metaphysically explanatory role. Principles must not only "explain what is true in particular cases without determining it," they must "determine what is true and explain it" (Dancy 1983: 533).

So, one thing the non-naturalist's account of moral laws must accommodate is that they must be responsible for particular moral facts. They must make the facts obtain. Before unpacking other desiderata of the non-naturalist account of moral laws, I want to point out an interesting implication of this.

As a rule, that which is grounded is ontologically dependent on

its grounds. On the view we're considering, moral laws are needed to ground particular moral facts. It follows these facts are ontologically dependent on moral laws, and would not obtain without the law obtaining. For example, without a general fact according to which suffering is bad, the relation between particular facts about suffering and particular facts about badness would not obtain. The facts about suffering *and* the moral law are *both* required to fully account for the facts about badness. Now, without *any* moral law, no moral fact would obtain. There being wrongness at all, thus depends (in part) on there being moral laws. For all its apparent boldness, it's hard to see how there could be an account on which moral laws do metaphysical grounding work that does not have this implication (on the laws of nature analogy, two objects attract each other with the force they do in part because of the masses they have and the distance between them, and in part *because* of the law of universal gravitation).

What other positive desiderata does the non-naturalist account of moral laws need to meet? Well, since they enter intro grounding relations, they must be facts. And since they have moral content, they are *moral* facts. Now, according to the non-naturalist, moral facts are mind-independent. That is, they are facts about the world. Thus, the claim that giving to charity is good represents the world as being a certain way, and if that claim is true, that is in virtue of a certain kind of worldly fact: that giving to charity is good. Similarly, if it is true that giving to charity alleviates suffering, this is so in virtue of some other worldly fact. Now consider a moral principle: giving to charity is good *because* it alleviates suffering. This seems to be true as well. But if we accept that giving to charity alleviates suffering and giving to charity is good are both worldly facts, to say that giving to charity is good because it alleviates suffering is to say that one worldly fact obtains because another worldly fact obtains. Because this 'because' relation holds between two worldly facts, this 'because' relation seems like it must, itself, be worldly (yet non-natural). And the same for other moral laws. For the non-naturalist, moral laws are mind-independent aspects of the world, the truthmakers of claims where a moral property is supposed to obtain *because* some natural property obtains.

The final desideratum is that moral laws need to supply a necessary connection between distinct existences. Since that is how the non-naturalist conceives of the natural and the moral. To meet this, the non-naturalist simply asserts there *can* be necessary relations between distinct existences, at least when the distinct existences are normative on one side, and natural on the other (Enoch 2011: 147). The moral laws, to be understood as extra, *sui generis* facts about the world, 'hook them up'. The non-naturalist does not have an answer to how this could be, but denies she has to give one. The non-naturalist is indeed committed to something brute, but the bruteness, she claims, is exactly where it's supposed to be. So it's not (really) costly. After all, one might think,

something has to be fundamental, and necessary laws seem like good candidates. They are metaphysically basic, where reality starts out, by definition having no full metaphysical explanation. Fundamental laws governing the natural-normative grounding relation are a metaphysical fundamental explainer, on a par with, e.g., the constitution relation. Simply not the sorts of things that can, in principle, have a metaphysical explanation. Rock-bottom grounding relations are explainers, not things that need to be explained.

This ends our search for an answer to the question what moral laws must be like for Principles as Partial Grounds to be vindicated. They must be *sui generis* worldly facts about what grounds what, the most fundamental of which are ontologically basic. Next to being facts about what grounds what, they must be *responsible* for the particular moral fact they help to ground.[5]

5. *Why this picture of moral laws is inconsistent with denying that* sui generis *ontology plays a role in moral justification*

With all this in place, it's not hard to see why Partial Grounds entails that non-naturalism has *sui generis* ontology play a role in moral justification. It follows from that picture that moral laws not just describe particular moral facts, but *make* them the case. And on non-naturalism, moral laws *just are sui generis* items that occur in one's ontology. So it falls straight out of Principles as Partial Grounds and the definition of non-naturalism that non-natural ontology plays such a role in moral justification.

One might object that grounding, an explanatory notion, might not obviously be related to justification. But as Wedgwood (2017: 91) writes, "explanatory characterizations" of normative reasons "associate reasons with a justificatory story—that is, with a story that explains the truth about which action or attitude one has, all things considered, most reason to do". Elstein (ms) similarly suggests that normative explanation "coincides" with justification. Normative explanations are (perhaps among other things) justifications: at least some explanatory reasons why a normative fact holds must provide normative reasons for certain responses, or be features in the light of which those responses are apt or fitting or the like.[6] Normative explanations are explanations of why things have the normative features they to do: they aim to explain why things have properties such as rightness and wrongness. Most of us are inclined to think that such facts in a way involve reasons: considerations that justify actions. If so, then we would want normative explanations, too, to cite such considerations, and to be in-

[5] See Berker (2019) for an argument that this is an incoherent combination. See Enoch (2019) and Fogal and Risberg (2020) for replies.

[6] See Väyrynen (2019) for an argument for this claim.

complete otherwise (Väyrynen 2015: 173).

The argument of this paper has proceeded in two relatively simple steps: non-naturalism needs moral laws to play a role in grounding making moral facts the case. And on non-naturalism, moral laws are *sui generis* pieces of ontology. Therefore, non-naturalism is committed to *sui generis* pieces of ontology playing a role in justification. I hope to have said enough about why the first premise is true, and why 'making the case' (grounding) is linked to justification. In closing, I want to respond to an objection to the second premise. Can the non-naturalist deny that her notion of moral facts, of which moral laws are a subset, comes with ontological commitment?

On pain of becoming a version of quietist normative realism, it seems not. The Robust Realist claims moral facts exist in the same sense as chemical facts, physical facts, and all the rest. She explains what moral judgments are about, and explains their truth conditions, by postulating non-natural moral properties. But if one claims that moral facts exist in the same sense as physical facts and are as 'ontologically respectable' as them, this reply is not an option. For then what could it mean when they claim that "in whatever sense there are physical facts, there are normative ones; in whatever sense there are truths in biology, there are in normative discourse" (Enoch 2011: 5)?

In support of this interpretation, consider how FitzPatrick (2018: 555) explains his motivations for adopting non-naturalism: "We are skeptical about capturing everything we want without relying on some irreducibly evaluative or normative facts about standards or good-or right-makingness; so we posit such apparently 'non-natural' facts and properties at the bottom of all this." What could "at the bottom of all this" mean, if not the bottom of a *chain of justifications*? On such a view, non-natural properties are the truthmakers for the normative truths about which natural properties are normatively significant in which ways (Chappell 2019: 131). So when two natural properties differ in normative valence, this is ultimately reflected in them having different non-natural properties. But if non-natural properties are where moral justifications hit bottom, it seems misplaced to say deny that non-natural laws are moral grounds.

Similarly, Richard Chappell (2019: 125) clarifies that (according to him), "the role of non-natural properties is not to *be* responded to, but to 'mark' which natural properties it is *correct* for us to respond to in certain ways." This is consistent with the mentioned appeals to eliteness—the thought that differences in alignment with non-natural properties can settle moral disagreements between communities. But if non-natural properties make it morally correct or incorrect to care about certain things and not others, it's very hard not to see them as higher-level reasons why—as a reason why we should care about things like happiness and love (they share a non-natural property) and not about handclapping and blade-counting (they do not).

I conclude that non-naturalism must insist that an act was wrong because it caused suffering, *and* that *not only* because of that, but *also* because there is a non-natural, *sui generis*, extra fact about the world which makes it true that suffering is bad. Common replies that non-naturalism occurs no such commitment are belied by the view's grounding structure. For non-naturalism needs moral laws—which account for the patterns of distribution of non-natural properties—to pull their weight in doing metaphysical grounding work. It won't fly then, to, when responding to the normative objection, claim that actually these laws are explanatory idle and all we need is the natural facts. Perhaps (as David Enoch has suggested) grounding pluralism can be of help, but, in responding to the objection that non-naturalists are "leaving their first-order views hostage to a non-natural realm" (Bedke 2022: 13), they cannot, as some have wished, do with just the natural facts.

References

Bedke, M. S. 2020. "A Dilemma for Non-Naturalists: Irrationality or Immorality?" *Philosophical Studies* 177 (4): 1027–1042.

———. 2022. "Kowtowing to a Non-Natural Realm." *Journal of Moral Philosophy* 19 (6): 559–576.

Berker, S. 2019. "The Explanatory Ambitions of Moral Principles." *Noûs* 53 (4): 904–936.

Blanchard, J. 2019. "Melis Erdur's Moral Argument Against Moral Realism." *Ethical Theory and Moral Practice* 22 (2): 371–377.

Chappell, R. Y. 2019. "Why Care about Non-Natural Reasons?" *American Philosophical Quarterly* 56 (2): 125–134.

Dasgupta, S. 2017. "XV—Normative Non-Naturalism and the Problem of Authority." *Proceedings of the Aristotelian Society* 117: 297–319.

Enoch, D. 2011. *Taking Morality Seriously*. Oxford: Oxford University Press.

———. 2019. "How Principles Ground." In R. Shafer-Landau (ed.). *Oxford Studies in Metaethics Volume 19*. Oxford: Oxford University Press, 1–22.

———. 2021. "Thanks, We're Good: Why Moral Realism Is Not Morally Objectionable." *Philosophical Studies* 178 (5): 1689–1699.

Enoch, D. and McPherson, T. 2017. "What Do You Mean 'This Isn't the Question'?" *Canadian Journal of Philosophy* 47 (6): 820–840.

Eklund, M. 2020. "Reply to Critics." *Inquiry* 63 (5): 1–27.

Elstein, D. (ms). "Normative Regress and Normative Relevance". Unpublished.

Erdur, M. 2016. "A Moral Argument Against Moral Realism." *Ethical Theory and Moral Practice* 19 (3): 591–602.

FitzPatrick, W. 2018. "Representing Ethical Reality: A Guide for Worldly Non-Naturalists." *Canadian Journal of Philosophy* 48 (3–4): 548–568.

Fogal, D. and Risberg, O. "The Metaphysics of Moral Explanations." In R. Shafer-Landau (ed.). *Oxford Studies in Metaethics, Volume 15*. Oxford: Oxford University Press, 170–194.

Hayward, M. K. 2019. "Immoral Realism." *Philosophical Studies* 176: 897–914.

Horn, J. 2020. "On Moral Objections to Moral Realism." *The Journal of Value Inquiry* 54 (2): 345–354.

Kim, J. 1994. "Explanatory Knowledge and Metaphysical Dependence." *Philosophical Issues* 5: 51–69.

Leary, S. 2016. "Non-Naturalism and Normative Necessities." In R. Shafer-Landau (ed.). *Oxford Studies in Metaethics Volume 12*. Oxford: Oxford University Press, 76–105.

Litland, J. 2015. "Grounding, Explanation and the Limit of Internality." *Philosophical Review* 124 (4): 481–532.

Mackie, J. 1977. *Ethics: Inventing Right and Wrong*. Harmondsworth: Penguin Books.

Maguire, B. 2015. "Grounding the Autonomy of Ethics." In R. Shafer-Landau (ed.). *Oxford Studies in Metaethics Volume 10*. Oxford: Oxford University Press, 188–215.

McKeever, S. and Ridge M. 2006. *Principled Ethics: Generalism as a Regulative Ideal*. Oxford: Oxford University Press.

McPherson, T. 2011. "Against Queitist Normative Realism." *Philosophical Studies* 154: 223–240.

Parfit, D. 2011. *On What Matters Volume 2*: *The Berkeley Tanner Lectures*. Oxford: Oxford University Press.

———. 2017. *On What Matters Volume 3: The Berkeley Tanner Lectures*. Oxford: Oxford University Press.

Rosen, G. 2017a. "Ground by Law." *Philosophical Issues* 27: 279–301.

———. 2017b. "What Is a Moral Law?" In R. Shafer-Landau (ed.). *Oxford Studies in Metaethics Volume 12*. Oxford: Oxford University Press, 135–159.

———. Forthcoming. "What is Normative Necessity?" In M. Dumitru (ed.). *Metaphysics, Meaning and Modality: Themes from Kit Fine*. Oxford: Oxford University Press: https://www.academia.edu/9159728/Normative_Necessity

Väyrynen, P. 2015. "I—Grounding and Normative Explanation." *Aristotelian Society Supplementary Volume* 87 (1): 155–178.

———. 2018. "Normative Commitments in Metanormative Theory." In J. Suikkanen and A. Kauppinnen (eds.). *Methodology and Moral Philosophy*. Routledge, 193–219.

———. 2019. "Reasons Why in Normative Explanation." *Inquiry* 62 (6): 607–623.

———. 2021. "Normative Explanation and Justification." *Noûs* 55 (1): 3–22.

Wedgwood, R. 2007. *The Nature of Normativity*. Oxford: Clarendon Press.

Croatian Journal of Philosophy
Vol. XXIII, No. 68, 2023
https://doi.org/10.52685/cjp.23.68.5
Received: November 24, 2022
Accepted: July 14, 2023

Unwanted Arbitrariness

STIJN BRUERS
KU Leuven, Leuven, Belgium

I propose a new fundamental principle in ethics: everyone who makes a choice has to avoid unwanted arbitrariness as much as possible. Unwanted arbitrariness is defined as making a choice without following a rule, whereby the consequences of that choice cannot be consistently wanted by at least one person. Other formulations of this anti-arbitrariness principle are given and compared with very similar contractualist principles formulated by Kant, Rawls, Scanlon and Parfit. The structure of arbitrariness allows us to find ways to avoid unwanted arbitrariness. The two most important implications of the anti-arbitrariness principle are discussed: non-dictatorship and non-discrimination.

Keywords: Ethical principles; Kantianism; contractualism; categorical imperative; dictatorship; discrimination.

1. *Introduction*

From Kant (1785) and Bentham (1789) to Scanlon (1998) and Parfit (2011), philosophers have a long tradition of searching for the most fundamental ethical principles. This article fits in that tradition, by defining a new core concept in ethics: unwanted arbitrariness. With this concept, the anti-arbitrariness principle states that everyone who makes a choice has to avoid unwanted arbitrariness as much as possible. This is a new proposal of a fundamental principle in ethics. It is fundamental in three senses. First, the principle offers a necessary (but not necessarily sufficient) condition for an act to be right or a choice to be moral. In other words, a violation of the anti-arbitrariness principle is a sufficient condition for an act to be wrong or immoral. Second, the principle applies to all choices, including for example the choices of moral rules and moral theories. Hence, it is a meta-principle: a principle about principles. Third, unwanted arbitrariness refers to a lack

of moral justifications, valid reasons or acceptable rules. When such reasons or rules are absent, we basically leave the realm of morality. The concept of unwanted arbitrariness is so crucial, that it can be said to demarcate morality, to distinguish the moral and immoral from the amoral. This reason-based or rule-based approach to ethics fits in Kantian and contractualist traditions of ethics (Kant 1785; Scanlon 1998; Rawls 2005; Parfit 2011). As such fundamental principles do not tell you what to do or what is moral (i.e. do not make substantive moral claims or judgments), but rather give you a procedure or method to determine what to do, what is right or what is good, such fundamental principles are useful in the field of procedural or formal justice. Instead of offering substantive claims how to solve each case, the principle entails, for example, that one should treat like cases alike. Instead of offering which specific rights a person has, the principle imposes the condition that everyone should have equal rights.

The search for the fundamental principle(s) in ethics is a very ambitious project. It requires giving precise formulations, offering justifications, discussing implications, presenting applications and making comparisons with other proposals of fundamental principles in the moral philosophy literature. That requires a whole book. The main focus of this article is the first step: formulating the anti-arbitrariness principle as precise and unambiguous as possible. The justifications, implications, applications and relations with the existing literature will be briefly sketched, but are mainly left for future work. That means those issues are not yet fully developed in this article. Similarly, whether this anti-arbitrariness principle is a mere reinterpretation or reformulation of Kantian and contractualist theories, or contains substantial differences with such theories, will also be left for future research. Even if it is a mere reformulation of a fundamental ethical principle already proposed in the literature, it could help clarify that proposed principle and more clearly enable us to see certain implications of it.

In this article, I will define the concepts of unwantedness and arbitrariness, give several formulations of the anti-arbitrariness principle, briefly compare them with very similar fundamental ethical principles formulated by Kant (1785), Scanlon (1998), Rawls (2005) and Parfit (2011), use the structure of arbitrariness to look for options how to avoid unwanted arbitrariness, and discuss the two most important consequences of the anti-arbitrariness principle: non-dictatorship and non-discrimination. The unwantedness of a violation of the anti-arbitrariness principle gives us a reason why dictatorship and discrimination are morally wrong.

2. *Definitions of unwantedness and arbitrariness*

Unwantedness for an individual means being incompatible with that individual's largest consistent set of strongest subjective preferences. An individual is a being who has subjective preferences. A subjective

preference is a conscious value judgment or evaluation that has a subjective strength (to be distinguished from, e.g. a mere unconscious behavioral disposition). For example, being told a lie is incompatible with a preference for knowing the truth. Two preferences are mutually inconsistent when it is unfeasible or logically impossible to satisfy them both. Consider a reluctant drug addict, who is torn between two preferences: wanting the drugs and wanting to be clean or healthy.[1] This inconsistency in preferences is what makes such drug addiction problematic. When the drug addict values being clean more than the excitement from taking an extra dose of drugs, but still takes the drugs due to the addiction, that behavior can be said to be irrational.

To respect autonomy, an individual can freely choose the method to construct their own strongest consistent set of preferences. One method to construct your individual consistent set is: consider the list of everything you want, ranked according to personal value or strength, with the strongest preferences at the top. Move down the list and delete the items on the list that are incompatible with the higher non-deleted items. The remaining items form a consistent set of your strongest preferences. Everything that is not logically compatible with this remaining list of your strongest preferences is unwanted and cannot be consistently wanted by you. Everything that is compatible can be consistently wanted. Saying that you cannot consistently want something can be interpreted as being equivalent to saying that you can reasonably object against it.

Arbitrariness means selecting an element (or subset) of a set without using a selection rule. A selection rule is a rule that logically determines the selection. It is an if-then statement that consists of a set of conditions with logical operators (conjunctions, disjunctions, negations). For example: "If element X has conditions A and B or not C, then select X." If the question "Why selecting element X instead of element Y?" has no answer that refers to a selection rule (for example if the only answer is "Therefore!"), then selecting X is arbitrary.

Combining the above definitions of unwantedness and arbitrariness, we can define unwanted arbitrariness as making a choice without following a selection rule, whereby the consequences of that choice are unwanted (i.e. cannot be consistently wanted) by at least one person. Here, a choice can be defined as a conscious decision. Making a choice means consciously selecting an element from a choice set, the set of eligible options. These eligible options can be feasible actions but also for example preferences, allocations, ideas, moral theories or ethical principles.

[1] I thank an anonymous reviewer for this example.

3. *Formulations of the anti-arbitrariness principle*

The anti-arbitrariness principle states that:

When making a choice, we have to avoid unwanted arbitrariness as much as possible.

To avoid arbitrary exclusion of choices, this principle applies to all possible choices, including very specific actions ("Sit at seat 5 on bus 42 at 1 pm Friday"), to more general choices ("Use public transport"), to justifications ("Take a seat because the seat is empty and you paid for a ticket"), to higher level moral choices ("Choose the action allowed by a contractualist ethic"), to moral theories ("Choose the theory of act-utilitarianism"), to even very basic choices of premises and logical deduction rules used in justifications ("Use deontic logic to determine the validity of an argument"). For practical reasons, we do not have to consider impossible choices ("Avoid unavoidable unwanted arbitrariness").

This anti-arbitrariness principle does not yet say what happens if we don't avoid unwanted arbitrariness. Also, the "as much as possible" hints at the possibility that sometimes unwanted arbitrariness may not be avoidable. Therefore, we can give a more exact formulation of the anti-arbitrariness principle, in a strong and a weak version.

Anti-arbitrariness principle, universal formulation, strong version:

> If you do not avoid avoidable unwanted arbitrariness when making a choice, you are not allowed to make that choice.

The weak version can be derived from this strong version. Suppose unwanted arbitrariness is unavoidable. You have to make a choice that involves unwanted arbitrariness. What about other people making other choices? Are you allowed to determine the choices of others, to impose your choice on them? Are you allowed to choose who may make the choice? Choosing yourself as the dictator who dictates the choices of others, would involve unwanted arbitrariness again. To avoid this new unwanted arbitrariness, you are not allowed to be the dictator. You have to accept the choices made by other people.

Anti-arbitrariness principle, universal formulation, weak version:

> If you cannot avoid unwanted arbitrariness when making a choice, you are allowed to make that choice but other people may make other choices from the same choice set (i.e. you have to tolerate that other people make other choices).

The above formulations are universal, in the sense that everyone and everything must abide by this principle. No arbitrary exceptions are allowed. The principle applies to everyone and everything that is able to make choices based on selection rules. It also applies, for example, to artificial intelligent machines. Of course, when someone cannot make a choice, that is an exception, but not an arbitrary exception because it is justified using an "ought implies can" rule: "If you cannot do something, you have no obligation to do it."

We can give another, personal formulation of the anti-arbitrariness principle:

> For every choice you make, you have to be able to give a justification rule such that you and everyone can consistently want that everyone follows that rule in all possible (including hypothetical) situations (i.e. you and everyone can accept the consequences of a universal compliance by everyone of the justification rule).

This is a personal formulation, because it refers to what you can want. Hence, this formulation applies to everyone who is not only able to make choices, but also able to want something, i.e. someone with personal preferences.

Whereas the first, universal formulation referred to selection rules, this second, personal formulation refers to justification rules. A justification rule is a selection rule that is used in moral reasoning, to justify to other people one's choices. Therefore, a justification rule for (im)permissibility of a choice should be used in a logical deduction. That means a justification rule is basically an if-then statement that consists of a set of conditions: "If conditions C apply, then it is permissible to choose X."

The above second formulation does not yet say what to do when you are not able to formulate a justification rule. Therefore, as with the first, universal formulation, we have to make this second, personal formulation of the anti-arbitrariness principle more precise. And as with the universal formulation, this personal formulation also comes in two versions, of which the weak one can be derived from the strong version.

Anti-arbitrariness principle, personal formulation, strong version:

> If, when making a choice, you cannot give a justification rule of which you would accept universal compliance, then you are not allowed to make that choice nor follow that rule.

Anti-arbitrariness principle, personal formulation, weak version:

> If, when making a choice, you cannot give a justification rule of which everyone would accept universal compliance, then you must accept or tolerate that other people make other choices from the same choice set and follow other justification rules for making those choices.

There are many similarities between the universal and personal formulations of the anti-arbitrariness principle, such that they can be said to be roughly equivalent.

First, there is a correspondence between the selection rule and the justification rule. The universal formulation works with a selection rule to avoid arbitrariness. In the personal formulation, arbitrariness is avoided by the justification rule and by the idea that if you may follow that rule in a specific situation, then everyone may follow that rule in all possible situations. Suppose that the "everyone" and "all possible situations" were no requirements. Replacing them by "some people"

and "some situations" would introduce arbitrariness, because arbitrary subsets of the sets of all people and all situations can be chosen.

Second, both formulations look for what can be consistently wanted. The condition "everyone can consistently want that everyone follows that rule in all possible situations" is the opposite of unwanted arbitrariness. Suppose you choose option A arbitrarily and person Y is in a position P in which s/he cannot consistently want that arbitrary choice. If we consider everyone and all possible situations, this includes the situation where person Y chooses A and you are in the same position P that Y had, in which case you cannot consistently want A.

A third similarity between the two formulations, is that they both come in a weak and a strong version. Unwanted arbitrariness may not always be avoidable, because there may always be someone who cannot consistently want a choice that cannot be based on a selection rule. Consider for example the choice of moral theory. There are many, equally consistent theories. Choosing one theory, such as act-utilitarianism, would be arbitrary. And some people may not like that theory. Similarly, it may not be possible to find a justification rule of which everyone can accept universal compliance. The condition that everyone follows the rule in all hypothetical situations, may be too demanding. In these cases, people must tolerate that other people make other choices, for example choose another consistent moral theory (unless for example the act-utilitarians can argue that their chosen theory is not arbitrarily chosen, but chosen by a selection rule).

A final similarity is that both formulations apply to all possible choices, including the choice of selection and justification rules (in particular the choice of conditions in those rules). That means a selection meta-rule should be given to select the selection rule from the set of all selection rules. Similarly, a justification meta-rule should be given to that justifies the chosen conditions in a justification rule. With the application to all possible choices and the resulting necessary inclusion of such meta-rules (and higher order meta-meta-rules), the anti-arbitrariness principle becomes perhaps the most fundamental principle in ethics.

An example might give some clarification. Consider the situation of taking a seat on the bus. If you choose to take a seat, the rule could be: "If you are white, you may take the seat," or "If you have permission by person X, you may take the seat." But the choice of these conditions is arbitrary (they refer to one skin color or person arbitrarily chosen from the sets of skin colors and people). A better rule would be: "If the seat is empty and you have permission by the people who have a special relationship with the seat, you may take the seat." We have to specify what counts as a special relationship. This can again be done by considering relationships of which everyone can consistently want that they are part of the conditions in the justification rule. Examples of such a special relationship could be "being the owner of the bus" or "having

reserved the seat". Having permission could mean "having paid for a ticket" (or generally: "abiding by a system of property rights that does not privilege one person over others").[2]

4. *Connections with other fundamental ethical principles*

The anti-arbitrariness principle is related to other fundamental ethical principles proposed by, e.g. Kant (1785), Scanlon (1998), Rawls (2005) and Parfit (2011). These principles are fundamental, in the sense that they are meta-principles that refer to ethical principles or rules to guide our actions. This section briefly compares the anti-arbitrariness principle with some other proposed principles. Whether the anti-arbitrariness principle is a mere reformulation or contains substantial differences with the other proposals in the literature, is left for future research.

Kant's first formulation of his famous categorical imperative (unconditional obligation), reads: "Act only according to that maxim whereby you can, at the same time, will that it should become a universal law" (Kant 1785). A maxim is a subjective principle of action, i.e. what the agent believes to be the reason for his or her action. A maxim consists of the act (e.g. "lying") and the motivation (e.g. "for a benefit"). When you do an action, find your maxim and imagine a world where everyone (who is able and is in a similar position as you are) follows that maxim. Only if everyone can follow that maxim without contradictions and you can rationally will that everyone follows that maxim, you are allowed to do that action.

This universalizability formulation of the categorical imperative implies for example that when making an action, you cannot make an exception for yourself. You cannot say that you are the only one who may follow your maxim. A universal law does not allow for arbitrary exceptions. This reflects an avoidance of unwanted arbitrariness. We end up with the anti-arbitrariness principle if an act is interpreted more generally as a choice (such that the choice for inaction or allowing something to happen are also considered), a maxim is interpreted as a justification rule and "rationally willing a universal law" is interpreted as "consistently wanting or accepting a universal compliance of the justification rule".

One important difference between Kant's principle and the anti-arbitrariness principle, is that Kant in a sense only considers the most general maxims. Kant claimed that lying is always wrong because a contradiction or irrationality occurs when everyone lies for a benefit. Although Kant not explicitly derived his position of impermissibility of lying from his categorical imperative, such a derivation is only possible by considering only a general maxim such as "lying for a benefit" (for a similar discussion of this point, see e.g. Carson 2010). The anti-

[2] I thank an anonymous reviewer for this point.

arbitrariness principle, in contrast, considers more maxims or justification rules. As a consequence this anti-arbitrariness principle allows for lying in some situations, for example in order to save a life (e.g. when a murder asks you the hiding place of his target victim). I can consistently want that everyone follows the justification rule "if the lie saves the life of an innocent person and has no serious negative side-effects, then you may lie." Kant, if he were to derive his anti-lying conclusion, would only consider a justification rule "if the lie has a benefit, then you may lie", and I cannot consistently want universal compliance of this rule.

Scanlon formulated a contractualist principle of wrongness: "An act is wrong if and only if any principle that permitted it would be one that could reasonably be rejected by people moved to find principles for the general regulation of behavior that others, similarly motivated, could not reasonably reject" (Scanlon 1998: 4). This can also be turned into the anti-arbitrariness principle, when "a principle that permitted the act" is interpreted as the justification rule for a choice, "could reasonably be rejected" is interpreted as "cannot be consistently wanted when universally complied", and "by people" means "by at least one person". Scanlon's theory is reason-based, where a reason must be one "no one could reasonably reject as a basis for informed, unforced, general agreement" (1998: 153). The general agreement contains an anti-arbitrariness condition: an agreement by everyone is required, without arbitrary exceptions.

One important difference between Scanlon's principle and the anti-arbitrariness principle, is that Scanlon only considers a restricted group of people that could reasonably reject a principle, namely those people who are moved to find principles. This reflects a contractualist position, as only those people are able to mutually agree to a "contract", i.e. a set of principles for the general regulation of behavior. In contrast, as the definition of unwantedness refers to someone's subjective preferences, the anti-arbitrariness principle includes everyone who has preferences. That includes, e.g. young children and non-human animals.

As Scanlon, Rawls (2005) also proposed a contractualist principle which is characterized by its reason-giving nature, where a reason must be one others can "reasonably be expected to reasonably endorse" (2005: 450). Such an endorsed reason could be reinterpreted in terms of the unwanted arbitrariness principle, where the reason refers to a selection rule that justifies the selection of an element such that this selection is not arbitrary, and the endorsement of the reason refers to the selection rule not being unwanted by anyone.

Parfit made an important attempt to unify the Kantian and contractualist moral theories with a third theory, rule consequentialism, by suggesting that their fundamental principles could be interpreted in a converging way. This "Triple Theory" is summarized as (Parfit 2011: 412):

> An act is wrong if and only if, or just when, such acts are disallowed by some principle that is
> 1. one of the principles whose being universal laws would make things go best
> 2. one of the only principles whose being universal laws everyone could rationally will, and
> 3. a principle that no one could reasonably reject.

The second and third conditions represent Kantianism and Scanlonian/Rawlsian contractualism. The first condition refers to rule consequentialism (which says that everyone following the obligatory rules or principles generates the best consequences). Again, its reference to principles being universal laws reflects an anti-arbitrariness condition, but the words "making things go best" require more translation work to arrive at the anti-arbitrariness principle. Perhaps what makes things go best is a kind of preference satisfaction, such that a bridge can be built with the notion of unwantedness.

Expressed in a shorter "Kantian contractualist" formula, Parfit (2011: 342) claims: "Everyone ought to follow the principles whose universal acceptance everyone could rationally will." This unified formula turns into the anti-arbitrariness principle, by translating "Follow the principles", "universal acceptance" and "everyone could rationally will" into respectively "give justification rules", "everyone follows those rules in all possible situations" and "everyone can consistently want." This suggests that the anti-arbitrariness principle is like Parfit's Triple Theory, a kind of unification of Kantian, contractualist and rule consequentialist fundamental ethical principles.

5. *The structure of arbitrariness*

We can study unwanted arbitrariness by the most simple but sufficiently general structure: a choice set containing two elements {X,Y}. One could choose both elements, in which case there is no arbitrary selection of elements (there is only one way to select both elements). Or one could choose one element, either X or Y. This allows room for arbitrariness: if X is chosen, one could ask for the selection rule why X instead of Y is chosen. Finally, one could choose none of the elements, in which case there is no arbitrariness possible. All the options can be grouped together in the power set of all subsets: {{X,Y},{X}, {Y}, {} }. This power set has a hierarchy with several levels:

- Top level (no arbitrariness possible): {X,Y} (the full set of all elements)
- Intermediate level (arbitrariness possible): {X} or {Y} (the subsets of individual elements)
- Bottom level (no arbitrariness possible): {} (the empty set)

Only at the intermediate level is arbitrariness possible. This arbitrariness can be called first-order or horizontal arbitrariness, because there is another, meta-level arbitrariness possible, namely the choice of the

level. We can consider the set of levels: {Top level, Intermediate level, Bottom level}. If one chooses the top level without following a selection rule, that choice is arbitrary. This second-order arbitrariness can be called vertical arbitrariness. One could use a selection rule, such as "choose the level that does not allow for horizontal arbitrariness and contains at least one element", that uniquely selects the top level. Now the choice for the top level is no longer arbitrary (i.e. no vertical nor horizontal arbitrariness), but the choice of the selection rule can be arbitrary, because one could equally choose a selection rule such as "choose the level that does not allow for horizontal arbitrariness and contains no elements" (which selects the bottom level) or "choose the highest level where horizontal arbitrariness is possible" (which selects the intermediate level). Hence, there is a third-order arbitrariness. Avoiding this arbitrariness requires a fourth level, where a fourth-order arbitrariness occurs. This indicates that there will always be some arbitrariness: there will always be some level n with an n-order arbitrariness. It is impossible to avoid all arbitrariness.

6. *How to avoid unwanted arbitrariness?*

Horizontal arbitrariness involves choosing an element from a choice set. One way to avoid unwanted horizontal arbitrariness is by choosing the full set of choices (the top level) or choosing the empty set (the bottom level). However, it may not always be possible to choose the full or the empty set, because of some logical inconsistency. It may also be less desirable to choose the top or the bottom level. This undesirability happens in a general sense when at least someone cannot consistently want the full set or the empty set, or it happens in a more strict sense of "preference dominance" (similar to "Pareto dominance"): when those who cannot consistently want the intermediate level also cannot want the top or bottom level, and at least one person who can consistently want the intermediate level cannot consistently want the top or bottom level (in this case the top or bottom level is preference dominated by the intermediate level). We can categorize the situations where choosing the intermediate level is unavoidable or more desirable.

The full set and empty set are impossible: these situations often involve a choice set {do X, don't do X}. Of course, choosing both or choosing neither, is impossible.

The full set is impossible, the empty set undesirable (i.e. not wanted by at least someone): consider a choice between moral theories {moral theory X, moral theory Y}. Moral theories, such as a utilitarian welfare ethic and a deontological rights ethic, are based on universal principles. We may have a choice between {maximize total welfare, minimize the use of people against their will as merely a means to someone else's ends}. Respecting both principles of both utilitarian and deontological theories is logically impossible: there are cases when maximizing welfare involves using people as a means against their will. Choosing none

of the principles and moral theories is not impossible, but it is undesirable, because it is likely that at least someone cannot consistently want an anything goes situation without guiding ethical principles.

The full set is undesirable, the empty set impossible: suppose that helping both persons X and Y is impossible, and one faces a choice between {don't help X, don't help Y}. It is possible to choose both, but if both people want to be helped, this is less desirable than choosing either one of the options.

The full set and the empty set are undesirable. An instructive example is the choice of road traffic laws, such as the choice set: {make driving left permissible, make driving right permissible}. Choosing none of the options implies a prohibition of driving, and there are people who want to drive. Choosing both options results in more unwanted traffic accidents. Another example is: {eliminate starvation by feeding hungry people, eliminate starvation by killing hungry people}. Hungry people cannot consistently want the empty set, because that means not eliminating starvation. And they do not want the full set either, as that involves killing hungry people.

If choosing the intermediate level is unavoidable or more desirable, we might face horizontal arbitrariness, unless we are able to use a selection rule that selects one of the elements at the intermediate level. We can look for a rule "If a set of conditions C are satisfied, then choose X instead of Y." Now the challenge becomes choosing a proper set C of selection rule conditions that everyone can consistently want (otherwise, the choice of the selection rule itself generates unwanted arbitrariness). If such conditions cannot be found, then we have truly unavoidable unwanted arbitrariness.

One starting point for the selection rule could be: "If choosing X can be consistently wanted by most people, then choose X." It is already possible that everyone can consistently want this condition C that represents the majority criterion. If there remain some people who can reasonably object against this majority criterion, then they can propose another criterion (i.e. another set of conditions for the selection rule). Now we face the choice of selecting an element from the set {majority criterion, another criterion}. Choosing both elements (the full set) is impossible, choosing the empty set undesirable. To avoid horizontal arbitrariness, we need another, higher level selection rule that selects either the majority criterion or the other criterion. This process can continue to even higher levels. We can go on as far as is feasible, to minimize unwanted arbitrariness. But the further we go, the more important the choice of a higher level selection rule becomes, the more depends on it, and the harder it becomes to have reasonable objections against the choice. The preferred higher level selection rule becomes so fundamental, that one is likely to have a strong preference for it. It is for example already difficult to have a stronger preference for another criterion than the majority criterion. That means the majority criterion

selection rule is likely consistent with someone's largest consistent set of that person's strongest subjective preferences.

With the above line of reasoning, we can apply the anti-arbitrariness principle to itself. The choice set involves the two options {avoid unwanted arbitrariness as much as possible, don't avoid unwanted arbitrariness as much as possible}. Choosing both or none of the options is impossible. So we are stuck at the intermediate level, where we can arbitrarily pick one of the two options. But picking the second option (not avoiding unwanted arbitrariness) immediately becomes extremely unwanted. Allowing avoidable unwanted arbitrariness has so many ramifications, that it is likely in contradiction with anyone's largest consistent set of strongest subjective preferences. So you cannot consistently want the arbitrary choice for the second option.

To see this in more detail, suppose that you disagree with the anti-arbitrariness principle. You say that avoidable unwanted arbitrariness is permissible. But then you cannot give reasonable counterarguments when I allow unwanted arbitrariness in my moral choices. I may follow arbitrary principles that you cannot consistently want. When I impose my choices on you, you are not able to complain. You are not able to give justified arguments against the imposition of my choices, because you acknowledged that unwanted arbitrariness is allowed, and hence that it is permissible to arbitrarily ignore or violate someone else's largest consistent set of strongest preferences.

If you permit unwanted arbitrariness, I can say to you that your moral values and judgments are not valid. And if you complain and say that your ethical theory is valid, then I can reply that if you are allowed to arbitrarily exclude other moral views and make an ad hoc exception for your own moral rules, then so am I. So I may even make the exception that everyone's moral views should be respected, except yours. All your objections can easily be bounced back by saying: "If you are allowed to arbitrarily do that, then so am I, and so is everyone. What would make you so special that you are allowed to arbitrarily exclude others but I am not? You should not arbitrarily pick yourself from the set of all individuals and say that you are the only one who may do that thing." In summary: rejecting the anti-arbitrariness principle while avoiding irrationality, is extremely difficult, if not impossible. The above discussion applies to the cases where the top and bottom levels are impossible or undesirable. There are two other interesting categories to consider.

The full set is possible and not clearly undesirable, the empty set is undesirable or impossible. A prime example is the choice set {I decide, you decide}, or {I have a right to vote, you have a right to vote}. Someone has to decide, and at least someone wants to vote, so the bottom level is impossible or undesirable. But choosing the intermediate level and arbitrarily choosing one of the options results in a kind of dictatorship where one person can decide or vote.

The full set is impossible or undesirable, the empty set is possible and not clearly undesirable. Here we deal with choice sets such as {harm person A, harm person B} or {privilege A over B, privilege B over A}. It is undesirable to harm both A and B and it is not possible to privilege A over B and B over A at the same time, so the top level is undesirable or impossible. But choosing the intermediate level and arbitrarily choosing one of the options results in a kind of discrimination where one person is harmed or disadvantaged.

As the anti-arbitrariness principle deals with choices and rules, we are confronted with two important questions. Who decides or chooses the choices and rules? And who is affected by the choices and rules? These two questions relate to the dual problems of dictatorship and discrimination. The next two sections discuss how the anti-arbitrariness principle implies the non-dictatorship and non-discrimination principles.

7. Implication 1: Non-dictatorship

The non-dictatorship principle says that no-one should have the unconditional power to always unilaterally make decisions that negatively affect some other people. A vote is a power (or right) to influence a decision (the outcome of a decision process) made by a group, such that the outcome is more in accordance with one's personal preferences. In a dictatorship, there is at least one individual whose vote is excluded from the decision process and who does not want this exclusion. A dictatorship clearly violates the anti-arbitrariness principle, because the choice for the dictator is arbitrary (as the dictator's power is unconditional, no rule was followed to grant that power), and unwanted (when there are affected people who do not want the decisions made by the dictator).

Suppose person X wants to make choice A, but person Y cannot consistently want the consequences of that choice, and hence prefers choice B. Instead of the principle might makes right, which is a dictatorship of the most powerful, those people can look for other methods to decide who gets to decide. One such alternative method is generating justifications by giving arguments. Instead of the strongest person winning, now the strongest reason, justification or argument wins. The principle that the best argument wins, is also arbitrary, just like the principle that might makes right, but it is less likely to be unwanted.

Person X can simply claim: "I, person X, decides." This is the moral rule: "If the person is X, then that person may choose." Person Y does not want that, and counters: "No, person Y decides." The justification rule proposed by person X refers to X, and that choice should be justified as well. So person X can claim the meta-rule: "Person X decides who decides." But here again, person Y can complain, and the meta-rule arbitrarily refers to person X again. This discussion can go on to infinity. For practical relevance, the anti-arbitrariness principle should state that an infinite regression of justification rules is not allowed.

The non-dictatorship principle can also be applied to moral theories. These theories are logical systems of ethical principles that represent moral intuitions or values. There are different moral theories, such as a deontological rights ethic, a consequentialist utilitarian welfare ethic, a libertarian ethic or pluralist ethics that combine several ethical principles. But which theory should we choose? The anti-arbitrariness principle sets strong constraints on a moral theory. The theory should be coherent in the sense that it should be constructed following some rules, such as:

1) One should not arbitrarily limit the ethical principles to an arbitrary group of objects, beings or individuals.
2) One should not arbitrarily give weaker (less strongly felt) moral intuitions stronger priority. One should not arbitrarily change or exclude basic moral judgments.
3) One should not arbitrarily allow inconsistencies and gaps in the ethical system.
4) One should not arbitrarily introduce ambiguous or vague principles that one can interpret and apply arbitrarily in concrete situations.
5) One should not arbitrarily add artificial, complex, ad hoc constructions and exceptions to save the moral theory from counter-intuitive implications.

These construction rules for a coherent theory can be consistently wanted. If, for example, I allow inconsistencies, gaps, ambiguities or arbitrary exceptions in my theory, then I have to accept that your moral theory also contains such things. With such an incoherent theory, you can easily justify choices that I cannot consistently want. An incoherent theory always contains avoidable unwanted arbitrariness that should be rejected.

To avoid dictatorship, everyone is allowed to construct a coherent moral theory that best fits one's moral intuitions and values. Incoherent theories are impermissible. But there are many possible coherent moral theories. We do not have a rule that determines which of those coherent theories is the best. If we are against unwanted arbitrariness, we have to recognize that every equally coherent moral theory is equally valid. I cannot say that my coherent theory, based on my moral intuitions, is better than yours if both our theories are equally coherent. I prefer my theory, but I cannot impose my theory upon you, because what would make me so special that I would be allowed to do that? And the same goes for you and everyone else. It would be an unwanted kind of arbitrariness if I claim that my moral theory is special without good reason.

So picking one of the coherent moral theories always involves unavoidable arbitrariness. The non-dictatorship principle says that we should democratically choose which moral theory to apply. And if you follow a coherent moral theory without being able to give a justification

rule that selects that theory, you should tolerate that other people follow other coherent moral theories. We should be tolerant towards all other coherent ethical systems, no matter how much they go against our own moral intuitions.

A choice for an incoherent system, on the other hand, does not have to be condoned, because you can give a justification rule "If the theory is incoherent, then it is impermissible to choose it," and everyone can consistently want that everyone follows this rule. If you choose to follow an incoherent theory, I am allowed to reject that theory and impose my theory on you, and you are not able to complain. You are not able to give reasonable or justified counterarguments against the imposition of my ethical principles, because by following your incoherent theory, you are acknowledging that unwanted arbitrariness and hence arbitrary exclusion are allowed. That means it is also permissible to arbitrarily exclude your moral theory and ignore your moral views and ethical principles. You can only give a valid complaint or argument if you accept the anti-arbitrariness principle. Without that principle, any critique becomes invalid and complaints become impossible.

As the ethical systems of, e.g. racists, rapists or religious fundamentalists contain inconsistencies, avoidable arbitrariness, unscientific beliefs and vague principles, they can easily be rejected. If your ethical system is more coherent than theirs, then you can rightfully say that your ethical system is better than theirs and then you may oppose their incoherent systems.

The prohibition of incoherent theories allows us to avoid an extreme form of moral relativism that says that all moral theories, including incoherent ones, are equally valid. This extreme relativism implies that everything would be permissible, and we cannot consistently want that. The non-dictatorial claim that coherent moral theories are equally valid is a kind of weak moral relativism, which is a consequence of the anti-arbitrariness principle.

How do we deal with that plentitude of coherent ethical systems that are equally valid? Everyone (who is able to do so) constructs their own coherent ethical systems, and we can aim for a consensus or democratic compromise between everyone's system by using a democratic procedure. In a democracy, everyone has one vote, and everyone's vote is equally important, because we cannot say that one vote (one coherent theory) is better than someone else's. But those who cannot provide a coherent moral theory that does not contain unwanted arbitrariness, lose their vote. In other words: in this moral democracy it is not allowed to vote for parties who have incoherent moral theories, such as racist parties. Those parties cannot participate in elections.

Note that the coherence of moral theories imposes very strong constraints on the construction of moral theories. We can expect that the resulting theories that people construct, if they follow the anti-arbitrariness principle carefully, are not extremely divergent from each

other. This strong selection and convergence of moral theories makes a democratic choice of theory more feasible.

So there are two reasons why our moral theories should not contain unwanted arbitrariness. First, if it contains such arbitrariness, someone else is allowed to arbitrarily reject our theory and we are not able to complain. Second, the avoidance of unwanted arbitrariness puts strong constraints on the possible moral theories, which makes a democratic consensus between the resulting coherent moral theories more feasible.

8. *Implication 2: Non-discrimination*

Discrimination can be defined in different ways, suitable for different contexts (see e.g. Altman 2016). One could for example define discrimination merely as a different treatment of two individuals (or groups of people), but then we must distinguish permissible versus impermissible discrimination and define the latter. The following definition of arbitrary discrimination is suitable to derive the non-discrimination principle from the anti-arbitrariness principle.

Arbitrary discrimination of individual (or group) A relative to B by discriminator C is a systematically different treatment of A and B, whereby

1) B is given more advantages by C than A,
1) C believes A has a lower moral status than B (e.g. A has less intrinsic value or weaker rights than B) in the sense that C would not tolerate swapping positions (treating A as B and B as A), and
3) there is no justification or the justification of the difference in treatment refers to morally irrelevant criteria (properties that are not acceptable motives to treat A and B differently in the concerned situation), whereas A and B both meet the same morally relevant criteria to treat and value them more equally.

The first two conditions reflect unwantedness. The discriminated person A does not want the disadvantage, but also the person C who discriminates does not want swapping positions of A and B. The third condition reflects arbitrariness, i.e. the lack of a justifying rule. Discrimination is based on arbitrariness, and this arbitrariness is avoidable and unwanted, because the discriminated people do not want their negative treatment, their arbitrary exclusion from the moral community.

The anti-arbitrariness principle specifies what counts as morally irrelevant criteria. A criterion or property is morally irrelevant in a specific context (such as political elections or job opportunities), when it is arbitrary (in the sense that there is no non-circular rule that selects the property out of a multitude of similar kinds of properties), or it has a high risk of introducing arbitrariness. The latter happens with, for example, ambiguous properties, properties that are inherently impossible to detect, define or delimit, or non-empirical properties for which there are no objective or scientific criteria and methods—not even in

principle—to clearly see whether the property is present. With such properties, there is the risk that one arbitrarily assigns the property to individuals as one pleases. Consider a non-natural property such as a soul, and the claim that only beings that have a soul have rights. The danger is that one can arbitrarily assign a soul to some preferred entities or persons.

With the anti-arbitrariness principle we can derive which properties are morally irrelevant in which contexts and hence result in discrimination in those contexts. Some properties that are irrelevant in, for example, the context of political voting are: physical characteristics and appearances (e.g. skin color, behavior, gender), genetic properties (e.g. race, ethnicity, genetic kinship), supernatural properties (e.g. having a soul), preferences (e.g. sexual, political), and belonging to an arbitrary group.

As a concrete and important example of the non-discrimination principle, consider the choice of moral community: the subset of all entities in the universe that have moral status (in the sense of, e.g. having moral rights). Consider only living beings. According to the biological classification, we can classify living beings in a vertical taxonomic hierarchy, with the taxonomic rank "life" at the top, followed by ranks such as "domains" (e.g. eukaryotes), "classes" (e.g. mammals), "orders" (e.g. primates), and finally the taxonomic rank "populations" (races, subspecies) at the bottom. A white supremacist first chooses the lowest level in this hierarchy (the populations or ethnic groups), and then picks a subset at this level (the ethnic group of whites). Similarly, a speciesist first selects the level of the species, and then selects a specific species (e.g. *Homo sapiens*) as the moral community. If no selection rules were followed, these two choices involve respectively vertical and horizontal arbitrariness. We can first ask the non-trivial question: "Why choosing a species and not, e.g. a biological order or a phylum?" And at the level of the species, we can ask: "Why choosing *Homo sapiens* (humans) and not, e.g. *Sus scrofa* (pigs)?" One could answer: "Because most humans have the capacity for moral thought", but it is possible that this answer also applies to some levels up or down in the hierarchy. If, for example, there are less than 14 billion primates alive, containing more than 7 billion humans with the capacity for moral thought, then the majority of primates have this capacity. Hence, one could equally well first select the level of orders and then the order of primates. By selecting a biological group as a moral community, it is not easy to avoid arbitrariness.

The definition of discrimination means you can avoid discrimination in three ways: either treating A and B equally, tolerating swapping their positions or justifying the preferential treatment using non-arbitrary criteria.

If you tolerate swapping the positions of A and B, you give them equal moral value. This implies that some kinds of partiality are not (yet) discriminatory. Consider a burning house dilemma where you can

either save Alice or Bob from the flames. Suppose you want to save Bob first because he is your child, whereas Alice is a child from another country, with another skin color. Non-discrimination does not imply that you should flip a coin and give each child an equal 50% survival probability. You are not a racist or sexist (at least not necessarily) if you want to save Bob, as long as you do not condemn someone else who wants to save Alice. If you criticize someone who saved Alice, and you do so by using arbitrary criteria such as skin color or gender, then you discriminate and then it becomes racism or sexism. It permissible for you to show partiality to Bob for reason r (for example, because you feel attachment to Bob) if you tolerate others failing to show partiality to Bob for reason r.

Considering the above, we can formulate the following ethical principle of tolerated partiality: when helping others, you are allowed to be partial in favor of one individual or group (e.g. your own child), as long as you tolerate someone else's choice to help the other party (e.g. another child). In this sense, saving your child is not inconsistent with the claim that all children have an equal moral value. Two children can have different personal values for you, but they inherit an equal moral value when a tolerated symmetry (swapping their positions) is satisfied. Having a stronger empathic connection for one individual or having a stronger inclination to save one individual instead of the other, and acting on those feelings, is not necessarily discrimination.

This principle of tolerated partiality can be derived from the unwanted arbitrariness principle: everyone should tolerate your preference for saving the people you hold dear, even if your selection of those people is arbitrary (e.g. from my perspective), because everyone can consistently want to be able to save the people they hold dear.

What if you do not tolerate swapping the positions of Alice and Bob? Suppose Bob is your child and Alice is the name of my car. You would not tolerate me saving the car. The definition of arbitrary discrimination implies that to avoid discrimination, there must be a valid reason or justification, based on non-arbitrary criteria, why one entity (the child) is more important or valuable than the other (the car). In this example you can easily give a valid reason: the child has preferences to be rescued, to keep on living and to avoid the pain from the flames, whereas the car does not care at all about being burned or rescued.

Similarly, suppose you give a piece of chocolate to Bob, a child, instead of Alice, a dog. You have a non-arbitrary justification: chocolate is unhealthy for dogs. Being able to safely eat chocolate is a non-arbitrary criterion, because both the dog and the child prefer safe food. Non-discrimination does not say that we must treat everyone the same and give everyone the same food.

However, some reasons are invalid in cases when you do not tolerate swapping positions. For example, the reason to save Bob instead of Alice because Bob belongs to a certain social group or believes in a certain God. Those invalid reasons refer to arbitrary criteria, such as

skin color, religious beliefs or group membership. A white supremacist might help Bob instead of Alice (and does not tolerate someone saving Alice instead of Bob) based on their skin colors, but what does skin color have to do with a preference for being helped? Skin color is but one bodily characteristic, and it is arbitrary to claim that this particular characteristic relates to subjective preferences.

In summary, when swapping positions is not tolerated, the reason should not be arbitrary. When swapping positions is tolerated, the arbitrariness of the reason is not problematic. 'Being your child' may be an arbitrary reason to save your child, because what does that have to do with a preference for being helped? So if you use this as your reason, then you have to tolerate swapping positions (i.e. someone else saving another child).

To avoid discrimination, we have to expand the moral circle (Singer 2011). This expansion visualizes the traditional approach in a rights-based ethic. One traditionally starts with the list of rights and then asks the question: what are the entities in the world that should get these rights? Then we see an expanding circle: from the individual to the family to the tribe to the ethnic group to the species, ending up with the Universal Declaration of Human Rights. But selecting some entities or persons is arbitrary, and the consequences of this selection cannot be consistently wanted by individuals who are not selected.

The anti-arbitrariness principle suggests a reverse approach: to avoid arbitrary exclusions, we first start with the condition that everyone and everything gets rights. Then we ask the question: what are the basic rights that should be granted to all entities in the world?

Of course we cannot grant all possible rights to all entities, because that results in contradictions. Hence, the choice of rights might involve unavoidable arbitrariness. To avoid unwanted arbitrariness, we can look for the rights that are least unwanted or that can be selected following some rule.

Consider the right not to be killed. This right is trivially satisfied for non-living things, but if all living things get this right, we are no longer allowed to kill and eat plants. We can restrict this right to the right not to be killed against one's will. The addition of "against one's will" is possible, because everyone can consistently want such addition. Assuming plants do not have a consciousness and hence no will, this right is trivially satisfied for plants: even if we eat them, we do not kill them against their will and hence do not violate their right. We can easily grant plants this right.

But this right can still be unwanted: there are situations where we can save many people, only by accidentally or unintentionally killing one person against his will. When that person is a rightholder who has the right not to be killed against his will, the presence of that person imposes a cost on others: the other people can no longer be saved. They lose the freedom to be saved. The rightholder becomes an obstacle: it

would have been better for the other people if that one person was absent or did not exist.

As argued by Walen (2014), there is however another right that does not impose costs on others: the right not to be used as a means against one's will. One is used as a means for someone else's ends if one's existence and presence is necessary to achieve the ends. If everyone has this right instead of the right not to be killed, it is still allowed to save people by accidentally killing someone. Bringing into existence a person who has that right is not costly or harmful for others, because other people would not have been better-off if the person were absent. Consider the case of an unwanted pregnancy: abortion violates the right of the embryo not to be killed, but not the right not to be used as merely a means. Performing an abortion, the embryo is not used as a means, because the woman could still achieve her end (i.e. not being pregnant) if the embryo did not exist. In contrast, when the pregnancy is unwanted, one could say that the mother is used as a means against her will: her existence is necessary for the embryo to live. The embryo uses the body of the mother against her will.[3]

This right not to be used as means against one's will reflects a Kantian mere means principle (see Kant 1785 and Parfit 2011): if "use" generalizes to "treat" and "against one's will" translates into "merely", the no-mere-means right says that we should not treat someone as merely a means for someone else's ends.

Now we can formulate a selection rule to select this no-mere-means right: choose the right that refers to the person's will and does not impose costs on others (in the sense that others would not be better-off and cannot be made better-off if people who have the right were absent). The absence of costs means that it is difficult to complain against granting people this no-mere-means right. With the selection rule and the difficulty to complain, choosing this no-mere-means right is likely to avoid unwanted arbitrariness. Everything gets this right, but the right is only non-trivial for individuals who have a will (which consists of subjective preferences). This includes children and animals. As a practical result, this right imposes a duty of veganism. If animals have negative experiences when their bodies are used for food, their no-mere-means right is violated. And if only humans and some preferred non-human animals get this no-mere-means right, we are guilty of discrimination.

9. *Conclusion*

The anti-arbitrariness principle states that everyone who makes a choice has to avoid unwanted arbitrariness as much as possible. This principle strongly relates to Kantian, Scanlonian and Parfitian cate-

[3] Note that in this sense, using someone as a means does not have to be intentional.

gorical imperatives (Kant 1785; Scanlon 1998; Parfit 2011). Its most important implications are non-dictatorship and non-discrimination.

I will leave this discussion with some open questions for further research. Could the anti-arbitrariness principle be too strong in the sense that it prohibits too many ethical principles and choices that we deem to be valid and permissible? Could we find some kinds of arbitrariness that can still be justified, even if someone cannot consistently want them? Are there other fundamental ethical principles, conflicting with the anti-arbitrariness principle, that everyone can consistently want? If yes, can those other principles be justified? And when people have different coherent moral theories but cannot find a democratic consensus, how can we select the best moral theory? The latter moves us to the area of "normative uncertainty" (MacAskill 2014).

References

Altman, A. 2016. "Discrimination." In Edward N. Zalta (ed.). *The Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/archives/win2016/entries/discrimination/.

Bentham, J. 1789. *An Introduction to the Principles of Morals and Legislation*. London: T. Payne & Son.

Carson, T. L. 2010. "Kant and the Absolute Prohibition Against Lying." In T. L. Carson. *Lying and Deception: Theory and Practice*. Oxford: Oxford University Press, 67–88.

Kant, I. [1785] 1993. *Grounding for the Metaphysics of Morals: Third Edition*. Translated by James W. Ellington. Indiana: Hackett.

MacAskill, W. 2014. *Normative Uncertainty*. PhD thesis. University of Oxford. https://ora.ox.ac.uk/objects/uuid:8a8b60af-47cd-4abc-9d29400136c89c0f/files/m0e6d06ceaf493f85c33c6faee369d19b

Parfit, D. 2011. *On What Matters*. Oxford: Oxford University Press.

Rawls, J. 2005. *Political Liberalism*. New York: Columbia University Press.

Scanlon, T. 1998. *What We Owe to Each Other*. Cambridge: Harvard University Press.

Singer, P. 2011. *The Expanding Circle: Ethics, Evolution, and Moral Progress*. Princeton: Princeton University Press.

Walen, A. 2014. "Transcending the Means Principle." *Law and Philosophy* 33 (4): 427–464.

Croatian Journal of Philosophy
Vol. XXIII, No. 68, 2023
https://doi.org/10.52685/cjp.23.68.6
Received: October 28, 2022
Accepted: July 2, 2023

Reassessing the Exploitation Charge in Sweatshop Labor

HUSEYIN S. KUYUMCUOĞLU*
Kadir Has University, Istanbul, Turkey

One common argument against sweatshops is that they are exploitative. Exploitation is taken as sufficient reason to condemn sweatshops as unjust and to argue that sweatshop owners have a moral duty to offer better working conditions to their employees. In this article, I argue that any exploitation theory falls short of covering all standard cases of sweatshops as exploitative. In going through the most prominent theories of exploitation, I explain why any given sweatshop can either be wrongfully exploitative or not, depending on the exploitation theory being considered and the circumstances of the application. I conclude by suggesting that sweatshop critics had better find other reasons besides the charge of exploitation to protest or interfere with these workplaces.

Keywords: Exploitation; theories of exploitation; sweatshops; sweatshop criticism.

1. *Introduction*

One widespread reason to protest sweatshops is the exploitative working conditions within.[1] This protest is based on the argument that

* I am grateful to Christopher Morgan-Knapp for his support and help in writing this article and for asking me challenging questions regarding the arguments involved. I would also like to thank my colleagues in the Society for Practical Philosophy, Turkey (Çağlar Çömez, Umut Eldem, Maya Mandalinci, Beşir Özgür Nayır, Seçil Aracı, Arzu Formánek) for their valuable feedback on an initial version of the article.

[1] Exploitation is not the only reason to protest sweatshops. Other reasons might be coercion, background injustice, inhumane working conditions, the moral requirement of attaining better conditions, or any other anti-globalization political motive. This article focuses exclusively on exploitation.

sweatshop workers are exploited, and this fact constitutes one reason to protest sweatshops or interfere with them to ameliorate the working conditions.

In this article, I will go through some of the prominent theories of exploitation and argue that for all these theories, there are some circumstances, though maybe different for each theory, under which specific sweatshops are wrongfully exploitative while others are not. Hence, I argue against the contention that all standard cases of sweatshops are exploitative even for a sturdy theory of exploitation. I further claim that sweatshop critics had better find other moral grounds *beside* the exploitation charge to protest sweatshops if they want a more substantial moral ground for their criticism.

To build my argument, I will analyze the prominent definitions of exploitation in the literature and discuss how each explains the alleged exploitation in sweatshops. I will show that there are at least some circumstances for all these theories, and not necessarily the same circumstances for each theory, under which specific sweatshops are wrongfully exploitative. Nevertheless, no theory among the ones I investigate marks all standard cases of sweatshops as exploitative. I will also point at some theoretical flaws of each theory.

In the first and second sections below, I will distinguish between two main approaches to the concept of exploitation: definitions that focus on the outcomes and definitions that focus on attitudes. The first set of definitions takes exploitation to be related to the fairness of an interaction between the parties involved. On the other hand, the second set of definitions focuses on the attitudes of the involved parties during their interaction regardless of the fairness of the outcome.

2. *Exploitation and outcomes*

Outcome-based exploitation accounts focus on whether or not the allegedly exploitative relationship distributes the common surplus fairly. Here, what is referred to as the common surplus is any value created due to the interaction between parties A and B that would not be created had these parties not interacted. A general definition that these accounts would agree on can be formalized as this:

> (E1): A exploits B if and only if A benefits from a transaction with B, in which A takes unfair advantage of B.

The difference between the accounts under this category stems from their approaches to what constitutes unfairness in an interaction.

2.1. *Wertheimer's theory of exploitation*

Alan Wertheimer maintains that a transaction between two parties is unfair if the way they share the surplus conflicts with the shares that the parties would have had there been a hypothetical competitive market for the goods and services they transact (Wertheimer 1996:

230–236). Wertheimer's account of fairness is transaction-specific and does not consider how well-off the agents are in comparison to each other apart from the terms of the transaction itself.

Many authors have criticized this theory for using competitive markets as fairness criteria. Ruth Sample, for instance, finds Wertheimer's criterion conservative and questions his claim that exploitation consists of violating market practices: "Wertheimer does not explain why the market price should be regarded as the fair price and why nonmarket prices are exploitative" (2003: 24). Moreover, even if we accepted the fair market price as a criterion to detect exploitation, we would find cases where talking about such a price does not make sense.

For example, take a porcelain collector who lacks only one piece in his extensive collection of rare pieces. This piece stands in a store whose owner raises its price after hearing about the interest of this collector. The collector wants the piece so badly that he buys it for the new exorbitant price.[2] He seems overcharged because the store owner raised the price after learning about the collector's interest in this piece. However, to talk about exploitation, in this case, would require us to imagine the piece's price in a hypothetical competitive market.

In such a market, the piece would have the price "that an informed and unpressured seller would receive from an informed and unpressured buyer" (Wertheimer 1996: 230). There are multiple buyers and sellers in such a hypothetical competitive market to ensure that "neither party takes *special* unfair advantage of the particular defects in the other party's decision-making capacity or special vulnerabilities in the other party's situation" (Wertheimer 1996: 232, emphasis in the original). However, if there were multiple buyers and sellers of this porcelain piece, then there might be less of an incentive to make a collection of it. After all, people who are into collections tend to collect rare items, for which, by definition of a rare item, there cannot be multiple buyers and sellers. Hence, imagining a hypothetical competitive market for such items is difficult.

To be charitable to Wertheimer's definition, we can ignore the clause on "multiple buyers and sellers" and focus on the part where he talks of taking *special* unfair advantage of special vulnerabilities in the other party's situation (cf. 1996: 232). We can blame the store owner for taking a special unfair advantage, thus exploiting the collector when they use the information that reveals this collector's vulnerability.

However, unless the store owner's act also involves unfairness, it would bear the problems of an account of exploitation based on the idea of vulnerability. I explain such an account as (E2) and discuss its disadvantages below. Otherwise, if it involves unfairness, it has to account for how to come up with a hypothetical price for the porcelain piece in question.

[2] One can find a similar example in Sample (2003: 14).

It is, then, unclear whether the porcelain collector is exploited, as per Wertheimer. It is even dubious whether he is overcharged because we need a standard price to talk about overcharging, which is nonexistent in this case. He might well "feel" overcharged, yet this does not offer much help for the normative analysis I am pursuing here.

The flaw of Wertheimer's account of exploitation in explaining the moral issues in exchanging rare items might disclose a theoretical weakness on its side. Nevertheless, we still have to examine the explanatory power of this theory to analyze the exploitation charge in sweatshops.

Determining a hypothetical market price for sweatshop labor is difficult. One difficulty stems from how hypothetical we imagine this market to be. In other words, we must find out how much of a deviation from an actual market situation in a given region of sweatshops we want to imagine.

If we were to imagine a hypothetical market on a global scale, say for just the garment industry, keeping other factors constant, we would imagine a global labor market of workers and capitalists from all countries worldwide. Under such circumstances, other things equal, one would expect the more densely populated regions to have lower market wages than more sparsely populated regions since the labor supply is higher in these densely populated regions. This situation might result in a neighborhood in Manhattan having a lower hypothetical market wage for the garment industry than Brahmanbaria, a city in Bangladesh similar in size to Manhattan but with a population of at least one-tenth. The counterintuitive implication would be that while workers in Brahmanbaria are exploited with the current real wages they receive, the workers in Manhattan exploit their employers.

Holding everything else stable and focusing on population density to imagine a hypothetical market might not do justice to what Wertheimer had in mind. However, the point of the example is to mark the difficulty of imagining a hypothetical market for sweatshop labor globally.

Sample maintains that Wertheimer's account would fail to distinguish the wage-labor exchange between the MNE decision-makers and the sweatshop workers as exploitative. According to her, Wertheimer's account would fail in these cases because "[t]here is a competitive market for labor in Pacific Rim countries, but there are more workers than there are capitalists. Thus the competitive market price for labor is relatively low" (Sample 2003: 24).

Sample's claim might require more explanation here. The fact that there are more workers than capitalists should not suffice to say that the relevant labor relation is exploitative unless one believes that a labor-capital relationship is necessarily exploitative since the number of workers is always more than the number of capitalists in any capitalist part of the world. We could interpret Sample's idea to mean that the capitalist/worker ratio in the Rim countries is below a certain ratio, above which it would not make sense to talk of an exploitative labor

relationship according to Wertheimer's hypothetical market price criterion for exploitation.

Cohen's idea of "collective unfreedom" can help us here understand the significance of this capitalist/worker ratio.[3] Cohen explains collective unfreedom as follows: "a group suffers collective unfreedom with respect to a type of action A if and only if performance of A by all members of the group is impossible" (Cohen 1983: 16). He maintains that although each worker in a capitalist portion of the world is free to stop being a worker and become an entrepreneur using their savings or other loans, if possible, such freedom cannot be used by all workers. Thus, workers suffer from collective unfreedom to stop being workers.

Cohen adds, "[c]ollective unfreedom comes in varying amounts, and it is greater the smaller the ratio of the maximum that could perform A to the total number in the group" (1983: 16). One would expect this ratio to be quite small in a region where sweatshops are widely used in production. So, we would not expect many workers to be free to stop being workers and start their own sweatshops. Hence, workers would suffer from a higher level of collective unfreedom in such a region.

Now, we can take Sample's idea of the capitalist/worker ratio to follow Cohen's idea of collective unfreedom. So, as Cohen advocates, the lower the capitalist/worker ratio in any given region, the more exploitative conditions are in that place. However, unless we want to accept that all possible paid labor is exploitative, we need to mark a threshold for the capitalist/worker ratio at which labor relations become exploitative. Therefore, we still need to say more than what Cohen theorizes about workers' collective unfreedom.

Sample might be wrong in her assumption that there is a problem with Wertheimer's theory just because it fails to mark sweatshop labor as exploitative. Still, her criticism gives us a new idea for interpreting the criterion of a hypothetical market. We can now take the hypothetical market to imply a given region rather than the whole world and the hypothetical market wage as the wage resulting from an "ideal capitalist/worker ratio" in this region. This idea would be a better interpretation of Wertheimer's theory.

This result is still counterintuitive, according to Jeremy Snyder. He argues against a hypothetical market price definition of exploitation: "An interaction may be fair by the standards of a hypothetical fair market (or another standard of micro fairness), but leave workers without sufficient income to meet their basic human needs" (Snyder 2010: 199). He concludes that the fact that workers cannot meet their basic human needs in the presence of a hypothetical market price for labor is sufficient to reject Wertheimer's theory of exploitation.

Snyder is correct that a hypothetical market wage might leave workers with insufficient provisions. However, there would be nothing inconsistent in Wertheimer's theory to say that this wage is non-

[3] I thank an anonymous reviewer for reminding me of this concept.

exploitative. Wertheimer could still concede that other moral wrongs might be involved (like a history of colonization or a corrupt government) or some unfortunate events (like famine) that caused the circumstances under which a hypothetical market wage is so low. Thus, he would insist that no exploitation is involved as long as the wage in question is at a level commanded by a hypothetical labor market in this region.

Let me go back to the interpretation of the hypothetical market wage that depends on the idea of an "ideal capitalist/worker ratio." The theory has to hold that a hypothetical market wage would be obtained if such a ratio were maintained in a given region. This ratio ensures that an unpressured seller takes a price from an unpressured buyer. The difficulty of determining such a ratio is evident. If we point at a particular country or region of the world as an ideal example of such a ratio because there is no exploitation there, this will beg the question. We can define exploitation by looking at a given non-exploitative relationship only if we accept this particular non-exploitative case as a foundation for our theory. This argument is different from what Wertheimer defends.

Even if we could determine an ideal capitalist/worker ratio by looking at a particular part of the world to help us determine a non-exploitative hypothetical market wage, we would not be able to use it in other cases. Imagine that in a specific region where living conditions are harsh, power plant companies struggle to employ engineers because engineers prefer to live in parts of the world where living conditions are better than in this region. If a capable engineer asks for an exorbitant wage from a power plant company in such a region, we could consistently hold that this is an exploitative offer because the wage this engineer demands would be higher than the average wage engineers receive in the part of the world where the capitalist/worker ratio is ideal.[4] However, we would then lose Wertheimer's condition of a hypothetical market: an unpressured seller takes a price from an unpressured buyer. This result would constitute another reason we cannot take an existing location to indicate an ideal capitalist/worker ratio.

Another method would be giving another criterion, such as workers' capacity to meet their basic needs, to indicate this ratio. However, now, there would not be any need for the hypothetical market criterion because this new criterion, i.e., meeting the basic needs, would do the conceptual work. Thus, the hypothetical market criterion for determining fairness does not help clarify Wertheimer's exploitation theory.

Indeed, Wertheimer provides his readers with an alternative way to interpret fairness in a transaction, although he refuses to use it to develop this position further. According to this interpretation, a transaction between two parties is unfair if the way the surplus is shared

[4] I thank an anonymous reviewer for bringing this thought experiment into the discussion.

conflicts with the shares that the parties would end up having, according to the rational bargaining theory (Wertheimer 1996: 218–221).

Sollars and Englander (2018) develop an account of fairness based on this idea. They define the reservation prices of the seller and buyer of a product as R_S and R_B, respectively. R_S is the price below which the seller would not wish to transact, and R_B is the price above which the buyer would decline the transaction. For a transaction to occur, R_S has to be less than or equal to R_B; $R_S \leq R_B$. As a result of any transaction, a surplus will be defined by their difference, $R_B - R_S$. According to their theory, a fair transaction would divide the surplus equally; $(R_B - R_S)/2$. Any deviation from this amount would make the transaction exploitative.[5]

Their exploitation theory helps them distinguish between a company that makes enormous profits using sweatshops in their production line and another case where a relatively small domestic company uses them. In both cases, workers might find it challenging to make ends meet. Nevertheless, while we may rightfully blame the big company for exploiting its workers, we might not say the same for the small company.

The difference stems from the surplus ratio $(R_B - R_S)$. The big company is thought to have a higher R_B, so it is expected to yield a higher wage for its workers. If this company does not share half of the surplus with its workers, it is charged with exploitation. The small company could also be charged with exploitation unless it shares half of the surplus. However, even if the small company renounces all the surplus favoring its workers, this might make a minimal change in workers' wage levels. So, although the workers of the big company might make more money than those of the small company, the former are exploited while the latter are not, according to this theory (cf. Sollars and Englander 2018).

This account seems more stable than the hypothetical market criterion for fairness. Nevertheless, the problem lies in the initial presuppositions of the theory. Rather, indeed, the problem lies in what the initial presuppositions neglect. Rational bargaining theory neglects the background circumstances that affect the reservation prices of the transacting parties. As Wertheimer contends, "[...] rational bargaining always reflects the prebargaining position or endowments of the parties [...]" (Wertheimer 1996: 220). This theory ignores why sweatshop workers have a reservation price that is insufficient to fulfill their basic human needs.

A defender of Wertheimer's theory of exploitation could respond that those background circumstances enable a transaction between large MNEs and poor sweatshop workers in the first place, and thus

[5] Sollars and Englander explain that this random division of the surplus into equal portions is just an easy starting point for their argument: "we stress that we choose this criterion mainly for the ease of exposition. We believe that the selection of the best criteria for dividing the surplus within the context of sweatshops is a matter for future research" (2018: 24).

including these circumstances in a theory of exploitation to mark them morally wrong would be ignoring the basic requirements for foreign direct investment. However, this rejoinder misses the point of the need for a theory of exploitation that purports to explain a mutually advantageous albeit morally wrong relationship.

If a mutually advantageous relationship involves unjust background circumstances in which the parties find themselves, a theory of exploitation is expected, at least not to ignore them. I do not mean to defend that any transaction made under unjust background conditions is necessarily exploitative. A theory of exploitation must explain why and when they are irrelevant if that is what the theory claims. Unfortunately, Wertheimer's theory of exploitation falls short of this research goal.

This last response would only satisfy some, and I wish not to push it further since I do not necessarily need it to make my main argument. So, regarding Wertheimer's theory, I can say that until a better account of hypothetical market criterion is given, his theory works best with the approach suggested by Sollars and Englander, and their method concedes that we can rightfully blame some sweatshops, viz., some of those that make big profits, for wrongfully exploiting their workers. Moreover, many other sweatshops, especially ones that do not make huge profits, will be marked as non-exploitative in Wertheimer's approach.

2.2. *Improving the fairness account using the vulnerability criterion*

One way to improve (E1) is to add the "vulnerability" element to this definition to better distinguish cases like the porcelain collector from ones like sweatshops. This amendment would enable us to consider the background conditions and filter out cases in which the alleged exploitee is not necessarily vulnerable in a significant way to the transaction in question. The new definition of exploitation would look like this:

> (E2): A exploits B if and only if A benefits from a transaction with a vulnerable B, in which A takes advantage of this vulnerability.[6]

Robert Goodin and Thomas Christiano both have theories of exploitation that one can call vulnerability accounts. Goodin claims that "flagrant violation of duty to protect the vulnerable constitutes the essence of interpersonal exploitation" (Goodin 1987: 188). Moreover, Christiano defines exploitation as a violation of a duty to the vulnerable (Christiano 2015: 263). Both authors emphasize the vulnerability of the exploitee to the exploiter.

[6] (E2) is not an outcome-based account of exploitation. However, I still want to discuss it here, rather than in the next section, not because I consider it is an alternative outcome-based account, but because it is a helpful argumentative step to reach (E3).

The general criticism against vulnerability accounts is that they would create false positives, viz., mark many of our ordinary transactions as exploitative. For instance, Richard Arneson argues that there is nothing objectionable in using someone's vulnerability for advantage as long as there is no unfairness involved in the interaction (Arneson 2016: 10). After all, our specific vulnerabilities and necessities enable many of the transactions in which we get involved. For example, in Arneson's "Cancer Treatment" example, a cancer patient consults the only qualified surgeon in his town (Arneson 2016: 10). The surgeon offers to operate on the patient and save his life for a better-than-fair price. In this example, although the patient is vulnerable to the surgeon, the interaction is not exploitative. Of course, one might argue that this is what the Goodin-Christiano line is arguing for, i.e., the surgeon is not violating her duty of making a fair offer to the vulnerable patient. Again, however, we would fall back to the fairness account, which would explain why this transaction is not exploitative.

It seems plausible to combine the unfairness element with the vulnerability criterion to benefit from both criteria's strengths.

> (E3): A exploits B if and only if A benefits from a transaction with a vulnerable B, in which A takes unfair advantage of B's vulnerability.

It might seem evident that if A can exploit B, then something about B's situation already creates vulnerability on her account. Hence, adding vulnerability to the definition of exploitation might seem redundant. However, talking about an exploitative relationship for some types of vulnerabilities would not make sense. For example, assume I sell chess lessons online and get paid in terms of donation, viz., my customers pay me whatever they deem suitable as a fair price.[7] Now and then, some customers would pay a meager price or not pay at all while benefiting from the lessons. The situation fits (E3) well; I seem to be exploited by these customers. In this case, although I have made myself vulnerable to unfair transactions, I am not vulnerable in any special sense or any other sense regarding the background conditions of the parties involved. So, adding vulnerability as a separate criterion does not add much to the exploitation charge unless the definition of vulnerability is related to some background conditions.

Satz seems to agree that "*underlying* extreme *vulnerabilities* of the transacting parties" (Satz 2010: 97, emphasis in the original) rather than any random vulnerability can lead to exploitation. She adds that "widely varying resources or widely different capacities to understand the terms of their transaction" (Satz 2010: 97) cause such extreme vulnerabilities.[8] Therefore, the exploitation account must build the connection between the exploitee's background conditions and vulnerability to avoid false positives.

[7] A similar example is in Arneson (2016: 27).

[8] I thank an anonymous reviewer for reminding us of Satz's contribution to the discussion.

2.3. *Arneson's theory of exploitation*

Arneson gives an account of exploitation that would fit (E1) and include the background circumstances into the definition of unfairness. He formulates a prioritarian criterion of the best outcome regarding the distribution of social surplus in an interaction. He holds that the best outcome is obtained by measuring the weighted well-being of the parties involved in the transaction in question. Moreover, the weighted well-being score "is fixed in this way: obtaining a benefit (or avoiding a loss) for a person has more value, (1) the greater the well-being gain it achieves for the person, (2) the worse-off in lifetime well-being the person would otherwise be absent this benefit, and (3) the more deserving the person is in life-time terms" (Arneson 2016: 17). So, a transaction between parties A and B would be less fair the farther it is from the best outcome criterion that Arneson provides.

Arneson's theory of exploitation is superior to Wertheimer's in including the lifetime well-being of the parties when making an exploitation claim.[9] This criterion of how agents would do in lifetime well-being absent the transaction in question would imply a comparison in background circumstances. "This account will yield the result that Poor's driving hard bargain with Rich when Poor has a bargaining advantage is more fair than Rich's driving hard bargain with Poor when Rich has a bargaining advantage (on the assumption that greater wealth tends to lead to greater lifetime well-being)" (Arneson 2016: 18).

Nicholas Vrousalis criticizes Arneson for creating false negatives, i.e., missing to mark exploitative cases as exploitative. The case Vrousalis uses as a counterexample to Arneson's account is a version of Ant-Grasshopper cases.[10] In this version, although the Grasshopper is much worse-off compared to the Ant, she is undeserving to enjoy the benefits of their interaction because she spent all summer lazing around, and this fact makes her "completely undeserving (absolutely or comparatively, pluralistically or monistically)" (Vrousalis 2016: 535). Hence, according to Arneson's theory of exploitation, even if Ant charges the

[9] Wertheimer does not consider his account of exploitation to be weak just because it does not consider background circumstances (or overall well-being of the parties, for that matter) relevant to the charge of exploitation; he instead sees this fact as a virtue of it (see Wertheimer 2007: 261).

[10] "As in the fable, Ant works hard all summer and has ample provisions for the winter. Grasshopper lazes about and in January has an empty cupboard. As it happens, cardinal interpersonal comparisons of desert and well-being can be made. Without interaction, Grasshopper will end up with welfare level two, which amounts to dire misery, and Ant with three, bare sufficiency, and in this scenario Ant is comparatively more deserving; the gap between the welfare level Ant has and what he deserves is far greater for him than is the comparable gap for Grasshopper. Ant proposes to sell some provisions to Grasshopper at a very high price. Grasshopper accepts the deal, though he would prefer to pay less and get more. With this deal in place, Grasshopper ends up with welfare level three and Ant with twelve (Ant buys a cell phone). Even after this transaction, Ant's welfare level is less than he deserves, by comparison with the situation of Grasshopper" Arneson (2016: 535).

Grasshopper exorbitantly, she is not accused of exploitation because of Arneson's account's "desert" element.

Vrousalis also criticizes Arneson's account of exploitation for creating false positives, i.e., marking non-exploitative cases as exploitative. He compares cases of unfair free-riding that do not include domination to those that include domination (Vrousalis 2016: 536). Vrousalis maintains that Arneson's account marks these cases as exploitative, although they are intuitively not exploitative. He gives the example of someone, A, who escapes from persecution and hides in B's boat. When B rows her boat from one coast to another, A free-rides B's rowing efforts and hence takes unfair advantage of her for A's benefit. Vrousalis maintains that although A seems to be exploiting B according to Arneson's unfairness account, this is intuitively wrong.

Vrousalis gives another example where some villagers take unfair advantage of other villagers while at the same time dominating them. In this thought experiment, villagers take turns to stand sentry at the village's gates against bandits. Some villagers refuse to serve in the scheme, though, and thus free-ride the efforts of others. They do so knowing that the villagers who live close to the village's periphery will suffer the most when bandits attack. To defend the village safely, those contributing to the sentry scheme need the free riders' contribution. So, the free riders have power over the contributors. Vrousalis holds that in this example, the free riders can be rightly accused of exploiting the contributors' efforts to maintain safety.

Vrousalis also argues that even if Arneson were right that these free rider cases were both exploitative, something of normative significance would be lost by not distinguishing them from each other: "Only power-grounded advantage-taking constitutes exploitation, on this view" (Vrousalis 2016: 536).

Putting aside its theoretical handicaps, using Arneson's prioritarian criterion of fairness in (E1) must be tested to see if it renders sweatshops exploitative. This fairness account seeks to achieve the best weighted-well-being score in a transaction. For example, compare two possible wage levels that an MNE can pay to its workers in a sweatshop that it runs, w and w + x, where w is the current wage paid, and w + x is the wage that the MNE can pay without causing much of a change in other aspects of its budget.[11] Paying w + x to the workers would provide higher marginal well-being than the marginal well-being that paying w to workers would give to higher-level managers and consumers who would otherwise buy the same products slightly cheaper.

This calculation is accurate, at least in the short term. However, in the long term, paying w + x at the cost of lower manager compensation packets and a slightly higher product price may pull down wages paid to sweatshop workers. If further empirical work shows that this is the

[11] Arnold and Hartman count multiple possible ways MNEs can increase wages without much loss to their profits in Arnold and Hartman (2006).

case in the sweatshops in question, then the first criterion of Arneson's account would not agree that paying w + x creates greater well-being for workers. Whether the first criterion of Arneson's prioritarian fairness account supports paying w + x to sweatshop workers depends on empirical data relevant to the particular circumstances in question.

The second criterion seeks the distribution that will provide the best well-being in life-time terms. In line with the first criterion, sweatshop workers would lose much more than high-level managers and customers in general, and in lifetime terms, if w is paid to the workers instead of w + x.[12] This scenario would be valid unless the long-term consequences of higher worker wages drive profits down to drive the MNE in question down in the competition among the regional companies. Such a consequence would result in many sweatshop workers losing their jobs and thus having less well-being in lifetime terms. Hence, whether Arneson's account's second prioritarian criterion supports higher worker wages also depends on the pertinent empirical data.

The third criterion is more challenging to add to the calculation. This difficulty would be due to the ambiguity of comparing the "desert" element in this account of fairness to the other two criteria. The main determining factor in the Marxist and liberal divide in the approach to exploitation stems from their economic approaches to the calculation of desert.[13] According to the liberal approach to desert, the supply and demand curves of labor power determine the wage laborers deserve in that particular market. So, according to the liberal approach, if the supply and demand curves depict wage w as the equilibrium point, then the workers deserve to be paid w instead of w + x. Alternatively, if there is room in the liberal theory of desert, as some authors claim, to include w + x in the deserved wage interval, then the third criterion would hold that the fair wage for sweatshop workers is w + x and not w.[14]

As a result, according to Arneson's theory of exploitation, whether sweatshop labor is exploitative depends on empirical evidence. There is evidence in both directions.[15] Some evidence shows that paying higher

[12] It is plausible that relatively less well-off consumers also consume sweatshop products. However, it is a reasonable assumption that the primary consumers of these products are individuals in more affluent Western countries. If that were not the case, there is a chance that Arneson's second criterion does not support paying w + x to sweatshop workers. Whether this second criterion would support paying w + x would depend on comparing the change in the well-being between poor sweatshop workers and poor consumers after the change in the wage level.

[13] Jon Elster explains the difference in plain economic terms in Elster (1978: 3–17).

[14] Sollars and Englander (2007), who argue against increasing the minimum wage paid to sweatshop workers on moral grounds, also concede that there is room for firms to increase the market-level wages up to a point.

[15] Both Kates (2015) and Coakley and Kates (2013) give evidence to support a moderate to no effect of higher worker wages on worker layoff. On the other hand, Sollars and Englander (2007) point to other economic literature which shows that increasing minimum wage levels may lead to worker layoffs.

wages to sweatshop workers will lead to worker layoffs in that sweatshop. However, even in this case, Michael Kates argues that this is not necessarily harmful to less developed countries' sweatshop workers overall. According to him, some layoffs due to wage increases might bring greater benefits for the less developed countries' sweatshop workers (cf. Kates 2015).

There is empirical evidence favoring higher wages paid to sweatshop workers, leading to their greater well-being. The same evidence favors the idea that the overall population working in sweatshops would be worse off *in a lifetime* without a higher wage. These two premises are pertinent to Arneson's first two criteria in his prioritarian approach to exploitation. Moreover, even theoreticians working within the limits of classical liberal theory concede that some increase in the market-determined wage level is possible. So even if we accept the liberal economic criterion of desert for the third element of Arneson's theory of exploitation, the theory concedes that paying a wage level (w) below the economically feasible maximum (w+x) would plausibly assign the charge of exploitation to the MNE managers.

Therefore, there is some support in the relevant literature to claim that Arneson's account of exploitation would find some standard cases of sweatshops to be exploitative while marking others as non-exploitative.

3. *Exploitation and attitudes*

The accounts I will investigate in this section claim that exploitation is unrelated to how the parties in a transaction distribute the resulting surplus. These accounts question whether there is an attitude-based wrong in the type of interaction in question.

3.1 *Sample's theory of exploitation*

Sample presents her attitude-based account of exploitation as follows: "The basic idea is that exploitation involves interacting with another being for the sake of advantage in a way that degrades or fails to respect the inherent value in that being" (2003: 57). So, according to her account, A exploits B if and only if A benefits from interaction with B while degrading or disrespecting B.

> (E4): A exploits B if and only if A benefits from a transaction with B in which A degrades or disrespects B.

She provides three possible ways A can disrespect B: 1) A can fail to respect B by neglecting what is necessary for B's well-being or flourishing, 2) A can fail to respect B by taking advantage of an injustice done to B, and 3) A can fail to respect B by commodifying or treating as a fungible object of market exchange, an aspect of B's being that ought not to be commodified (Sample 2003: 57).

Sample asserts that respect for other persons does not require us to love them, and in this sense, respect is a limited relationship with

others. Just as respect requires us to engage in a limited but positive manner, she argues; the duty to refrain from exploiting others requires us to constrain ourselves in specific ways when interacting with them. So it is not enough not to harm the person we interact with; ignoring their needs is also as disrespectful as harming them. Such an act of ignoring the needs of others can be exploitative if we also benefit from the interaction, Sample maintains. Moreover, this is why she argues that exploitation is worse than neglect: We fail to fulfill our duty to respect the inherent value of our interactor when we are involved in exploitation, even if it is mutually beneficial.

Ignoring the basic needs of someone whose needs are at stake is acting disrespectfully towards that person, holds Sample.[16] So if A chooses to interact with B and B's basic needs are at stake, B is vulnerable to A. From this moment on, if A does not meet B's basic needs, A exploits B when A also benefits from the interaction. Sample admits that if, in the only possible mutually advantageous interaction, the basic needs of B cannot be met but are taken into account, then the interaction is not exploitative (2003: 75).[17]

The needs of people are related to their well-being, according to Sample. Therefore, when interacting with someone vulnerable, we must consider their well-being. The understanding of well-being that Sample bases her account of exploitation on is the capabilities approach advocated by Martha Nussbaum. Hence "exploitative interactions are those in which the capabilities of our interactors are ignored in the pursuit of our own advantage" (Sample 2003: 81). Still, "nonexploitation does not require that we ensure the capabilities of our interactors in individual transactions. Rather it requires that we in some way take their needs into account" (Sample 2003: 81). Hence, Sample reminds us that not ignoring the capabilities of our interactors means taking their needs into account.[18]

> In cases of mutually beneficial exploitation, the exploited person either does not benefit sufficiently from the transaction that results in her exploitation or, in some other way treated as having less value than she actually possesses. Either the resources obtained from the interaction fail to contribute

[16] Sample does not explain what she means by basic needs being "at stake," but a charitable reading could suggest that she is talking about transacting with someone whose basic needs are unmet or would be unmet after the transaction.

[17] This second criterion softens the one that comes before it. According to Sample's theory, a transaction between two people, both of whom lack their basic needs, would not necessarily be exploitative as long as they consider each other's basic needs. This criterion is attitude-based, although it seems to worry about the outcomes. One can think of cases where the fair transaction leaves at least one of the parties without enough provision to meet their basic needs. Sample's account assigns a duty to the transacting parties, in such cases, to go beyond what is required by fairness and have a particular attitude towards the other party, viz., one that fulfills their basic needs.

[18] Sample is probably talking about the needs relevant to contributing to these capabilities here.

> to that person's capabilities in a relevant way, or else the nature of the transaction itself—as opposed to the resources or social surplus of the transaction—degrades her. (Sample 2003: 81)

Such is different for repeated transactions, however. Sample argues that in the case of repeated transactions, where A is taking advantage of B and B is vulnerable to A, B's capabilities must be ensured: "If an employer fails to compensate an employee in a way that provides her with adequate income when such compensation is possible, then the relationship is exploitative" (Sample 2003: 81).

Sample does not cash out the exact circumstances under which such compensation is possible for the employer. Thus, her account needs to be more explicit about how much compensation an employer of a local sweatshop that makes quite a small profit owes to their workers. It is more evident, though, that managers of MNEs ought to fulfill the capabilities criterion of Sample's account. If such compensation were not possible for large MNEs, it would not be possible for any organization because of the high-profit range of these enterprises.

Sample concedes that her account of exploitation bears the result that since exploitation is a violation of a duty to respect others, agents might prefer not to interact whenever the interaction would be exploitative, even if it is mutually advantageous. In her account, avoiding interaction is permissible, and she clearly emphasizes that exploitation is worse than neglect. On the other hand, Sample notes that if an agent "always" prefers to neglect when the other option is mutually advantageous exploitation, then this agent does not fulfill the imperfect duty of beneficence and is morally blameworthy (Sample 2003: 72).

According to Sample, her account of exploitation explains why we take exploitative interactions to be worse than neglect. She believes that interacting with someone else burdens us with special duties towards that person and equips our interactor with specific claims against us. She maintains that this aspect of her account is in line with the intuition that killing is worse than letting die (Sample 2003: 61).

Another advantage of her account, she claims, is that it can explain exploitative systems besides exploitative transactions. In such systems, exploitative behavior can be part of the routine and be accepted as ordinary, yet the exploitation claim should not be waived. In these systems, the exploitee might not *feel* exploited, even when they indeed are.

Like other theories of exploitation, Sample's account has faced criticisms and counterarguments. Wertheimer criticizes all the three accounts Sample gives of why an interaction might be exploitative. For the first account of exploitation, Wertheimer draws on Sample's example of the teller. Sample claims that I would exploit the teller if I interacted with her to profit from this interaction but still tolerated the wages upon which she could not decently live (Sample 2003: 69). Wertheimer disagrees. He maintains that "it is not clear why the mere fact that A enters into an arguably limited transaction with B requires A to be quite so responsive to B's life needs" (Wertheimer 2007: 216).

For Sample's second account, Wertheimer gives the example of a person, B, who has recently suffered a malicious attack on her home. She interacts with a carpenter, A, to get her home fixed. According to Wertheimer, there is nothing wrong if A takes advantage of the interaction with B as long as A does not take unfair advantage, viz., A does not charge B an exorbitant price for the work. In this example, although B has suffered an injustice, there is nothing wrong with A taking advantage of the interaction with B.

Against Sample's third account, Wertheimer gives the example of a female sex worker, B, who sees herself as a professional who interacts with A, someone who seeks B's services. In this example, B knows how A regards her but does not care about it and even finds A pathetic for his needs. Wertheimer argues that B is not exploited in her interaction with A.

Leaving the possible handicaps of her account of exploitation aside, for the time being, Sample gives a detailed explanation of the first type of disrespect in her definition of exploitation, viz., one in which A fails to take B's well-being into account in their interaction.[19] Her criticism of Wertheimer's theory of exploitation was that the criterion of a hypothetical market misses marking sweatshop jobs as exploitative. Moreover, she argues that these jobs are exploitative because of her first account of disrespect in her definition of exploitation.

> [...] There is little doubt that much global trade today involves interacting with people on exploitative terms. Even if, as defenders of globalization argue, expanded trade improves the situation of many or all of those in developing countries, the terms of such trade may be inadequate for meeting their basic needs and generally demonstrating respect for their personhood. (Sample 2003: 169)

Exploitation as degradation or disrespect, as in (E4), can explain the exploitation of sweatshop workers at the bottom of the global production chain. However, besides its theoretical flaws, the degradation account focuses only on the worst-off to miss the exploitation of rather better-off workers, whose basic and serious needs have been met. This theoretical choice might or might not be a drawback of (E4) depending on whether a better interpretation of disrespect that covers the adverse circumstances of better-off workers is given.

3.2. *Vrousalis' theory of exploitation*

Another attitudinal account of exploitation, which might cover the plight of the better-off workers as well, is given by Vrousalis. He gives a domination account of exploitation. He believes exploitation is "a form of domination for self enrichment" (Vrousalis 2013: 1). He adds that when A dominates B, "[a] necessary condition for domination is power-induced injury to B's status or some form of servitude on B's part"

[19] She barely expands the second account and admits that some implications of the third account are open to discussion (like the case of sex work).

(Vrousalis 2016: 529). Like Sample, Vrousalis denies that unfairness in the division of social surplus created by the parties' interaction has anything to do with exploitation. So his definition can be formulated as follows:

> (E5): A exploits B if and only if A benefits from a transaction with B in which A dominates B.

Arneson has a handful of counterexamples against the domination account. He argues that B is vulnerable to A in each of these examples, and A's behavior towards B can reasonably be called domination. Nevertheless, although A enriches herself, none of these examples correspond to our intuitions about exploitation. One such example is what he calls the "Utility Company" (Arneson 2016: 10). In this example, a utility company is the monopoly heat supplier in a town with a cold climate. The residents have no choice but to buy heat from this company. The company charges the residents fairly and makes a profit from the sales. According to Arneson, although the heat company might be said to dominate the residents, there is no exploitation in this case *because* the company charges them fairly.

The defender of (E5) could respond to Arneson, holding that the residents are not exploited in the "Utility Company" example because they are not experiencing power-induced injury to their status, nor are they going through any form of servitude under the heat supplier. However, these defenders would need to show how and when someone would experience these forms of domination apart from being charged unfairly in an exchange. This is what Vrousalis claims to be doing. He also accuses the defenders of the unfairness view, (E1) or (E3), of confounding unfair treatment with exploitation. "On the fairness view, by contrast, 'unfair treatment' and 'exploitation' are used interchangeably. What is the extra purchase of saying that A exploits B, on this view? Arneson's answer is 'not much.' Indeed, he uses 'unfair treatment' and 'exploitation' as synonymous throughout his essay [...]" (Vrousalis 2016: 537).

Vrousalis argues the distinct wrong that (E5) points should be analyzed under a specific category. Thus such cases would need a different response than those covered by unfairness views like (E1) or (E3).

> I claim that there is a concept distinct from Arneson's, call it *shmexploitation*, whose contours are defined by the domination view, which takes cases like *Pit* and *Ant and Grasshopper* as instances of wrongful advantage-taking. If I am right about these instances, then shmexploitation captures instances of wrongful advantage-taking that are surplus to exploitation: shmexploitation is explanatorily superior to exploitation in that respect. We should think of cases like *Pit* in terms of shmexploitation, not exploitation. (Vrousalis 2016: 537, original italics)[20]

[20] Here is Vrousalis' example of the Pit: "A and B are alone in the desert. A finds B lying at the bottom of a pit. A proposes to extract B, on condition that B works for A for a wage of $1/day for the rest of B's life" (Vrousalis 2016: 527).

According to (E5), MNE managers exploit sweatshop workers because they have dominating power over the workers. According to this power relationship, MNE managers can get sweatshop workers to agree to the managers' terms. That is why such a relationship creates some servitude for the workers.

One alleged drawback of Sample's theory of exploitation was its lack of explanatory power for the plight of better-off workers. Vrousalis' theory can explain their situation as exploitative due to their employers' dominating power over them. This result does not imply that sweatshop workers and better-off workers are exploited to the same degree. Since it is plausible to claim that domination comes in degrees, it follows that exploitation as domination also comes in degrees. Hence it is consistent to expect a correlative degree of moral indignation towards different degrees of exploitation.

So, besides explaining the exploitation of sweatshop workers at the bottom of the supply chain, (E5) can explain the plight of better-off workers who are also claimed to be exploited. According to this definition, managers exploit these better-off workers as long as they have no say in their work conditions.

The domination account of exploitation is better equipped to mark many standard cases of sweatshops as exploitative. However, what seems to be an advantage of the domination account becomes a disadvantage in sweatshops. The domination account can place sweatshops on a scale marking them as more or less exploitative. Nevertheless, many sweatshop critics refer to the concept of exploitation to determine with which workplaces to interfere. If the domination account marks all labor as exploited, we still need a further criterion to determine which workplaces are "exploitative enough" to deserve a proper protest and interference. Until such a criterion is defined, the domination account is of little use to sweatshop critics.

4. *Conclusion*

Exploitation is one strong reason to protest and interfere with sweatshops. However, reviewing the prominent theories of exploitation proves that each theory includes some theoretical flaws. For this reason, it is not straightforward to point at the correct definition of this concept that we can apply to all sweatshops. Moreover, each prominent exploitation theory proves that only some sweatshops are exploitative while missing to mark some other intuitively exploitative cases as exploitative.

This article opposes the contention that all standard cases of sweatshops are exploitative according to a robust theory of exploitation. I have gone through the most prominent theories in the literature to mark their theoretical flaws and strengths and evaluate their verdict on whether the standard cases of sweatshops are exploitative. I have shown that each of these theories marks at least some sweatshops as exploitative while missing to mark others.

In conclusion, my argument shows that the exploitation charge alone is not a solid moral ground to interfere with all sweatshops. Going over the prominent theories of exploitation gives us two results. First, there is no theory of exploitation without any theoretical flaws. All these theories either miss to mark some intuitively exploitative cases as one or wrongly mark an intuitively non-exploitative case as exploitative. Some theories need clarity in matching their claims with empirical data. Other theories have some ambiguity in their definitions and how they use concepts.

Second, the best of these prominent theories mark sweatshops as exploitative under some circumstances, while not in others. So, even if one were to pick their favored theory of exploitation as the right one and try to defend it against counterarguments, it might still not ascertain that all standard sweatshop cases are exploitative according to this theory.[21]

This conclusion leaves sweatshop critics, who wish to charge sweatshops with exploitation, with two possible paths. First, one can pick and defend a theory of exploitation against others and then apply it to a particular case to see whether that case is exploitative. Alternatively, one can pick a particular case and try to see how these prominent theories mark the case on the exploitation scale, later to take the case as exploitative if enough (or just one or all) of the theories agree on it.

I will not argue in favor of any one of these methods here. What I have done until this point has demarcated the theoretical landscape for finding out what kind of sweatshop cases are marked as exploitative according to these prominent theories. One has to find a method to follow to discover whether a particular case is exploitative. If the case turns out to be exploitative, then this constitutes a *prima facie* reason to interfere with the sweatshop in question. In the end, sweatshop critics had better find other reasons *besides* the charge of exploitation to protest or interfere with these workplaces.

References

Arneson, R. 2016. "Exploitation, Domination, Competitive Markets, and Unfair Division." *The Southern Journal of Philosophy* 54: 9–30.

Arnold, D. and Hartman, L. 2006. "Worker Rights and Low Wage Industrialization: How to Avoid Sweatshops." *Human Rights Quarterly* 28: 676–700.

[21] Take Sample's theory of exploitation, for instance. Even in such a tight theory, we can imagine a sweatshop run by a local individual who is not significantly wealthier than their workers. We can further imagine that this sweatshop is run in a country that is not poor necessarily because of past injustice but because it does not attract enough attention from wealthy investors. Her exploitation criteria might not necessarily pinpoint this workshop as exploitative, although another definition of sweatshops can mark this place as one.

Christiano, T. 2015. "What Is Wrongful Exploitation?" In D. Sobel, P. Vallentyne and S. Wall (eds.). *Oxford Studies in Political Philosophy, Volume 1*. Oxford: Oxford University Press, 250–275.

Coakley, M. and Kates, M. 2013. "The Ethical and Economic Case for Sweatshop Regulation." *Journal of Business Ethics* 117 (3): 553–558.

Cohen, G. A. 1983. "The Structure of Proletarian Unfreedom." *Philosophy and Public Affairs* 12 (1): 3–33.

Elster, J. 1978. "Exploring Exploitation." *Journal of Peace Research* 15 (1): 3–17.

Goodin, R. 1987. "Exploiting a Situation and Exploiting a Person." In A. Reeve (ed.). *Modern Theories of Exploitation*. London: Sage Publications, 166–200.

Kates, M. 2015. "The Ethics of Sweatshops and the Limits of Choice." *Business Ethics Quarterly* 25 (2): 191–212.

Powell, B. 2014. *Out of Poverty: Sweatshops in the Global Economy*. Cambridge: Cambridge University Press.

Sample, R. J. 2003. *Exploitation: What It Is and Why It's Wrong*. Lanham: Rowman and Littlefield.

Satz, D. 2010. *Why Some Things Should Not Be for Sale: The Moral Limits of Markets.* New York: Oxford University Press.

Snyder, J. 2010. "Exploitation and Sweatshop Labor: Perspectives and Issues." *Business Ethics Quarterly* 20 (2): 187–213.

Sollars, G. G. and Englander, F. 2007. "Sweatshops: Kant and Consequences." *Business Ethics Quarterly* 17 (1): 115–133.

Sollars, G. G. and Englander, F. 2018. "Sweatshops: Economic Analysis and Exploitation as Unfairness." *Journal of Business Ethics* 149 (1): 15–29.

Vrousalis, N. 2013. "Exploitation, Vulnerability, and Social Domination." *Philosophy and Public Affairs* 41 (2): 131–157.

Vrousalis, N. 2016. "Exploitation as Domination: A Response to Arneson." *The Southern Journal of Philosophy* 54 (4): 527–538.

Wertheimer, A. 1996. *Exploitation*. Princeton: Princeton University Press.

Wertheimer, A. 2007. "Ruth J. Sample, Exploitation: What It Is and Why It's Wrong." *Utilitas* 19 (2): 259–261.