

# dialectica

International Journal of Philosophy

## Contents

JARED WARREN, <i>Gruesome Counterfactuals</i> . . . . .	314
ARINA PISMENNY, <i>When Is Jealousy Appropriate?</i> . . . . .	331
BRYCE GESSELL, <i>The Legend of Hermann the Cognitive Neuroscientist</i> . . . . .	359
MARIA SEKATSKAYA & GERHARD SCHURZ, <i>Alternative Possibilities and the Meaning of ‘Can’</i> . . . . .	379
DEBORAH RAIKA MÜHLEBACH, <i>Neopragmatist Inferentialism and the Meaning of Derogatory Terms—A Defence</i> . . . . .	409
GUIDO MELCHIOR, <i>Sensitivity and Inductive Knowledge Revisited</i> . . . . .	441
ULRICH SCHWABE, <i>Review of Imhof (2014)</i> . . . . .	463
TONI RØNNOW-RASMUSSEN, <i>Review of Lutz (2016)</i> . . . . .	469

# dialectica

International Journal of Philosophy

Official Organ of the European Society of Analytic Philosophy

founded in 1947 by Gaston Bachelard, Paul Bernays and Ferdinand Gonseth

## Editorial Board

Jérôme Dokic, EHESS, Paris, France

Pascal Engel, EHESS, Paris, France

Manuel García-Carpintero, Universitat de Barcelona, Spain

Diego Marconi, Università di Torino, Italy

Carlos Moya, Universitat de València, Spain

Martine Nida-Rümelin, Université de Fribourg, Switzerland

François Recanati, Collège de France, Paris

Marco Santambrogio, Università degli Studi di Parma, Italy

Peter Simons, Trinity College Dublin, Ireland

Gianfranco Soldati, Université de Fribourg, Switzerland

Marcel Weber, Université de Genève, Switzerland

## Editors

Fabrice Correia, University of Geneva

Philipp Blum, University of Lucerne

Sharon Casu, University of Fribourg (managing editor)

## Review Editors

Stephan Leuenberger and Philipp Blum

## Editorial Committee

Philipp Blum (né Keller), Claudio Calosi, Fabrice Correia, Matthias Egg, Martin Glazier, Aleks Knoks, Jörg Löschke, Arturs Logins, Giovanni Merlo, Robert Michels, Ryan Miller, Jonathan Mitchell, Paolo Natali, Donnchadh O'Conaill, François Pellet, Edgar Phillips, Jan Plate, Stephanie Rennick, Mario Schärli, Thomas Schindler, Marco R. Schori, David Schroeren, Mike Stuart, Daniel Vanello.

## Consulting Board

Johannes Brandl (Salzburg), João Branquinho (Lisboa), Elke Brendel (Bonn), Ingar Brinck (Lunds), Eros Corazza (Ikerbasque and Carleton), Josep Corbi (València), Michael Esfeld (Lausanne), Dagfinn Føllesdal (Stanford and Oslo), Frank Jackson (Australian National University, Canberra), Max Kistler (Paris I), Max Kölbel (Wien), Jan Lacki (Genève), Karel Lambert (Irvine), Paolo Leonardi (Bologna), Fraser Macbride (Manchester), Josep Macià (Barcelona), Genoveva Martí (Barcelona), Élisabeth Pacherie (Institut Jean Nicod, Paris), David Piñeda (Girona), Wlodek Rabinowicz (Lund), Barry Smith (Buffalo), Thomas Strahm (Bern), Christine Tappolet (Montréal), Neil Tennant (Ohio State), Mark Textor (King's College London), Achille Varzi (Columbia University), Alberto Voltolini (Torino), Timothy Williamson (Oxford).

September 2021

## Contents

JARED WARREN, <i>Gruesome Counterfactuals</i> . . . . .	314
ARINA PISMENNY, <i>When Is Jealousy Appropriate?</i> . . . . .	331
BRYCE GESSELL, <i>The Legend of Hermann the Cognitive Neuroscientist</i> . . . . .	359
MARIA SEKATSKAYA & GERHARD SCHURZ, <i>Alternative Possibilities and the Meaning of ‘Can’</i> . . . . .	379
DEBORAH RAIKA MÜHLEBACH, <i>Neopragmatist Inferentialism and the Meaning of Derogatory Terms—A Defence</i> . . . . .	409
GUIDO MELCHIOR, <i>Sensitivity and Inductive Knowledge Revisited</i> . . . . .	441
ULRICH SCHWABE, <i>Review of Imhof (2014)</i> . . . . .	463
TONI RØNNOW-RASMUSSEN, <i>Review of Lutz (2016)</i> . . . . .	469



# Gruesome Counterfactuals

JARED WARREN

One of the most popular answers to the grue puzzle appeals to counterfactual dependence. An observed green emerald is grue. But while it still would have been green even if unobserved, if unobserved it would not have been grue. Because of this, or something like it, we can project “green” but not “grue” to the unobserved emeralds. Counterfactual theories have been offered by Frank Jackson and Peter Godfrey-Smith, among others. But there is a worry that all counterfactual approaches to grue fail for the same reason—counterfactual symmetry. The grue theorist can endorse symmetrical counterfactuals: an observed green emerald would have been grue but not green if unobserved. It then seems that counterfactual responses to grue beg the question. Here I argue that there are two ways to understand this challenge and that they both fail, but for different reasons. I close by drawing some general lessons about philosophical fair play regarding the twentieth century’s many broadly semantic, broadly skeptical challenges—grue, quus, gavagai, and the like.

## 1 Counterfactuals and Grue

All of the emeralds we’ve observed have been green. They’ve also been grue—either green and observed or blue and unobserved. Yet we take our observations to support or confirm the hypothesis that all emeralds are green, but not the hypothesis that all emeralds are grue. We inductively project the “green” predicate, but not the “grue” predicate. The new riddle of induction is roughly the challenge of explaining and vindicating these inductive policies.<sup>1</sup>

One popular strategy for answering the riddle appeals to *counterfactual* differences between green and grue. For a given observed green emerald, *e*, the following counterfactuals seem true: Observed(*e*) Green(*e*)

(1)  $\neg$  Observed( $\$e$ )

---

<sup>1</sup> The riddle derives from Goodman (1946, 1955).

(2)

 $\} \{ \#one \} \neg \text{Observed}(e) \square \rightarrow \text{Green}(e)$ (II)  $\neg \text{Observed}(e) \square \rightarrow \neg \text{Grue}(e)$ 

If  $e$  hadn't been observed, it still would have been green. But if it hadn't been observed, it wouldn't have been grue. On this strategy, counterfactuals like these are key to why projecting "green" is warranted but projecting "grue" is not.

The counterfactual strategy comes in different forms. Not all of them appeal to (1) and (2).<sup>2</sup> And not all that appeal to (1) or (2) do so in the exact same way. For some it is our *knowledge* of such counterfactuals that is key.<sup>3</sup> For others it is enough that we *believe* such counterfactuals.<sup>4</sup> For still others, the *truth* of (1) and (2) is what matters.<sup>5</sup> These differences are important, and will be relevant at several points below, but I won't belabor them. It has been claimed that *every* counterfactual approach to grue fails for the same reason—counterfactual symmetry.<sup>6</sup> I will respond to the challenge here, but obviously, even if my response is successful, counterfactual responses to grue might fail for other reasons.

## 2 Counterfactual Symmetry

The grue theorist can respond to counterfactual approaches by rejecting (1) and (2) and instead accepting:

(III)  $\neg \text{Observed}(e) \square \rightarrow \neg \text{Green}(e)$ (IV)  $\neg \text{Observed}(e) \square \rightarrow \text{Grue}(e)$ 

Given that the antecedent is possible— $\diamond \neg \text{Observed}(e)$ —(1) and (III) are incompatible, as are (2) and (IV). The challenge is roughly that if we can vindicate our practices by appealing to (1) and (2), the grue theorist can vindicate gruesome practices by instead appealing to (III) and (IV). The exact nature of the appeal will, of course, depend on the details of the counterfactual

<sup>2</sup> See Freitag (2015, 2016) for a counterfactual approach that doesn't appeal to (1) or (2).

<sup>3</sup> See Jackson (1975).

<sup>4</sup> See Schramm (2014) and Okasha (2007). In Jackson and Pargetter (1980) justified belief is appealed to.

<sup>5</sup> See Godfrey-Smith (2003, 2011).

<sup>6</sup> See Roskies (2008) and Dorst (2016, 2018); Schwartz (2005) makes some related points.

account on offer. Taken generally, the claim is that the gruesome practice is symmetrical to our non-gruesome practice, in the epistemically relevant way.

Some counterfactual theorists will be fine with this. Alfred Schramm merely attempts to show that at most one out of “all emeralds are green” and “all emeralds are grue” is confirmed by *our* evidence.<sup>7</sup> And Samir Okasha’s position makes warrant for an inductive inference relative to our beliefs.<sup>8</sup> If you believe (1) and (2), you are warranted in concluding that all emeralds are green. If you instead believe (III) and (IV), you are warranted in concluding that all emeralds are grue. These versions of the counterfactual strategy aim for a somewhat limited conclusion. They want to show that projecting “green” but not “grue” is warranted *for us*, given *our beliefs*.

This is compatible with gruesome practices being warranted relative to alternative beliefs. Admitting this is not paradoxical, nor need it collapse into epistemic relativism. Nearly everyone accepts this kind of distinction. Different background beliefs warrant different conclusions. The kind of epistemic warrant being used here is broadly *internalist*; it is based on factors internal and accessible to the agent, such as their beliefs. Call this kind of internalist epistemic warrant *justification*. There is also an externalist kind of epistemic warrant, call it *entitlement*. Entitlement can depend on factors completely outside of and inaccessible to an agent. In the case at hand, perhaps the mere truth of (1) and (2) means that we are entitled to “green” projections, but not “grue” projections.

So both justification and entitlement to “green” projections can be secured using counterfactual strategies. This is significant; most attempted replies to grue don’t even get this far. But while it certainly shouldn’t be dismissed, we may hope for a bit more. On the internalist front, we might hope for non-question-begging arguments that beliefs in (1) and (2) are justified, while beliefs (2) and (IV) are not. On the externalist front, we might hope for non-question-begging arguments that (1) and (2) are true, while (2) and (IV) are not. These and other hopes seem to hinge on our being able to successfully defend a counterfactual theory based on (1) and (2) against a gruesome counterfactual theory based on (2) and (IV).

Can counterfactual attempts to solve grue successfully rebut the challenge of counterfactual symmetry? This is our question. To answer it, we’ll need to

---

7 In Schramm’s (2014) terminology, people who believe (III) and (IV) while disbelieving (1) and (2) have *different evidence* than people—like us—who have the opposite beliefs.

8 See Okasha’s (2007) commentary on Jackson (1975).

distinguish between two different ways of pushing the symmetry challenge. I think that both ways fail, but they fail for different reasons.

### 3 The Epistemological Version

Goodman originally defined “grue” as being either green and observed before *t*, or blue and unobserved before *t*. Above, I followed the common practice of simply assuming that *t* is *now*. Either way, the definition of “grue” involves a temporal element. A natural first thought is that this temporal element is what defeats the projection of “grue”. In response to this, Goodman introduced a parallel term, “bleen”, meaning observed and blue or unobserved and green. He then noted that “green” and “blue” are definable using “grue” and “bleen” (Goodman 1955). In a language that starts with “grue” and “bleen”, it is “green” and not “grue” that has an explicit temporal element in its definition. This shows that whether a term’s definition includes a temporal element—and even whether a term is defined at all—is language-relative.

This point is much cleaner than the counterfactual symmetry point. All Goodman needed to show was that a definition existed for “green” that included an explicit temporal element. The correctness of Goodman’s definition is common ground in the debate. Some have doubted the possibility of a grue/bleen mother tongue, and others have pointed to lingering temporal asymmetries between “grue” and “green”, but nobody challenges the accuracy of Goodman’s definition of “green” in terms of “grue” and “bleen”. The definitional facts are agreed upon by all parties.

Not so for the counterfactual facts. Counterfactual claims are about the world and its features. This match *would* light, if struck. This sugar cube *would* dissolve, if placed in water. That star *would* collapse into a black hole, if it were twice as massive. Those who disagree with us about the world and its features may well disagree with us about which properties are independent of our observational procedures. So unlike grue theorists who adopt Goodman’s symmetrical definitions, grue theorists who endorse symmetrical counterfactuals disagree with us about *matters of fact*.

The question is whether there is truly symmetry here and, if so, what kind? One idea is that there is *epistemological* symmetry. Adina Roskies (2008) has pushed something like this form of the counterfactual symmetry point, calling it the problem of “counterfactual robustness”. Her direct target is Jackson’s theory, which required knowledge of counterfactuals like (1) and (2).



Roskies claims that any route to this knowledge is question-begging against proponents of (III) and (IV).<sup>9</sup>

This version of the counterfactual symmetry challenge isn't merely the original grue puzzle in another guise, at least on the most natural understanding. It is instead akin to a *skeptical* challenge based on grue. It asks us how we know that the world isn't radically different than we think it is. How do we know that color features don't depend on observations? The challenge is now to vindicate our overall picture of physical reality, not just to defend some of our local inductive policies.

Of course, from the earliest presentations, Goodman tied the grue puzzle to issues of laws and counterfactuals. So you might question my claim that the epistemological symmetry challenge substantively differs from the original grue puzzle. I agree that in its more expansive formulations, Goodman's puzzle concerns not only induction proper but also more general explanatory reasoning, including abduction or "inference to the best explanation." This point was perhaps first clearly made in the literature by John Moreland in an excellent but little-noticed discussion published in 1976:

What is misleading in Goodman's formulation of the Riddle is that it mixes questions of induction with questions of abduction. It is not just a question of which [hypothesis] to project. We have seen that in the appropriate circumstances either might be projected. We wish normally to reject [the gruesome hypothesis] out of hand (regardless of the evidence) because in most situations [the gruesome hypothesis] would not be accepted as an explanatory hypothesis; and this is a question of abduction, not induction. [...] it does seem important to distinguish between the question of whether or not [the gruesome hypothesis] is to be projected in a given situation and the question of whether or not [the gruesome hypothesis] would ever be formulated as an explanatory hypothesis and, thus, made a candidate for inductive confirmation. [Moreland (1976), 376.]<sup>10</sup>

In the case of Jackson's counterfactual theory, a division like this falls out quite naturally. We start with the question of which predicates we can project,

---

<sup>9</sup> A related argument is in Schwartz (2005), targeting Godfrey-Smith (2003).

<sup>10</sup> I tracked down and read Moreland's paper after first reading a detailed synopsis of it in Stalker's incredible annotated bibliography of the first fifty years of grue literature, found in Stalker (1994).

but answering that question involves an appeal to background knowledge of counterfactuals. This background knowledge is what is questioned by Roskies. There are at least two questions here, and they are not the same. Proponents of counterfactual theories are independently committed to distinguishing them.

In fact, *everyone* must distinguish between questions of induction and questions of abduction, not just counterfactual theorists.<sup>11</sup> So this isn't an *ad hoc* move of desperation in the face of refutation. Nor does distinguishing these questions mean that a unified inductive logic, covering all reasonable non-deductive reasoning, is impossible.<sup>12</sup>

With this distinction noted, Jackson can plausibly explain how it is that we know (2) (which is what, on his account, blocks the projection of "grue"). The overall answer is likely a very long story.<sup>13</sup> In short: an extended process of observation, induction, deduction, and—most crucially—abduction led to our overall theory of the natural world. This overall scientific story entails (2), so knowing this, our overall theoretical knowledge transfers from our background theory to (2). Unless *everything* is *always* up for grabs, it is perfectly legitimate to appeal to our fundamental beliefs about the natural world when evaluating some particular inductive inference involving a newly introduced predicate. No question is begged in the process. Toward the end of her discussion, Roskies herself indicates openness to this type of reply to her challenge. She says her goal was only to show that a Jackson-style counterfactual account required supplementation.

I don't disagree completely, but we should put the point somewhat differently. We should say that Jackson's account of projection is fine as it stands, but add that it appeals to background knowledge that must itself be explained, in the long run. That explanation will involve not the original anti-grue reasoning, but instead general explanatory reasoning about the world, so there is

- 
- 11 In addition to the introduction of the "IBE" terminology, Harman (1965) argued that enumerative induction should be understood using inference to the best explanation. This is either unacceptable or compatible with the point I'm making here, depending on exactly how the claim is understood.
- 12 van Fraassen (1989) has argued that IBE contradicts conditionalization, and so cannot be integrated into a standard Bayesian framework. But his argument is based on an implausible way of combining Bayesianism with explanatory reasoning. For a better strategy for integration, see Huemer (2009) and Weisberg (2009).
- 13 I believe Jackson was always aware of this. Douglas Stalker told me that Jackson once told him that a full development and defense of his (Jackson's) approach to grue would take a very long book, not just a paper.

no circularity. If general explanatory considerations tell against the overall grue position, including the alternative counterfactuals, then Jackson has an answer to the symmetry challenge.<sup>14</sup>

It is worth noting that in requiring *knowledge* of counterfactuals like (2), Jackson's account is extremely demanding.<sup>15</sup> Every other counterfactual theory of projection requires much less of us. This is important. I already mentioned that variant theories like Okasha's and Schramm's require only *belief* in the relevant background counterfactuals, not *knowledge*. And other counterfactual theories, like Godfrey-Smith's, require only the *truth* of the relevant counterfactuals. I highlight these points to stress that, by considering the symmetry challenge as aimed at Jackson's original counterfactual approach, we have been considering it in its strongest form. Other counterfactual theories should do *at least* as well at answering the challenge.

Whatever form the counterfactual theory takes, there is no epistemological symmetry between us and the grue theorists with respect to these counterfactuals. If all parties understand counterfactuals as we do, then there are good reasons for preferring (1) and (2) over (III) and (IV). These reasons are general and theoretical and explanatory, but they aren't question-begging. Of course, this assumes that the challenge is posed using *our* understanding of counterfactuals. There is another way of pressing the counterfactual symmetry challenge. This more radical approach has recently been pursued by Christopher Dorst in critical discussions of the theories of both Alfred Schramm and Wolfgang Freitag.<sup>16</sup> Here I'll be discussing the general merits of the challenge, not its justice as an objection to any *particular* counterfactual theory.

#### 4 The Similarity Version

Consider how we semantically evaluate counterfactuals like (1), (2), (III), and (IV). Obviously, we used and asserted and evaluated counterfactuals long

- 
- 14 In addition to Moreland's (1976) discussion, my response here also dovetails with Godfrey-Smith's (2011) response to related objections, which he attributes to Laura Schroeter and Ira Schall. Though in some ways his discussion makes the challenge he is addressing sound more like the alternative similarity version discussed below.
- 15 Jackson later altered his account in several ways—see Jackson and Pargetter (1980).
- 16 See Dorst (2016, 2018), Schramm (2014), and Freitag (2015, 2016). Schramm's approach is similar to that of Jackson (1975) in many ways, but different in others. But despite how Dorst interprets him, Freitag doesn't appeal to knowledge of or belief in either (1) or (2)—see Freitag (2019) for further clarifications about this.

before anyone came up with an explicit semantic theory for counterfactuals. Still, a semantics is useful for codifying the truth conditions our practices assign to counterfactuals. The usual counterfactual semantics derives from Stalnaker and Lewis and uses a similarity metric over the space of possible worlds.<sup>17</sup> Here's a simplified version of this kind of semantics:

A counterfactual  $\Box \phi \rightarrow \psi$  is true at a world  $w$  just in case in all of the most similar worlds where  $\phi$  is true,  $\psi$  is also true.<sup>18</sup>

What exactly *similarity* comes to here has been much discussed.<sup>19</sup> There is broad agreement over cases, but the precise analysis is tricky. Sometimes “similarity” is claimed as subjective—including by Goodman himself.<sup>20</sup> Yet if subjectivity about similarity is combined with a similarity-semantics for counterfactuals, and then fed into our scientific and inductive practices as the counterfactual strategy requires, absurdities result.<sup>21</sup> This will be illustrated below.

Let's first assume that the relevant notion of similarity, though context-sensitive, is not completely subjective. Given what *we* mean by “most similar” in this semantic clause, the only way for (III) and (IV) to be true while (1) and (2) are false is for the world to be wildly different in the manner discussed in the previous section. Yet there is another option. Grue theorists could appeal to radically different “similarity” judgments, and then use those judgments in their counterfactual semantics without otherwise disagreeing with us about the world.

This involves saying that a world in which an observed green emerald *e*—this very one—is unobserved and blue, is *more similar to our world* than is a world where *e* is unobserved but green. This is bizarre to us, given what we mean by “similar.” Perhaps there are possible worlds where *e*, this very thing, is blue and not green. But given everything we know and believe about physics, chemistry, optics, and more, such a world must be very dissimilar to

17 See Stalnaker (1968) and Lewis (1973). A relatively popular minority alternative analyzes counterfactuals using strict conditionals ( $\Box \phi \rightarrow \psi$ ) and context-sensitivity.

18 Stalnaker and Lewis make different framework assumptions that lead to differences in their respective counterfactual logics. These differences won't be of concern here. Lewis's approach doesn't require a sphere of similarity containing *only* antecedent worlds, but I have simplified.

19 See Fine (1975), Lewis (1979), and Bennett (2003) for important contributions.

20 Not by proponents of the similarity semantics for counterfactuals—see Lewis (1983b). For Goodman's treatment, see his (1972, ch. IX).

21 Something like this bullet is bitten by Ullian (1961).

our world. The imagined grue theorist denies this. They agree with us about which worlds are possible. And they also agree with us about the facts in this world, but they disagree about how similar certain worlds are to this world.

Something like this reply might be implicit in some of Goodman's later discussions of gruesome matters. More recently it has been explicitly pursued by Dorst in reply to recent counterfactual approaches:

We are thus examining the same world in both cases, so only one of the two counterfactuals can possibly be true. [...] But which one is true? That will evidently depend on the similarity metric we impose on the space of possible worlds. On our traditional understanding of 'similarity,' the closest (most similar) world where the emeralds in our evidence class were not observed before 2020 will be one in which they are green and not grue. Surely, however, a 'grue'-speaker would have exactly the opposite conception of 'similarity.' After all, he thinks grue things all "look alike," so it is only natural that his conception of similarity would reflect that. [...] So if we appeal to counterfactuals to justify the 'green' induction over the 'grue' induction, the 'grue'- speaker will have a precisely symmetric justification open to him. (Dorst 2016, 153)

22

This understanding of the counterfactual symmetry challenge differs from the epistemological understanding discussed above. In some ways it is a more radical and troubling challenge.

There is no accounting for taste, and maybe there is no accounting for weird similarity judgments either. Yet meaning is determined by use. It's plausible that anyone who clearheadedly used the term "similar" so differently would no longer mean what we mean by the term. If they then used their alternative notion of similarity in giving a counterfactual semantics, this difference in meaning will also infect terms like "would" and "counterfactual." But the real issue is not about semantic theory. The real issue is *use*—the use that the formal semantic theory was meant to codify. Drastic changes in use lead to changes in meaning. If these grue theorists use counterfactuals in a way that aligns with their "similarity" judgments and not ours, then *they no longer mean what we mean by counterfactuals*.

---

22 A similar passage also occurs in Dorst (2018, 181).

If this “change of meaning” charge is true, it provides a response to the similarity version of the counterfactual symmetry challenge. The response is that, in adopting this version of the challenge, grue theorists have changed their language significantly. They have changed it so much, in fact, that *they no longer disagree with us*. Our dispute has devolved into a merely verbal dispute, with no direct disagreement.

In order to see this, it is important to understand that the kind of linguistic change involved here is *not* the simple change of moving to a language in which there are primitives for “grue” and “bleen” but all else remains the same. In that type of language, “similar” still has the same meaning it has in our language. So those grue speakers will agree with us about (1), (2), (III), and (IV) or rather, about their *translations* into the grue language. That change did not amount to a difference in worldview, only a difference in language. This shows that the radical counterfactual similarity charge is not backed up by the possibility of grue/bleen languages of the kind discussed by Goodman.<sup>23</sup> Instead a much more radical linguistic change is required, one that systematically alters the truth conditions of counterfactuals.

Some may quibble. Has meaning *really* been changed, they will ask. Anyone who thinks meaning is closely tied to use will say *yes*. And since almost everyone thinks that meaning is closely tied to use, almost everyone will say *yes*. Even Quine, the arch-critic of analyticity, argued that drastic meaning changes undermine simple homophonic translations (Quine 1970).<sup>24</sup> So I don’t think my claims about meaning change beg any significant questions about analyticity or the like.

We could argue for meaning change here *theoretically*, by appealing to widely accepted theoretical principles of interpretation or translation—charity, humanity, rationality, and so on.<sup>25</sup> But the central point is probably best illustrated more simply, by reflecting on simple applications of our actual practices of translation and interpretation. Imagine that you encounter someone who clear-headedly makes “similarity” judgments that align with those of our imagined grue theorist. Even after all of the facts are in, they continue to disagree with you. They say that grue things are “more similar” to each other than green things, even more similar with respect to “color,” and that grue things, but not green things, “look alike.” After you convinced yourself that these divergences are not caused by some perceptual deficiency or a mistake

23 Dorst (2016, 2018) sometimes seems to deny this. See also Schwartz (2005).

24 See Warren (2018) for an updated version of the argument.

25 For such principles, see Grandy (1973), Lewis (1974), Hirsch (2011), and Warren (2016).

about the factual situation, you would conclude that your interlocutor simply spoke a different language than you did. They simply do not mean what you mean by “similar” or the like.

This meaning change diagnosis is the best and most appealing way to understand the apparent disagreement here. To some extent though, it can be left to one side. The crucial point concerns the differences in practical language use. Even those who think there is a difference of opinion, not a difference of meaning, must admit that the gruesome practices differ wildly from our own. Let me provide a concrete illustration of this by considering what happens *after* time  $t$ , where  $t$  is the time to which the definition of “grue” is indexed. Let’s update Goodman’s original definition with a predicate for “observed before 2020”:

$$\text{Grue}_{2020}(\alpha) \leftrightarrow (\text{Green}(\alpha) \wedge \text{Observed}_{2020}(\alpha)) \vee (\text{Blue}(\alpha) \wedge \neg \text{Observed}_{2020}(\alpha))$$

Something is grue<sub>2020</sub> just in case it is either green and observed before 2020 or blue and not observed before 2020. Since it is now past 2020, we can observe previously unobserved emeralds without them being observed<sub>2020</sub>. What happens when we do is instructive. On January 1st, 2020, the grue defender is committed to the following for previously observed emerald,  $e$ :

(III\*)  $\neg \text{Observed}_{2020}(e) \square \rightarrow \neg \text{Green}(e)$

(IV\*)  $\neg \text{Observed}_{2020}(e) \square \rightarrow \text{Grue}_{2020}(e)$

Now let us observe a previously unobserved emerald,  $m$ .

1.  $\neg \text{Observed}_{2020}(m) \wedge \text{Emerald}(m)$  (assumption)
2.  $\forall x(\text{Emerald}(x) \rightarrow \text{Grue}_{2020}(x))$  (inductive projection made by the grue defenders, backed up by (III\*) and (IV\*))
3.  $\text{Grue}_{2020}(m)$  (1,2)
4.  $\text{Blue}(m)$  (1,3 and the definition of “Grue<sub>2020</sub>”)

---

1.  $\neg \text{Observed}_{2020}(m) \wedge \text{Emerald}(m)$  (assumption)

2.  $\forall x(\text{Emerald}(x) \rightarrow \text{Grue}_{2020}(x))$  (inductive projection made by the grue defe

3.  $\text{Grue}_{2020}(m)$  (1,2)

4.  $\text{Blue}(m)$  (1,3 and the definition of “Grue<sub>2020</sub>”)

---

- (1)  $\neg \text{Observed}_{2020}(m) \wedge \text{Emerald}(m)$
- (2)  $\forall x(\text{Emerald}(x) \rightarrow \text{Grue}_{2020}(x))$
- (3)  $\text{Grue}_{2020}(m)$
- (4)  $\text{Blue}(m)$

In other words, these grue theorists can prove to themselves that  $m$  is blue, and then they look to the world and see that it's green. Saying that  $m$  is green does not beg any questions here. This is because the similarity-based symmetry challenge differs from the epistemological challenge. The radical grue theory under consideration is supposed to agree with us about all of the physical facts, including facts about the color of emeralds. They were supposed to differ from us only over similarity claims.

Related arguments have been used elsewhere in the massive grue literature, for different purposes.<sup>26</sup> The purpose here is to illustrate that when counterfactuals connect to induction, as proponents of the counterfactual strategy believe that they do, our practice of evaluating counterfactuals is not isolated. It instead feeds into a cluster of related physical notions, including *nomological modality*, *laws*, *dispositions*, and *causes*.<sup>27</sup> So if you change what you count as “relevantly similar,” you change a great deal indeed.<sup>28</sup> Someone who changes what is meant by counterfactual terms would be ill-advised to fit their alternative “counterfactual” notions into the same conceptual space as our notions. Doing so leads to radically different ways of reasoning about and interacting with the same natural world.

Claiming that similarity itself is entirely subjective doesn't change this. If you say that, and then use similarity-relations to analyze the counterfactuals

- 
- 26 From the comprehensive annotated bibliography in Stalker (1994), I learned that something like this reasoning has been used by Bayesians like Cohen (1989) to assign the grue hypothesis a low prior probability. Although there Cohen seemingly used a definition of “grue” more in line with Barker and Achinstein's (1960)—see Jackson (1975) for criticism of this. Also, Cohen's reasoning requires an additional step about grue-like predicates that goes beyond anything in my argument so far. That step concerns future grue predicates. Here we don't need that, since if the grue defender doubles down, with  $\text{grue}_{2021}$ ,  $\text{grue}_{2022}$ , and so on, the same situation recurs. This is secured by the assumption that the radical grue defender agrees with us about all particular, physical facts.
- 27 See Putnam (1990) and Maudlin (2007) for related views. An aside: my view of alethic modality is (at least) tripartite. Logical and conceptual modality is a projection of our conventions, while physical modality is fully factual and objective. Finally, metaphysical modality is a mixed case—see Warren (2022).
- 28 Hesse (1969) made some related points. Like her, I don't think the strangeness of the grue theorist's conceptual scheme is itself a response to grue. Recall though that the dialectic here is that we are answering an objection to counterfactual approaches to grue.



which underwrite justified inductive inferences, you descend into a subjectivist nightmare.<sup>29</sup> Use any alternative counterfactual practice and you will likely find yourself with many false beliefs and many frustrated expectations. You could get lucky, but I wouldn't bet on it. Neither would you. Induction is not a pointless game we play for our amusement, it is instead a crucial part of how we reason about and master the physical world that surrounds and includes us.

So on neither reading does the counterfactual symmetry charge lead to genuine and troubling symmetry between our position and the grue theory. If the symmetry challenge is posed using *our* counterfactual notions, then we have non-question-begging *epistemic* reasons for favoring our counterfactuals over theirs. And if it is posed using some *alternative* notion of counterfactual similarity, then we have non-question-begging *practical* reasons for favoring our counterfactual practice over theirs. Either way, we have non-question-begging reasons for favoring our practices over the gruesome practices. Counterfactual approaches to grue might fail for other reasons, but the counterfactual symmetry charge doesn't stick. And despite its superficial appeal, the argumentative strategy it exemplifies is quite risky. I will close by explaining this.

## 5 Philosophical Fair Play

Twentieth-century philosophy was replete with overtly *semantic*, broadly *skeptical* challenges. These challenges attacked some of our most cherished doctrines using clever semantic tricks, principally clever redefinitions of crucial terms. The targets differed, as did the particular semantic tactics employed. Yet a general similarity between these challenges is easily recognized, provided it isn't overstated. Quine's translation argument, Putnam's model-theoretic argument, and Kripkenstein's skeptical paradox all fit into this model.<sup>30</sup> So too, does Goodman's grue puzzle.

Seen from this perspective, the overall dialectic surrounding the counterfactual symmetry challenge is quite familiar. A challenge has been posed by a semantic skeptic. One of our treasured assumptions is under threat. We rush in gallantly to offer a defense. Alas, the semantic skeptic uses a version of the original re-definition move yet again. This time on the very defense we have

<sup>29</sup> This is arguably the exact path that Goodman followed to reach the radical irrealist position of his (1978). Even Goodman's most committed followers, for instance, Scheffler (2001), were unable to follow this path all the way to the end.


<sup>30</sup> See Quine (1960), Putnam (1980), and Kripke (1982).

offered. The defense itself is seen by the skeptic as “just more theory” to be reinterpreted, just more grist for the skeptical mill.<sup>31</sup>

More often than not though, this move is not quite fair. When a constraint is used to screen off some skeptical reinterpretation, reinterpreting the *statement of the constraint* misses the mark. If we respond to Kripkenstein by claiming we mean addition and not quaddition by “plus” because we execute the addition algorithm in response to “plus” queries, talk of “quaddition algorithms” misses the point. The constraint concerns what we *do*, not what we *say about what we do*. Likewise with Quine’s challenge, and Putnam’s. Likewise too, with Goodman’s grue challenge.

Almost the same exact dialectic pops up again and again, all across the philosophical landscape, so the point is worth belaboring. Skeptical reinterpretation is *risky*. Great care must be taken whenever the move is attempted. In the present context, we have seen that blithe appeals to gruesome counterfactuals come with baggage. The counterfactual symmetry claim has hidden costs. Either a commitment to absurd factual claims or an unnoticed change of topic. In contexts like this, we must always take care to tease out all ramifications of the skeptic’s maneuvering. The semantic skeptic’s tricks are ever so easy to apply, but they can very quickly take us into uncharted waters, where monsters lurk. In these waters, merely ersatz symmetry is often mistaken for the real thing.\*

Jared Warren

 0000-0003-4028-7969

Stanford University

jaredwar@stanford.edu

## References

- BARKER, Stephen F. and ACHINSTEIN, Peter. 1960. “On the New Riddle of Induction.” *The Philosophical Review* 69(4): 511–522, doi:[10.2307/2183485](https://doi.org/10.2307/2183485).
- BENACERRAF, Paul and PUTNAM, Hilary, eds. 1964. *Philosophy of Mathematics: Selected Readings*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc. Second edition: Benacerraf and Putnam (1983).

<sup>31</sup> Compare Lewis’s (1984) discussion of the dialectic surrounding Putnam’s model-theoretic argument.

\* Thanks to Rosa Cao, Christopher Dorst, Eli Hirsch, Alfred Schramm, Douglas Stalker, and Daniel Waxman.

- , eds. 1983. *Philosophy of Mathematics: Selected Readings*. 2nd ed. Cambridge: Cambridge University Press. First edition: Benacerraf and Putnam (1964).
- BENNETT, Jonathan. 2003. *A Philosophical Guide to Conditionals*. Oxford: Oxford University Press, doi:10.1093/0199258872.001.0001.
- COHEN, Laurence Jonathan. 1989. *An Introduction to the Philosophy of Induction and Probability*. Oxford: Oxford University Press.
- DORST, Christopher. 2016. "Evidence, Significance, and Counterfactuals: Schramm on the New Riddle of Induction [on Schramm (2014)]." *Erkenntnis* 81(1): 143–154, doi:10.1007/s10670-015-9733-2.
- . 2018. "Bet Accepted. A Reply to Freitag (2015)." *The Philosophical Quarterly* 68(270): 175–183, doi:10.1093/pq/pqx016.
- FINE, Kit. 1975. "Critical Notice of Lewis (1973)." *Mind* 84(335): 451–458. Reprinted in Fine (2005, 357–365), doi:10.1093/mind/lxxxiv.1.451.
- . 2005. *Modality and Tense. Philosophical Papers*. Oxford: Oxford University Press, doi:10.1093/0199278709.001.0001.
- FREITAG, Wolfgang. 2015. "I Bet You'll Solve Goodman's Riddle." *The Philosophical Quarterly* 65(259): 254–267, doi:10.1093/pq/pqu093.
- . 2016. "The Disjunctive Riddle and the Grue-Paradox." *Dialectica* 70(2): 185–200, doi:10.1111/1746-8361.12136.
- . 2019. "Why Doxastic Dependence Defeats Grue: A Response to Dorst's Reply [to Dorst (2016)]." *The Philosophical Quarterly* 69(274): 156–165, doi:10.1093/pq/pqy029.
- GODFREY-SMITH, Peter. 2003. "Goodman's Problem and Scientific Methodology." *The Journal of Philosophy* 100(11): 573–588, doi:10.2307/3655745.
- . 2011. "Induction, Samples, and Kinds." in *Carving Nature at Its Joints. Natural Kinds in Metaphysics and Science*, edited by Joseph Keim CAMPBELL, Michael O'ROURKE, and Matthew H. SLATER, pp. 33–52. Topics in Contemporary Philosophy n. 7. Cambridge, Massachusetts: The MIT Press, doi:10.7551/mitpress/9780262015936.001.0001.
- GOODMAN, Nelson. 1946. "A Query of Confirmation." *The Journal of Philosophy* 43(14): 383–385. Reprinted in Goodman (1972, 363–366), doi:10.2307/2020332.
- . 1955. *Fact, Fiction and Forecast*. Cambridge, Massachusetts: Harvard University Press.
- . 1972. *Problems and Projects*. Indianapolis, Indiana: Bobbs-Merrill Company Inc.
- . 1978. *Ways of Worldmaking*. Indianapolis, Indiana: Hackett Publishing Co.
- GRANDY, Richard E. 1973. "Reference, Meaning, and Belief." *The Journal of Philosophy* 70(14): 439–452, doi:10.2307/2025108.
- HAHN, Lewis Edwin, ed. 1992. *The Philosophy of A.J. Ayer*. The Library of Living Philosophers n. 21. LaSalle, Illinois: Open Court Publishing Co.
- HARMAN, Gilbert H. 1965. "The Inference to the Best Explanation." *The Philosophical Review* 74(1): 88–95, doi:10.2307/2183532.

- HESSE, Mary B. 1969. "Ramifications of 'Grue' ." *The British Journal for the Philosophy of Science* 20(1): 13–25, doi:10.1093/bjps/20.1.13.
- HIRSCH, Eli. 2011. *Quantifier Variance and Realism: Essays in Metaontology*. Oxford: Oxford University Press.
- HUEMER, Michael. 2009. "Explanationist Aid for the Theory of Inductive Logic." *The British Journal for the Philosophy of Science* 60(2): 345–375, doi:10.1093/bjps/axp008.
- JACKSON, Frank. 1975. "Grue." *The Journal of Philosophy* 72(5): 113–131, doi:10.2307/2024749.
- JACKSON, Frank and PARGETTER, Robert. 1980. "Confirmation and the Nomological." *Canadian Journal of Philosophy* 10(3): 415–428, doi:10.1080/00455091.1980.10715734.
- KRIPKE, Saul A. 1982. *Wittgenstein on Rules and Private Language*. Cambridge, Massachusetts: Harvard University Press.
- LEWIS, David. 1973. *Counterfactuals*. Cambridge, Massachusetts: Harvard University Press. Cited after republication as Lewis (2001).
- . 1974. "Radical Interpretation." *Synthese* 27(3): 331–344. Reprinted, with a postscript (Lewis 1983c), in Lewis (1983a, 108–119), doi:10.1007/bf00484599.
- . 1979. "Counterfactual Dependence and Time's Arrow." *Noûs* 13(4): 455–476. Reprinted, with a postscript (Lewis 1986b), in Lewis (1986a, 32–51), doi:10.2307/2215339.
- . 1983a. *Philosophical Papers, Volume 1*. Oxford: Oxford University Press, doi:10.1093/0195032047.001.0001.
- . 1983b. "New Work for a Theory of Universals." *Australasian Journal of Philosophy* 61(4): 343–377. Reprinted in Lewis (1999, 8–55), doi:10.1080/00048408312341131.
- . 1983c. "Postscript to Lewis (1974)." in *Philosophical Papers, Volume 1*, pp. 119–121. Oxford: Oxford University Press, doi:10.1093/0195032047.001.0001.
- . 1984. "Putnam's Paradox." *Australasian Journal of Philosophy* 62(3): 221–236. Reprinted in Lewis (1999, 56–77), doi:10.1080/00048408412340013.
- . 1986a. *Philosophical Papers, Volume 2*. Oxford: Oxford University Press, doi:10.1093/0195036468.001.0001.
- . 1986b. "Postscript to Lewis (1979)." in *Philosophical Papers, Volume 2*, pp. 52–66. Oxford: Oxford University Press, doi:10.1093/0195036468.001.0001.
- . 1999. *Papers in Metaphysics and Epistemology*. Cambridge: Cambridge University Press, doi:10.1017/CBO9780511625343.
- . 2001. *Counterfactuals*. Oxford: Basil Blackwell Publishers. Republication of Lewis (1973).
- MAUDLIN, Tim. 2007. *The Metaphysics Within Physics*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199218219.001.0001.
- MORELAND, John. 1976. "On Projecting Grue." *Philosophy of Science* 43(3): 363–377, doi:10.1086/288693.

- OKASHA, Samir. 2007. "What Does Goodman's 'Grue' Problem Really Show? ." *Philosophical Papers* 36(3): 483–502, doi:[10.1080/05568640709485211](https://doi.org/10.1080/05568640709485211).
- PUTNAM, Hilary. 1980. "Models and Reality." *The Journal of Symbolic Logic* 45(3): 464–482. Reprinted in Benacerraf and Putnam (1964) and Putnam (1983, 1–25), doi:[10.2307/2273415](https://doi.org/10.2307/2273415).
- . 1983. *Realism and Reason. Philosophical Papers, Volume 3*. Cambridge: Cambridge University Press.
- . 1990. "Is It Necessary That Water Is H<sub>2</sub>O?" in *Realism with a Human Face*, pp. 54–79. Cambridge: Cambridge University Press. Reprinted in Hahn (1992, 429–454).
- QUINE, Willard van Orman. 1960. *Word and Object*. Cambridge, Massachusetts: The MIT Press. New edition: Quine (2013).
- . 1970. *Philosophy of Logic*. Cambridge: Cambridge University Press. Second edition: Quine (1986).
- . 1986. *Philosophy of Logic*. 2nd ed. Cambridge, Massachusetts: Harvard University Press. First edition: Quine (1970).
- . 2013. *Word and Object*. Cambridge, Massachusetts: The MIT Press. First edition: Quine (1960).
- ROSKIES, Adina L. 2008. "Robustness and the New Riddle Revived ." *Ratio* 21(2): 218–230, doi:[10.1111/j.1467-9329.2008.00396.x](https://doi.org/10.1111/j.1467-9329.2008.00396.x).
- SCHEFFLER, Israel. 2001. "My Quarrels with Nelson Goodman." *Philosophy and Phenomenological Research* 62(3): 665–677, doi:[10.1111/j.1933-1592.2001.tb00084.x](https://doi.org/10.1111/j.1933-1592.2001.tb00084.x).
- SCHRAMM, Alfred. 2014. "Evidence, Hypothesis, and Grue." *Erkenntnis* 79(3): 571–591, doi:[10.1007/s10670-013-9524-6](https://doi.org/10.1007/s10670-013-9524-6).
- SCHWARTZ, Robert. 2005. "A Note on Goodman's Problem." *The Journal of Philosophy* 102(7): 375–379, doi:[10.5840/jphil2005102715](https://doi.org/10.5840/jphil2005102715).
- STALKER, Douglas F., ed. 1994. *Grue!* LaSalle, Illinois: Open Court Publishing Co.
- STALNAKER, Robert C. 1968. "A Theory of Conditionals." in *Studies in Logical Theory*, edited by Nicholas RESCHER, pp. 98–112. American Philosophical Quarterly Monograph Series n. 2. Oxford: Basil Blackwell Publishers.
- ULLIAN, Joseph S. 1961. "More on 'Grue' and Grue." *The Philosophical Review* 70(3): 386–389, doi:[10.2307/2183381](https://doi.org/10.2307/2183381).
- VAN FRAASSEN, Bas C. 1989. *Laws and Symmetry*. Oxford: Oxford University Press, doi:[10.1093/0198248601.001.0001](https://doi.org/10.1093/0198248601.001.0001).
- WARREN, Jared. 2016. "Internal and External Questions Revisited." *The Journal of Philosophy* 113(4): 177–209, doi:[10.5840/jphil2016113411](https://doi.org/10.5840/jphil2016113411).
- . 2018. "Change of Logic, Change of Meaning." *Philosophy and Phenomenological Research* 96(2): 421–442, doi:[10.1111/phpr.12312](https://doi.org/10.1111/phpr.12312).
- . 2022. "Inferentialism, Conventionalism, and *A Posteriori* Necessity." *The Journal of Philosophy* 119(10): 517–541, doi:[10.5840/jphil20221191034](https://doi.org/10.5840/jphil20221191034).

WEISBERG, Jonathan. 2009. "Locating IBE in the Bayesian Framework ." *Synthese* 167(1): 125–143, doi:[10.1007/s11229-008-9305-y](https://doi.org/10.1007/s11229-008-9305-y).



# When Is Jealousy Appropriate?

ARINA PISMENNY

What makes romantic jealousy rational or fitting? Psychologists view jealousy's function as preserving a relationship against a "threat" from a "rival". I argue that its more specific aim is to preserve a certain privileged status of the lover in relation to the beloved. Jealousy is apt when the threat to that status is real, otherwise inapt. Aptness assessments of jealousy must determine what counts as a "threat" and as a "rival". They commonly take for granted monogamous norms. Hence, compared with jealousy in monogamous relationships, norms of polyamory set the thresholds for what counts both as a "threat" and as a "rival" much higher.

When is it appropriate to feel jealousy in a romantic relationship?<sup>1</sup> To answer this question, I explore the various rational norms of jealousy in the light of rational norms generally applicable to emotions. In particular, I analyze the relevance of moral, prudential, social, and aptness assessments to jealousy. I attempt to elucidate the formal object of jealousy—the *jealousy-worthy*—to show that it lacks the moral dimension required to justify the moral desert that the jealous person often takes themselves to have with respect to the beloved. I argue further that the aptness norms of romantic jealousy are significantly influenced by the specific romantic ideology that is taken for granted in the majority of romantic relationships. I show that monogamy provides conditions for numerous cases of apt and inapt jealousy, whereas polyamory significantly reduces the possibility of apt jealousy. That seems to mark a respect in which the latter type of relationship is not inferior and may even be thought superior from a moral point of view.

In section 1 I construct a psychological profile of jealousy, outlining its defining features. Section 2 presents various rational assessments applicable to jealousy. Section 3 analyzes the aptness conditions of jealousy, and presents

---

<sup>1</sup> In this paper, I concentrate on romantic jealousy as opposed to other kinds of jealousy such as sibling jealousy, workplace jealousy, friendship jealousy, etc. I use "jealousy" to refer to romantic jealousy unless otherwise specified.



arguments for thinking that jealousy is not an intrinsically moral emotion. This conclusion undercuts arguments that present a moral justification of jealousy as a strategy to protect what is rightfully one's own—namely the affections of one's beloved. Section 4 outlines the ways in which monogamous and polyamorous ideologies affect the aptness conditions of jealousy. I argue that in polyamory the conditions for apt jealousy are minimal compared to monogamy. I conclude that the questionable moral character of jealousy raises concerns about the moral status of monogamy, which is a great facilitator of jealousy.

## 1 Jealousy—A Psychological Profile

Without endorsing any particular theory of emotion, I take emotions to be intentional states that represent organism-environment relationships (Prinz 2004). Emotions are quick automatic responses that inform the organism of how it is faring in the world by making particular features of the situation salient to it (DeSousa 1987; Deonna and Teroni 2012). The *phenomenology* of an emotion makes a crucial contribution to the achievement of that task. Although emotional episodes can be unconscious, an occurrence of an emotion defines the domain of relevant features, informing other kinds of cognition in the subject (Damasio 1994; DeSousa 1987; Goldman 1986). Furthermore, emotions have characteristic *action tendencies*, preparing an organism to respond to a particular situation in a meaningful way (Frijda 1987; Scarantino 2017).

The intentionality of emotions is characterized by two kinds of objects. The emotion is directed at a *particular object* or *target*, and represents the target as having a particular evaluative property—the *formal object* of the emotion (Kenny 1963; DeSousa 1987). Emotions can misrepresent their targets when the target lacks the properties that ground the formal object of the emotion. The intentionality of emotions necessitates that they have correctness conditions—aptness. An emotion is apt when it correctly represents the target as having a particular evaluative property that supervenes on a set of natural properties of the target. Fear, for example, is apt when its formal object—the *fearsome*—supervenes on the properties (the menacing teeth, attacking posture) of its target (the dog). An emotion is inapt when the target that it represents as having a particular evaluative property lacks natural properties sufficient to ground the evaluative property. Thus, an instance of fear of a dog is inapt when the dog poses no danger. On the basis of this characterization,

emotions have two functions: (1) to inform the organism of how it is faring by correctly identifying evaluative properties of the target, and (2) to prepare an organism to respond to the situation by facilitating a response appropriate to those evaluative properties—in the case of fear, flight, or some other means of evading the danger.

Jealousy can be characterized along these and other parameters pertaining to emotions. Emotions are individuated by their formal objects (DeSousa 1987; Deonna and Teroni 2012; Tappolet 2016). In order to zoom in on the formal object of jealousy, the *jealousy-worthy*, it is important to identify the *eliciting conditions* of jealousy. Jealousy's defining eliciting condition involves a love triangle composed of the lover, the beloved, and a rival.<sup>2</sup> The negative hedonic character of jealousy indicates that the presence of a rival is a threat to the relationship or some aspect of the relationship between the lover and the beloved (Ben-Zeév 1990; Protasi 2017). Jealousy makes salient the features of the situation that constitute this threat. I will have a lot more to say about the formal object of jealousy in the upcoming sections. For now, we can say that the formal object of jealousy is a *threat-of-a-loss* posed by a rival to one's romantic relationship. If this is right, then jealousy is apt when the threat posed by a rival is real, and inapt when it is not.

The action tendency of jealousy offers further support for thinking that jealousy is a response to a threat because in romantic contexts the jealous engages in a variety of behaviors that appear to constitute *mate guarding*. These include interrupting the interaction between the beloved and the perceived rival, aggression against the beloved, or withdrawing (Chung and Harris 2018). Given the eliciting conditions of jealousy and its action tendency, the *function* of jealousy seems to be warding off rivals in order to protect one's relationship.

The diversity of mate guarding behaviors raises questions about the *target* of jealousy. Is it directed at the rival or the beloved? The grammatical structure of jealousy says that one is jealous *of* the rival. However, one is a rival only if one is receiving affection and attention from the jealous subject's beloved. Furthermore, it is the beloved whom the lover does not want to lose. Mingi Chung and Christine Harris report that the actions of mate guarding tend to be directed at the beloved more often than the rival. They hypothesize that this is because it is easier to secure the beloved's faithfulness than to discourage all others from attempting to lure the beloved away. Since it is the stability

---

<sup>2</sup> The three-party relationship is one feature that distinguishes jealousy from envy. For discussion see Farrell (1980), Ben-Zeév (1990), Kristjánsson (2002, 2018), and Protasi (2017).

of the beloved's affections that the lover is trying to secure, it makes sense that jealousy should be directed primarily at the beloved. At the same time, the rivalrous nature of jealousy indicates competitiveness of the lover for a privileged status with respect to the beloved (Farrell 1980). Therefore, the target of jealousy is both the beloved and the rival. The particular strategy employed in a given case may be indicative of the *focus* of the lover's jealousy.

If jealousy is about responding to threats from rivals, how should these threats be characterized? In the psychology literature on jealousy it has been defined in the following two ways. First, a threat may be constituted by an actual transgression of the beloved with a rival—e.g., a flirtation or an affair. In this case one experiences *reactive jealousy*—a jealous response to an actual infidelity.<sup>3</sup> Second, an aspect of the situation may be construed as a potential threat to one's relationship. In this case one experiences *suspicious jealousy*—a jealous reaction to a potential infidelity of the beloved with a rival (Rydell and Bringle 2007; Attridge 2013).<sup>4</sup>

Reactive and suspicious jealousy are typically distinguished by their antecedent conditions. Each kind is also associated with different qualities of the lover's personality. Reactive jealousy is associated with dependency and trust, secure and avoidant attachment styles, and extroversion. Suspicious jealousy is associated with insecurity and low self-esteem, and is correlated with anxious attachment style, and neuroticism (Marazziti et al. 2006; Chung and Harris 2018). Given these associations and antecedent conditions of each kind of jealousy, it may appear that reactive jealousy is always apt because it correctly identifies a threat, whereas suspicious jealousy may be more prone to error since it arises in cases where the threat is not obvious. Furthermore, the association with neuroticism and anxious attachment style suggests that suspicious jealousy, as an occurrent emotion, can sometimes be regarded as manifesting a character trait: a *jealous person* is one who often endures episodes of unfounded jealousy. Trait jealousy is associated with particular individual dispositions—anxiety, distrust, and suspicion—and is better described as a dimension of personality. In fact, Chung and Harris propose to delineate the distinction between suspicious and reactive jealousy not by construing them as two types of jealousy but rather as two aspects of the

3 Of course, infidelity may not constitute a threat to the relationship if the lovers have an arrangement about allowing extra-dyadic sex. Here I am assuming that it does for the sake of argument.

4 For other kinds of characterizations of jealousy see Pfeiffer and Wong (1989), and Buunk et al. (2020).

same emotion. They maintain that since the function of jealousy is to detect a threat, suspicious jealousy may be thought to be the initial stage of jealousy, when the jealous is gathering and examining the evidence for a potential transgression by the beloved. This way suspicious jealousy, no less than reactive jealousy, fulfills the function of protecting the relationship from threats. If suspicious jealousy arises in circumstances that do not ground it, then it simply fails to perform its presumed function. However, it need not fail to do so. Distinguishing between reactive and suspicious jealousy does not commit us to thinking of one as always apt and the other inapt.

Thus far we can say that romantic jealousy is an emotion that arises in response to a perceived threat posed by a rival with respect to one's beloved. Jealousy aims to correctly identify the threat, and to facilitate action designed to protect one's relationship from the rival.

## 2 Jealousy and Norms

If jealousy has the twofold function specified above, how effective is it in warding off rivals and sustaining a relationship? To answer this question, we need to examine the different ways—moral, prudential, and social—in which jealousy's effectiveness is assessed. That is the aim of this section.

I begin with social attitudes to jealousy, illustrating their variability across and within cultures. I then examine the prudential value of jealousy, in terms both of its social meaning and of its biological function. Lastly, I assess jealousy from a moral point of view: can it be said to be a moral emotion? As will become clear, these assessments call into question the value of jealousy and lay the groundwork for the critical evaluation of the formal object of jealousy in the next section.

One finds a variety of *social attitudes* to jealousy across and within cultures (Hupka and Ryan 1990; Buunk et al. 2020). For instance, in the so-called honor cultures—cultures in which reputation and status matter greatly—men are thought to be justified in violent outbursts triggered by jealousy (Cihangir 2013; Canto et al. 2017). In the matriarchal society of Mosuo in Southwest China jealousy is frowned upon (Cai 2001). In the United States attitudes towards jealousy are mixed (Puente and Cohen 2003; Vandello and Cohen 2008).<sup>5</sup> On the one hand, jealousy is praised as an expression of love, care, attachment, and vulnerability (Buss 2000). The jealous lover is clearly invested

---

<sup>5</sup> For the study of honor culture in the U.S. see Nisbett and Cohen (1996).

in the beloved and the status of their relationship: they are hurt by the potential loss of the relationship. They wish to keep it and protect it from intruders who might take their beloved away from them. On the other hand, jealousy is disapproved of as it signals insecurity, low self-esteem, possessiveness, lack of trust, and immaturity. It portrays the lover as suspicious, mistrustful, and controlling (Salovey 1991).

The diversity of opinions on jealousy in the United States is partly explained by the changing attitudes towards the conception of men's honor and women's purity (Stearns 2010). At the same time, it is clear from these meta-attitudes that jealousy has multiple complex, conflicting social meanings. One way to attempt to reconcile them is to appeal to the distinction between apt reactive jealousy and inapt suspicious jealousy. Furthermore, pathological or morbid jealousy is associated with violence and homicide, and may stem from both reactive and suspicious jealousy (Pfeiffer and Wong 1989; Mullen 1993; Wilson and Daly 1996).<sup>6</sup> That might also be motivating negative attitudes towards jealousy.

While it is unlikely that negative and positive attitudes towards jealousy neatly map onto these distinctions, the purported function of jealousy might justify some of the positive attitudes towards it. We must look at the different ways in which this function is to be understood in order to assess some of the justifications of jealousy it might provide. The usefulness of jealousy can be construed in terms of its supposed biological, social, and personal functions. These are not mutually exclusive but differentiating between them sheds light on the utility of jealousy. I begin with the biological function of jealousy.

An evolutionary psychologist, David Buss, and his colleagues have argued that romantic jealousy is an adaptation. It was selected to ensure pair-bonding and successful childrearing in human reproduction, by securing sexual exclusivity from women and emotional exclusivity from men (Buss and Schmitt 1993; Buss 2000, 2006). Buss argues that the perceived sex differences with respect to jealousy reflect different evolutionary challenges for the sexes: securing paternity makes men more jealous of women's sexual infidelity, and securing resources makes women more jealous of men's emotional infidelity. Buss's findings have been challenged in the light of wide cultural variations with respect to sex differences in jealousy. Notably, in more egalitarian soci-

---

<sup>6</sup> Peter Stearns (2010) outlines other social changes in the twentieth century that contributed to the negative shift in opinion regarding jealousy.

eties both men and women care more about emotional fidelity (DeSteno and Salovey 1996; Harris 2003; see also Hupka and Ryan 1990).

Putting the question of the best explanation for sex differences in jealousy aside, there is reason to think that jealousy may be adaptive because it is universal and traceable in infants as young as three months old. Sybil Hart and her colleagues found that infants react negatively to their mother talking sweetly to a lifelike doll but not to a book, suggesting that the mechanism for jealousy is hardwired to enable infants to secure vital resources from their caregivers by taking their attention away from real rivals (Hart 2010). One can speculate whether the jealousy response in infants is co-opted in romantic jealousy or whether an innate disposition to sexual jealousy is already expressed in infant jealousy. Regardless, the universality of a trait and its presence in infants are insufficient to establish that it is adaptive. As illustrated by the presence of the vermiform appendix, and by our preference for fatty and sugary foods, some features of an organism, while they might have been adapted in our evolutionary past, are no longer adaptive and may even be deleterious in our current environments (DeSousa 2017). Lastly, the presumed adaptive function does little to justify a normative assessment of jealousy since the adaptiveness of a trait does not imply that it is socially or personally beneficial.

Another way to approach the functionality of jealousy is to think of the role it plays in society. Given the general twofold function outlined above, it may be that jealousy contributes to maintaining social structures such as families by sustaining pair bonds. However, we must ask to what extent jealousy is a successful strategy in preserving these institutions. Furthermore, we must weigh the costs placed on the members of these institutions to assess whether jealousy is a justifiable means to achieve those aims.<sup>7</sup>

Assessing the success of jealousy is difficult since numerous factors contribute to sustaining a relationship. One possible measure of jealousy's contribution is its correlation with relationship satisfaction—an individual's assessment of the quality of their relationship. Relationship satisfaction can serve as a predictor of the relationship's endurance (Hendrick 1988). In some studies of jealousy, after testing participants' jealousy responses to vignettes and asking them to assess the jealousy reactions of their partners to their potential infidelities, the participants were asked to answer questions rating their relationship satisfaction. Different studies found different correlations

---

<sup>7</sup> One could also question the value of the institutions jealousy is said to protect. See, for instance, Brake (2012; Brake 2016). But that project is well beyond the scope of this paper.

between jealousy and relationship satisfaction. For example, a study by Laura Guerrero and Sylvie Eloy found a negative correlation between all types of jealousy and relationship satisfaction (1992; see also Andersen et al. 1995). Others found that relationship satisfaction positively correlates with reactive jealousy but negatively correlates with suspicious jealousy (Barelds and Barelds-Dijkstra 2007; Dandurand and Lafontaine 2014).<sup>8</sup>

Furthermore, studies of jealousy expression and communication found that aggressive expression or manipulative behavior designed to control or hurt one's partner is negatively correlated with relationship satisfaction. The same was true for aggression against the rival. On the other hand, constructive communication that focused on discussing relationship issues and aimed at restoring the relationship was positively correlated with relationship satisfaction (Sheets, Fredendall and Claypool 1997; Guerrero, Hannawa and Babin 2011). This suggests that when thinking about the correlation between jealousy and relationship satisfaction, people report how they perceive different types of jealousy as well as how they react to communications of jealousy.

These reports shed light on people's attitudes to jealousy and its expression. The correlations tracked in these studies are inconclusive, however, because of the mixed results and also because correlation does not establish causation. Even if we assume that there is a positive correlation between reactive jealousy and relationship satisfaction and a negative correlation between suspicious jealousy and relationship satisfaction, it does not mean that reactive jealousy in fact improves the relationship. Yet these correlations are telling, because they demonstrate that for many people jealousy is an important part of the romantic love narrative. No doubt, for some jealousy is a sign of love and commitment. But violence perpetuated and justified by jealousy imposes a disproportionate cost on women in romantic relationships (Mullen and Maack 1985; Daly and Wilson 1988; White and Mullen 1989; Mathes and Verstraete 1993; Puente and Cohen 2003; Vandello and Cohen 2008). Thus, in light of current research, one is left doubting the social usefulness of jealousy.

While biological and social justifications of jealousy do not appear promising, one might assess the prudential value of jealousy on an individual level. Jealousy might improve a relationship by correctly identifying threats and employing successful strategies for securing it. It would then be contributing to relationship satisfaction. Many other things would have to be true for this

---

8 Dandurand and Lafontaine (2014) have found that people react more positively to their own jealousy that they direct at their beloveds, and more negatively when the jealousy of their beloveds is directed at them.

picture to be correct. Personality, character, attachment styles of the individuals involved, their beliefs about romantic love, a particular type of jealousy and jealousy expression, together with other factors will determine the prudential value of jealousy for those individuals. Therefore, while jealousy may have prudential value in particular cases, that value depends on numerous factors that are difficult to generalize.

The analysis of social attitudes towards jealousy and of jealousy's role on the biological, social, and individual levels puts pressure on the significance of jealousy and casts doubt on its functionality. However, despite its questionable utility, jealousy might turn out to have a positive *moral* value. The moral value of jealousy can be cashed out in two ways: (1) if jealousy's formal object is a moral property, and (2) if it turns out to be morally praiseworthy. These two ways in which jealousy might relate to morality are independent of one another but can overlap.

If jealousy is a moral emotion, its formal object—the *jealousy-worthy*—is a moral property. Its aptness conditions would be defined by considerations of whether the target of jealousy instantiates the moral property of the *jealousy-worthy*. Since jealousy aims to identify threats to one's relationship posed by a rival-beloved interaction, the *jealousy-worthy* could be a moral property if it designates an injustice constituted by the rival-beloved relationship. It is important to note that if jealousy turns out not to be a moral emotion, rejecting its aptness on moral grounds would amount to committing a moralistic fallacy (D'Arms and Jacobson 2000). That is, if jealousy is deemed irrational on the grounds that it is immoral to feel, there would be a conflation of moral assessment and the aptness norms of jealousy. This is the case regardless of whether the *jealousy-worthy* is a moral property. I explore these questions in the next section.

### 3 The Formal Object of Jealousy: Moral or Non-Moral?

Does the formal object of jealousy consist of a moral property? To answer this question I examine moral and nonmoral accounts of the formal object of jealousy. I consider the implications of characterizing jealousy as a moral and a nonmoral emotion. I argue that if jealousy is a moral emotion then it is always inapt. If jealousy is a nonmoral emotion, it can be apt but it is morally problematic.

An account of the formal object of jealousy as a *moral property* is defended by Kristján Kristjánsson. He construes jealousy as an Aristotelian virtue of



self-respect (2002, 2018). For Kristjánsson jealousy is a mean between two extremes: too much sensitivity to perceived disrespect, and too little sensitivity to the disrespect manifested by the beloved who responds favorably to a rival. According to this view, jealousy is an emotion that protects one's self-respect as a response to disrespect from others. It defends that which is due to and deserved by the lover. It is an emotion that responds to an injustice akin to anger and indignation, as opposed to fear, which responds to a danger. Kristjánsson construes jealousy as a moral emotion, the formal object of which is the violation of *moral deserts* (2002, 153). He argues that jealousy is necessary for a good life because it serves the function of preserving self-respect and respect from one's beloved. Therefore, jealousy is a moral emotion in virtue both of (1) the moral nature of its formal object, and of (2) its praiseworthy character.

In contrast to Kristjánsson, several accounts construe the *jealousy-worthy* as a nonmoral property. Daniel Farrell says, “[T]o be jealous is to be bothered by the very fact that one is not favored in some way in which one wants to be favored” (Farrell 1980, 543; see also Ben-Zeév 1990, 2010). More specifically, the jealous person perceives the beloved-rival interaction as a threat to their *privileged status* with respect to the beloved. Farrell’s view brings out the rivalrous nature of jealousy—the jealous person wants to be favored *more* than anyone else by the beloved in the ways that a romantic lover is favored. Farrell denies that jealousy is a response to a threat of a loss of a relationship since a person might still be jealous, even if they could be assured that they would not lose it. The formal object of jealousy in his view is a *threat-to-one’s-privileged-status*. It is not a moral property since it is not grounded in desert. Instead, it is simply a fact about human psychology.

Similarly, Sara Protasi describes jealousy as *threat-of-a-loss-of-comparative-advantage* to a rival. She says, “[T]he jealous is motivated to protect her comparative advantage, possibly by fending attacks from the rival and/or locking away the good” (2017, 323).

Both Farrell and Protasi point out that the formal object of jealousy reflects the *exclusivity* criterion associated with monogamy.<sup>9</sup> The monogamous framework requires that only one partner be the recipient of sexual and emotional favors from the beloved. The presence of a more favored rival threatens the privileged status of the lover. It devalues the goods of love and sex by undermining exclusivity. I will have more to say about these features of monogamy

---

9 By “monogamy” I mean a romantic relationship governed by the norms of sexual and emotional exclusivity.

below. For now, it is important to emphasize that in a romantic context, sexual and emotional exclusivity determine the status of being favored.

The three accounts just cited—from Kristjánsson, Farrell, and Protasi—illustrate ways in which the formal object of jealousy can be construed. I first turn to the moral property accounts.

Kristjánsson argues that jealousy is a moral emotion, whose formal object is a *threat-to-moral-desert* that supervenes on the beloved-rival interaction, and on the relationship between the lover and the beloved. Jealousy upholds one's self-respect when one is mistreated by the beloved. According to him, "jealousy can properly be felt by *A*, other things being equal, when *B* receives from *C* a favor that *A* deserves more than, or at least as much as, *B*" (Kristjánsson 2002, 163). But what determines whether *A* deserves favors from *C* more than does *B*? Kristjánsson says it is the expectations of fairness provided by rules of commitment and faithfulness in the romantic love institutions: "[E]xclusive affiliation is typically valued from the very start of a loving relationship, and indications of complete indifference in this matter are likely to be considered morally defective" (2002, 158–159). Hence, one deserves favors from one's beloved more than a stranger or friend does because one is in a romantic relationship with them. The desert is cashed out in terms of sexual and emotional exclusivity. That, we are to understand, is dictated by monogamy, the default mode of romantic relationships. The jealous person deserves not to be made jealous since if they are experiencing apt jealousy, they have been disrespected.

Kristjánsson thinks that following these rules of romantic relationships amounts to respecting one's romantic partner, while not reacting with jealousy towards the beloved's transgressions indicates a lack of self-respect. Kristjánsson recognizes that social rules dictate how self-worth should be understood, what boosts it and undermines it (2002, 161). Jealousy for him, therefore, as a protection of self-worth, is connected with one's reputation and status. For example, since cuckoldry is shameful, especially in certain cultures, jealousy is justified as a means to guard against it.

Kristjánsson's argument for the morality of jealousy goes as follows: there are social rules that govern relationship structures. These rules create expectations for the members of society. One such rule is about sexual and emotional exclusivity between romantic partners. When people enter romantic relationships, they take these rules for granted. Following these rules fulfills the expectations of the romantic partners. Violating these rules amounts to disrespecting one's partner because such violations undermine their expectations.

Therefore, one ought to follow the rules in place in order to treat one's partner well.

The argument assumes a moral obligation to uphold and follow social rules. This assumption is clearly indefensible: the moral status of such rules can always be questioned.<sup>10</sup> Hence it remains to be demonstrated that monogamous norms are morally defensible.<sup>11</sup>

Kristjánsson makes a leap from socially defined expectations to moral desert. In fact, his account seems, paradoxically, to imply that jealousy can never be apt. To see this, consider that the formal object of jealousy in his view, *threat-to-moral-desert*, is grounded in one's expectations, which are in turn grounded in social conventions (for Kristjánsson recognizes that they take different forms in different times and places).<sup>12</sup> But moral desert cannot be grounded in social norms. It follows that on Kristjánsson's account jealousy can never be apt since the value property it represents is not grounded in the features of the world he has in mind.

If we cannot ground moral desert in social norms, then we might characterize jealousy as representing not moral desert but a certain form of socially sanctioned *entitlement*. The formal object of jealousy would then be *threat-to-entitlement*. Entitlement arises from participating in social or legal institutions that specify how one ought to be treated (Feldman and Skow 2020). For example, a customer is entitled to a refund from a store when they are not satisfied with their purchase if the store's policy specifies that such refunds will be provided on this basis. An athlete is entitled to a gold medal if they have won the competition, and a gold medal is the way in which the winner is rewarded.

If jealousy is about entitlement and entitlement is not a moral property, then jealousy is not a moral emotion, and cannot be a virtue. But in any case, how strong is the case for the claim that the institution of monogamy entitles one to sexual and emotional exclusivity? Is it the kind of institution that can provide conditions for entitlement? The institutions presented in the examples above are formal institutions with explicit rules that can be enforced. Monogamy (in the restricted sense in which I have used the word) is an institution in a different sense—it is an informal institution, a widely

---

10 Slavery, segregation, and inequitable gender norms demonstrate this point.

11 For extensive criticism of monogamy see Brake (2017); Brunning (2016, 2020); DeSousa (2017, 2018); Jenkins (2017).

12 E.g., "In Mediterranean societies, for instance, people have tended to be extremely sensitive to pride and shame in matters concerning sexual fidelity [...] whereas transgressions of that kind may have been viewed more lightly in liberal France" (Kristjánsson 2002, 161).

accepted social practice. The rules are not explicit, and there is no formal way for them to be enforced except for the court of public opinion. In that sense, monogamy can be viewed as a social convention.

The practice of monogamy can be formalized through the formal institution of marriage. In marriage, the rules of monogamy are explicit and have been enforceable until the introduction of no-fault divorce. Is one entitled to exclusive affection and sexual attraction from one's spouse? Indeed, the marriage contract seems to entitle one to such exclusivity. However, it should now be clear that formal and informal social institutions on their own cannot *morally* justify a social practice. Simply accepting them without further argument ignores their variability across time and cultures, and commits one to embracing an objectionable social conservatism. If the aptness conditions of jealousy are simply defined by social norms, they tell us nothing about its moral value.

We could try to show that jealousy is a moral emotion by grounding *threat-to-moral-desert* in some other way. One possibility is to adopt a contractarian framework and cash out moral desert in terms of an implicit agreement to "terms and conditions" of a monogamous romantic relationship. The contractarian framework establishes rights and obligations for all parties involved. On this view, one's romantic partner has a moral claim to one's sexual and emotional favors that outweighs any such demands from third parties, by virtue solely of the romantic relationship's existence. The relationship entails rights, and jealousy is an emotion that guards those rights.

But can one really ever assert a right to be loved? That is surely questionable because love is neither a matter of desert, nor of the will (Neu 1980, ch. 3).<sup>13</sup> Construed in this way, jealousy is then always inapt since the *threat-to-moral-desert* is really a *threat-to-one's-rights*, and there are no such rights.<sup>14</sup>

It could be insisted that while one may not have a right to be loved, according to the romantic contract, one has a right to sexual and emotional exclusivity for as long as the partner can provide them. That is, if the beloved falls in love with someone else, the romantic contract is terminated since the conditions of the original agreement are no longer satisfied. The contract only lasts as long as its conditions endure.

13 This seems true even in the case of child-parent love. For discussion, see Liao (2015) and Protasi (2019).

14 Although jealous people might often feel that they do have such rights. For more details see Neu (1980, ch. 3), and Wren (1989).

Another possibility is to acknowledge that sexual and emotional attraction are not in fact exclusive. The monogamous contract prohibits *acting* upon attractions towards others. Pursuing them would violate the obligation of exclusivity. In this case, jealousy is the insistence that one honor the contract of exclusivity despite other attractions. Yet, jealousy is clearly not just about prohibiting the beloved to act upon their attractions. It is about being *preferred* to all others by the beloved. Can it be shown that one has a moral obligation to prefer one's partner to all others sexually and emotionally? It seems not, for, as we have said, there are no moral obligations to love, or to be exclusively sexually attracted to someone. Given these considerations, the contractarian framework cannot sustain jealousy's claim to be a moral emotion.

Another attempt might be made to show that jealousy is a matter of moral desert. Consider the concept of cheating. Cheating constitutes not only a transgression of the rules of a romantic relationship but a betrayal of the partner's trust. Why? Because the expectation of exclusivity was violated. How does one acquire such an expectation and why does one trust that it will be fulfilled? The expectation is a default assumption in a romantic relationship since monogamy is the default kind of romantic relationship. Through their actions and words, the partners lead one another to believe that both will be sexually and emotionally exclusive. As the relationship develops, the partners can explicitly state or otherwise indicate that they are "not seeing anyone else", thereby tacitly or explicitly endorsing monogamy. One reason why cheating is wrong is not because one's expectations are violated but rather because one's trust is.<sup>15</sup> Can it be said that a threat posed by a rival-beloved interaction is the kind of threat that endangers the trust between the lover and the beloved such that jealousy is an apt response to the situation? While it is clear that the threat to trust is real and that the beloved has a moral duty not to deceive the lover, the threat to trust does not make jealousy apt because jealousy is about deserving to be valued more than the rival. It is about having a greater claim to the affections of the beloved than the rival. A violation of trust constitutes a condition for apt anger and apt sadness but not apt jealousy.

In sum, jealousy construed as tracking injustice fails to be apt. To be sure, this reason is insufficient to rule out the possibility that the formal object of jealousy is *threat-to-moral-desert* or *threat-to-one's-rights*. It could well be that the formal object of jealousy is one of these moral properties. But if so, then jealousy is always unfitting because the properties that are supposed to

---

<sup>15</sup> For a discussion of how duties of trust arise in intimate relationships, see Wallace (2012).

ground the formal object thus specified fail to do so.<sup>16</sup> The same can be said about a *threat-to-entitlement*. I will not attempt to settle the matter of whether the formal object of jealousy is a moral property here.

Let us now move on to the critical analysis of the proposed nonmoral formal object of jealousy discussed by Farrell and Protasi. Recall that Farrell and Protasi construe the formal object of jealousy as a *threat-to-one's-privileged-status* and *threat-of-a-loss-of-comparative-advantage* respectively. According to them, jealousy is an emotion that aims to protect one's priority standing with respect to the beloved. It is a response to a threat to one's status by a rival. In their views, jealousy is apt when one's privileged status is actually threatened by a rival, and inapt when it is not. This seems like a very plausible account of jealousy because it captures the rivalrous nature of jealousy. It also does not attempt to justify it from a moral standpoint as it does not insist that the jealous person deserves to be valued this way.

Farrell raises the question of the intelligibility of jealousy. He points out that there is something strange about a mature adult having this emotion (Farrell 1980, 546). Indeed, this characterization makes the jealous person look selfish, self-absorbed, and insecure. Farrell suggests that being favored more than anyone else could be intrinsically pleasurable for some people just as it seems to be for children and nonhuman animals (1980, 553). While this may be so, it is still puzzling since children are discouraged from being jealous. Why should jealousy be an appropriate emotion in a romantic context?

Farrell's and Protasi's accounts present a plausible picture of the formal object of jealousy and its aptness conditions. There remains the question of whether jealousy is morally justifiable or praiseworthy. It would seem that the jealous person confuses being valued as special and being the only one valued. In addition, they want to be in a superior position to everyone else.

It might seem that jealousy is justified by a monogamous ideology because it is based on an underlying assumption that "true love" can only be for one person at a time. Such an assumption implies that if love is not exclusive then it is not really love, or a love that is worthwhile, as it is not true love. Whether one can experience romantic love for more than one person at a time is an empirical question. Given numerous polyamorous accounts, it seems

---

<sup>16</sup> Perhaps regret, contempt, grief, and hatred (if the latter is an emotion) are also examples of inherently inapt emotions. For discussion see Landman (1993), Bell (2013), DeSousa (2019), Price (2020), Brogaard (2020), and Aumer and Erickson (2022). Caroline Price makes a case for the rationality of grief (2018). But she reduces aptness to prudence.

that it is indeed possible (Brake 2017; Jenkins 2017).<sup>17</sup> Defining “true love” as necessarily exclusive, therefore, begs the question.

It might also seem that what the jealous person wants is to be valued as unique and special. It might seem that being the only one valued satisfies this desire since if one is the only one loved in this way, one appears to be preferred to everyone else. However, it is a mistake to equate exclusivity with being valued as unique because exclusivity does not by itself take care of the Problem of Trading Up—the idea that if someone better comes along, the lover will prefer them to their current beloved (Nozick 1989). To address the problem, we must move away from equating being valued as unique with being valued exclusively. Neither entails the other. Exclusivity by itself does not preclude one from regarding one’s beloved as fungible. Instead, valuing the beloved as unique is best captured by valuing them as irreplaceable where the lover simply refuses to compare the beloved to others (Grau 2004).<sup>18</sup> Valuing the beloved as unique is a normative attitude grounded in the love-attitude of the lover, and not in some set of features of the beloved. If uniqueness is characterized empirically, it is contingent.

It is also a mistake to think that one cannot be valued as unique if one’s partner has other lovers. Each one can be valued in this way by virtue of being loved. Therefore, exclusivity by itself does not provide conditions for being valued as unique or irreplaceable. Rather, it is the normative attitude of the lover that perceives their beloved as irreplaceable, i.e., not fitting for comparison or ranking.

One further defense of the claim that uniqueness stems from sexual and emotional exclusivity might appeal to the “relationship first” view elaborated by Niko Kolodny (2003). In Kolodny’s view, a relationship might be defined by a requirement of exclusivity; in such a case the uniqueness of the relationship might be due to precisely that defining commitment.<sup>19</sup> But that is true of any commitment mutually undertaken—never to use tobacco, or never to see an Orson Welles movie without the other. Such commitments might create “reasons of love” for that kind of exclusivity, but reasons of love may not be moral reasons. They give rise to disappointment and hurt, but that is

- 
- 17 Polyamory is a form of ethical nonmonogamy, in which individuals have (or are open to having) multiple romantic partners with voluntary informed consent of everyone involved.
- 18 J. David Velleman (1999) and Troy Jollimore (2011) make a similar point about what it means to being valued as unique or special.
- 19 I am grateful to an anonymous reviewer for noting the possibility that lovers might decide to make a relationship “intentionally exclusive, for whatever reason”.

very different from the moral indignation that is warranted in response to a moral transgression (see [Albrecht 2017](#); [Pismenny 2021](#)). Such intentional commitments, then, cannot amount to a *moral* entitlement for sexual or emotional (as opposed to any other kind of) exclusivity.<sup>20</sup>

The desire to be valued as unique or special, according to Farrell's and Protasi's accounts, is not the mark of jealousy. Rather, it is that the jealous wants to be valued *more* than anyone else. If they are the sole recipient of the beloved's sexual and emotional favors, they may be said to be loved more than anyone else, since no one else is getting those favors from the beloved. How should we assess such a desire? At the very least, it demands that the beloved close themselves off from other romantic opportunities. Such jealousy seems driven not so much by love or concern with the relationship as by egoism (see [Brunning 2020](#)). We can conclude that while Farrell's and Protasi's accounts present a plausible view of jealousy and its aptness conditions, they provide reasons to doubt its moral value.

To sum up, the accounts of the formal object of jealousy I have considered here all seem to suggest that the jealous person reacts to a threat to their privileged status with respect to the beloved, and aims to preserve that status from the encroachment of a rival. The moral accounts I have considered attempt to show that the jealous person has a moral claim on the beloved such that the jealous deserves to maintain their privileged status either because of existing expectations or because they have a right to be favored in this way. However, construing the formal object of jealousy as a moral property renders jealousy inapt because moral obligations cannot be grounded in social conventions, and because rights claims do not seem to apply to love and sexual desire. Social conventions cannot ground entitlement claims for exclusivity. Entitlement claims are not moral claims, and require further moral assessment.

The nonmoral accounts of the formal object of jealousy suggest that the jealous wants to be favored above all others, which is cashed out in terms of maintaining their privileged status or comparative advantage over others. While these accounts can provide for apt cases of jealousy, they bring out the ethically problematic nature of jealousy by showing that the jealous person is concerned with occupying a position of privilege which they aim to achieve through excluding everyone else. While romantic love is partial and cannot be directed towards everyone, the demands of jealousy are not justified by

---

<sup>20</sup> It might also be noted that on Kolodny's view, every relationship is trivially unique, inasmuch as no two different relationships could share a single history.



the partiality of love. The desire to be loved exclusively or to be loved more than anyone else is either based on a misconception about what it means to be valued as unique, or on a self-centered desire that does not cast the jealous in a favorable light.

The discussion of the formal object and aptness of jealousy sheds light on the intentionality of jealousy, on its representational content. How does this connect with the function of jealousy outlined by psychologists who say that its function is to identify and ward off threats from rivals? The intentional content of jealousy is supposed to fulfill the function of correctly identifying threats. What is not spelled out in psychological accounts is what that content amounts to. In particular, the examined psychological accounts say nothing about the privileged status that the lover is afraid to lose and aims to protect. Yet, it is clear that the threat to the relationship is understood by them in terms of maintaining an exclusive monogamous relationship. Therefore, what is threatened is the privileged status of the lover, understood as requiring exclusivity.<sup>21</sup> Furthermore, as should be clear, jealousy is different from other emotions that represent loss, such as sadness and grief (or even fear of a loss) because it is a rivalrous emotion. Thus, when describing the twofold function of jealousy, it is important to recognize its intentional content in order to capture its competitive nature.

#### 4 Romantic Norms and Aptness

As we have seen, the condition of exclusivity figures prominently in the discussion of the rationality of jealousy. It is often appealed to in order to show that jealousy is apt. In this last section, I consider the ways in which relationship norms influence the aptness norms of jealousy. I argue that the exclusivity norms infiltrate the aptness conditions of jealousy in monogamous relationships by specifying when and who counts as a rival. The “rivalry” conditions are determined differently in polyamorous relationships. I argue that the norms of polyamory provide fewer conditions for apt jealousy compared to monogamy.

Recall that the formal object of jealousy speaks to its representational content—representing the situation as a threat to one’s privileged status posed by a rival. It is apt when the threat is real and inapt when it is not. The threat

---

<sup>21</sup> This is true for Kristjánsson’s account as well, for he thinks that the lover deserves to be valued more than the rival.

is real when the lover could lose their privileged status. As we have seen, the privileged status in a monogamous relationship is cashed out in terms of sexual and emotional exclusivity between the lover and the beloved. Therefore, when the norms of exclusivity are violated, the privileged status of the lover is undermined.

The exclusivity conditions determine what counts as a threat, thereby informing the aptness conditions of jealousy. What about other kinds of relationships in which exclusivity is not one of the norms? There are numerous romantic relationship styles that are nonmonogamous.<sup>22</sup> Given the scope of this paper, I only consider the practice of polyamory—a form of ethical nonmonogamy in which individuals cultivate multiple romantic relationships with the consent of everyone involved.

Polyamorous relationships can take many different forms, and vary in degrees of sexual and emotional connection and intimacy. Some relationships have rigid hierarchical structures that specify the rules for primary and secondary partners. Primary partners might enjoy more intimacy and emotional connection than secondary partners. Typically, though not necessarily, primary partners spend more time with one another, run a joint household, and share financial resources. They also often have a direct influence on their primary partner's romantic activity with others by negotiating their rules of engagement with others. Other polyamorists have no such rules, and reject any kind of hierarchy.<sup>23</sup> They might still have nesting partners—partners with whom they live. But that is not necessarily an indication of a relationship priority. Others still form polyamorous families of which all members live together, engage with one another sexually and emotionally in various ways, and jointly co-parent all the children in the household.

What does jealousy look like in polyamorous relationships? When is it apt? Since the function of jealousy is to correctly detect and respond to threats that come from rivals, we need to identify conditions under which such threats are possible in polyamory. In a hierarchically-structured polyamorous relationship, the primary partner might be threatened by the secondary partner who might try to take their place, for the secondary partner might want to receive privileges of the primary partner from which they are excluded. However, generally, polyamorists do not consider other lovers to be rivals. The practice of polyamory rests on a number of principles that include honesty, open-

---

<sup>22</sup> They include swinging, certain instances of polygamy, group marriages, etc.

<sup>23</sup> This is called “relationship anarchy”. For discussion see Nordgren (2006), Barker and Langdrige (2010), and Heras Gómez (2018).

ness, communication, non-possessiveness, trust, and respect for the partners' autonomy. Other lovers, therefore, do not pose a threat to one's existing or potential relationships. Polyamorists value compersion—the feeling of joy one experiences when one's partner is made happy by another (DeSousa 2017; Brunning 2020).

Given these considerations, a threat posed by a rival is defined differently in monogamy and polyamory. In monogamy, the threshold for a threat is low—any potential mutual romantic interest between the beloved and a third party presents real danger to the privileged status of the lover. This is because love is perceived as either being possible or worthy only in a dyad. In polyamory the threshold for a threat is high—other lovers are not rivals, and, therefore, do not as such pose a threat to the lover. In monogamy one is likely to have numerous cases of apt and inapt jealousy because of the way in which interactions between the beloved and others are assessed. Since there are more possibilities of real threats, there are more opportunities for apt jealousy. Even if threats do not occur, one is likely to be more vigilant and engage in more mate guarding in a monogamous framework. By contrast, in polyamorous relationships there are fewer possibilities for apt jealousy since the ideology of polyamory rejects competitiveness and exclusivity. Nonetheless, polyamorists experience jealousy. Often jealousy can be *recalcitrant*—it occurs despite one's judgment that it is inappropriate (D'Arms and Jacobson 2003; Brady 2009; Döring 2015). Such an occurrence may be particularly prevalent for those who have transitioned from monogamous to polyamorous relationships. Most polyamorists are aware of the recalcitrance of jealousy; they learn to manage it in various ways.

Cases of apt jealousy are nonetheless possible in a polyamorous framework, especially in hierarchical polyamorous relationships.<sup>24</sup> Apt jealousy could also occur in cases where the lover has fallen out of love, and is pursuing someone else. In this case, one's "privileged status" would simply amount to being loved, rather than being loved *more* than others. Overall, given the polyamorous framework, other lovers of one's beloved are not rivals because they don't constitute a threat to one's relationship. In general, it rejects competition for a privileged position with respect to the beloved.

Jealousy aims to identify threats to one's privileged status. As I hope to have shown, the criteria for what counts as a threat is partly determined by the

---

<sup>24</sup> At the same time, hierarchical polyamorous relationships, and uneven distribution of time and attention could not trigger jealousy if everyone is happy with the arrangement.

norms of a particular romantic ideology. Social norms pertaining to romantic relationships infiltrate the aptness conditions of jealousy by specifying the threshold for threats from others. In monogamy the threat criteria are easy to satisfy, in polyamory, much less so.

## 5 Conclusion

When is jealousy appropriate? To answer this question, I have considered the twofold function of jealousy of correctly identifying a threat to the lover by a rival, and engaging in mate guarding in order to counter the threat. Given these functions, I have examined the value of jealousy from biological, social, and personal points of view. I have raised doubts about the value of jealousy in light of the inconclusive data regarding its contribution to relationship satisfaction, and its justification of violence disproportionately directed at women. Although it is possible that jealousy can sometimes be useful in helping partners maintain a relationship, it is difficult to determine the extent to which it does so reliably. Furthermore, there are better ways to maintain a fulfilling relationship such as communication, trust, respect, etc.


To zoom in on the nature of the threat to which jealousy is a response, and to explicate the relationship between jealousy and morality, I have examined a variety of ways in which the formal object of jealousy, the *jealousy-worthy*, could be defined. In specifying the formal object of jealousy, it became clear how dominant the norms of sexual and emotional exclusivity are in making sense of romantic jealousy.

If the formal object of jealousy is a moral property characterized as a *threat-to-moral-desert* or a *threat-to-one's-moral-rights*, jealousy is always inapt because social conventions of monogamy can never ground moral properties. The same is true if the formal object of jealousy is a *threat-to-one's-entitlement* because although that is not a moral property, the institution of monogamy is an informal institution, and cannot, therefore, ground strict entitlement.

If the formal object of jealousy is defined as a *threat-to-one's-privileged-status* or a *threat-to-one's-comparative-advantage*, jealousy is apt when that status is threatened and inapt when it is not. While characterizing the formal object of jealousy in this way allows for apt jealousy, I have questioned whether the emotion is morally praiseworthy. The desire to be loved more than everyone else is morally dubious, and it raises concerns about the person's character.

The painfulness of jealousy is intelligible when one assumes the monogamous framework, since it only allows for an exclusive dyad, and the beloved's new romantic interest may well indicate a loss of interest on the lover's part. In the monogamous ideology, love is a zero-sum game. Thus, protecting one's privileged status can be equated with protecting one's love status. This is why anyone in whom the beloved might express a romantic interest constitutes a threat to the lover. This is clear from the comparison of monogamy to polyamory. Polyamory sets a high bar for apt jealousy and discounts the majority of jealousy occurrences as recalcitrant because other romantic partners of one's beloved are not rivals, and therefore constitute no real threat to the lover. The moral problems raised by jealousy raise concerns about the moral standing of monogamy since it facilitates numerous occasions for apt or inapt jealousy.\*

Arina Pismenny

 0000-0001-8988-5121

University of Florida  
arinapismenny@ufl.edu

## References

- ALBRECHT, Ingrid V. 2017. "How We Hurt the Ones We Love." *Pacific Philosophical Quarterly* 98(2): 295–317, doi:10.1111/papq.12101.
- ANDERSEN, Peter, ELOY, Sylvie V., GUERRERO, Laura K. and SPITZBERG, Brian H. 1995. "Romantic Jealousy and Relational Satisfaction: A Look at the Impact of Jealousy Experience and Expression ." *Communication Reports* 8(2): 77–85, doi:10.1080/08934219509367613.
- ATTRIDGE, Mark. 2013. "Jealousy and Relationship Closeness: Exploring the Good (Reactive) and Bad (Suspicious) Sides of Romantic Jealousy." *Sage Open* 3(1), doi:10.1177/2158244013476054.

---

\* Special thanks to Julien Deonna for the conversation that led me to pursue this project. Thanks to the *Daily Nous* for the invitation to write a short piece that inspired some of the arguments made in this paper (2018). I am grateful for the useful questions and comments I received presenting parts of this work at the following conferences: "Emotions and Expressions" Workshop organized by the Expression, Communication, and the Origins of Meaning (ECOM) Research Group at the University of Connecticut (2018), the Canadian Philosophical Association annual meeting at the Université du Québec à Montréal (2018), the Intermountain Conference at the University of Utah (2018), the International Society for Research on Emotion at the University of Amsterdam (2019), and the Florida Philosophical Association annual meeting at the University of Florida (2019). Also, thanks to the anonymous referees who helped improve this article.

- AUMER, Katherine and ERICKSON, Michael A. 2022. "The Good and Bad of Love and Hate ." in *The Moral Psychology of Love*, edited by Arina PISMENNY and Berit BROGAARD, pp. 57–88. Lanham, Maryland: Rowman & Littlefield Publishers.
- BARELDS, Dick P. H. and BARELDS-DIJKSTRA, Pieternel. 2007. "Relations between Different Types of Jealousy and Self and Partner Perceptions of Relationship Quality." *Clinical Psychology & Psychotherapy* 14(3): 176–188, doi:10.1002/CP.532.
- BARKER, Meg and LANGDRIDGE, Darren. 2010. "Whatever Happened to Non-Monogamies? Critical Reflections on Recent Research and Theory." *Sexualities* 13(6): 748–772, doi:10.1177/1363460710384645.
- BELL, Macalester. 2013. *Hard Feelings. The Moral Psychology of Contempt*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199794140.001.0001.
- BEN-ZE'EV, Aaron. 1990. "Envy and Jealousy ." *Canadian Journal of Philosophy* 20(4): 487–516, doi:10.1080/00455091.1990.10716502.
- . 2010. "Jealousy and Romantic Love." in *Handbook of Jealousy. Theory, Research, and Multidisciplinary Approaches*, edited by Sybil L. HART and Maria LEGERSTEE, pp. 40–53. Oxford: Wiley-Blackwell.
- BRADY, Michael Sean. 2009. "The Irrationality of Recalcitrant Emotions." *Philosophical Studies* 145(3): 413–430, doi:10.1007/s11098-008-9241-1.
- BRAKE, Elizabeth. 2012. *Minimizing Marriage*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199774142.001.0001.
- , ed. 2016. *After Marriage. Rethinking Marital Relationships*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780190205072.001.0001.
- . "Is 'Loving More' Better? The Values of Polyamory." in *The Philosophy of Sex. Contemporary Readings*, edited by Raja HALWANI, Alan SOBLE, Sarah HOFFMAN, and Jacob M. HELD, 7th ed., pp. 201–220. Lanham, Maryland: Rowman & Littlefield Publishers.
- BROGAARD, Berit. 2020. *Hatred: Understanding Our Most Dangerous Emotion*. New York: Oxford University Press, doi:10.1093/oso/9780190084448.001.0001.
- BRUNNING, Luke. 2016. "The Distinctiveness of Polyamory ." *The Journal of Applied Philosophy* 35(3): 1–19, doi:10.1111/japp.12240.
- . 2020. "Compersion: An Alternative to Jealousy?" *The Journal of the American Philosophical Association* 6(2): 225–245, doi:10.1017/apa.2019.35.
- BUSS, David M. 2000. *The Dangerous Passion: Why Jealousy Is as Necessary as Love and Sex*. New York: The Free Press.
- . 2006. "The Evolution of Love." in *The New Psychology of Love*, edited by Robert J. STERNBERG and Karin WEIS, pp. 65–86. New Haven, Connecticut: Yale University Press.
- BUSS, David M. and SCHMITT, David P. 1993. "Sexual Strategies Theory: An Evolutionary Perspective on Human Mating ." *Psychological Review* 100(2): 204–321, doi:10.1037/0033-295X.100.2.204.

- BUUNK, Abraham P., DIJKSTRA, Pieterneel, LECKIE, Glenn and DIPOKARTO, Dascha. 2020. "Ethnic Differences in Jealousy in Surinam ." *Journal of Social and Personal Relationships* 37(4): 1136–1149, doi:10.1177/0265407519880287.
- CAI, Hua. 2001. *A Society without Fathers or Husbands: The Na of China*. New York: Zone Books. Translated by Asti Hustvedt.
- CANTO, Jesús M., ALVARO, José Luis, PEREIRA, Cícero, GARRIDO, Alicia, TORRES, Ana R. and PEREIRA, Marcos Emanuel. 2017. "Jealousy, Gender, and Culture of Honor: A Study in Portugal and Brazil ." *The Journal of Psychology* 151(6): 580–961, doi:10.1080/00223980.2017.1372344.
- CHUNG, Mingi and HARRIS, R., Christine. 2018. "Jealousy as a Specific Emotion: The Dynamic Functional Model." *Emotion Review* 10(4): 272–287, doi:10.1177/1754073918795257.
- CIHANGIR, Sezgin. 2013. "Gender Specific Honor Codes and Cultural Change ." *Group Processes & Intergroup Relations* 16(3): 319–331, doi:10.1177/1368430212463453.
- D'ARMS, Justin and JACOBSON, Daniel. 2000. "The Moralistic Fallacy: On the 'Appropriateness' of Emotions." *Philosophy and Phenomenological Research* 61(1): 65–90, doi:10.2307/2653403.
- . 2003. "The Significance of Recalcitrant Emotion (or, Anti-Quasijudgmentalism)." in *Philosophy and the Emotions*, edited by Anthony HATZIMOYSIS, pp. 127–146. Royal Institute of Philosophy Supplement n. 52. Cambridge: Cambridge University Press, doi:10.1017/S1358246100007931.
- DALY, Martin and WILSON, Margo I. 1988. *Homocide*. New York: Aldine-deGruyter.
- DAMASIO, Antonio R. 1994. *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: G.P. Putnam.
- DANDURAND, Cathy and LAFONTAINE, Marie-France. 2014. "Jealousy and Couple Satisfaction: A Romantic Attachment Perspective." *Marriage & Family Review* 50(2): 154–173, doi:10.1080/01494929.2013.879549.
- DEONNA, Julien Amos and TERONI, Fabrice. 2008. *Qu'est-ce qu'une émotion?* Chemins Philosophiques. Paris: Librairie philosophique Jean Vrin.
- . 2012. *The Emotions. A Philosophical Introduction*. London: Routledge. Substantially enlarged edition and translation of Deonna and Teroni (2008).
- DESOUSA, Ronald B. 1987. *The Rationality of Emotion*. Cambridge, Massachusetts: The MIT Press, doi:10.7551/mitpress/5760.001.0001.
- . 2017. "Love, Jealousy, and Compersion." In *Oxford Handbook of Philosophy of Love*, ed. Christopher Grau and Aaron Smuts. New York, NY: Oxford University Press. " in *The Oxford Handbook of the Philosophy of Love*, edited by Christopher GRAU and Aaron SMUTS. Oxford Handbooks. Oxford: Oxford University Press, doi:10.1093/oxfordhb/9780199395729.013.30.
- . 2018. "How to Think Yourself Out of Jealousy." in *Shadows of the Soul. Philosophical Perspectives on Negative Emotions*, edited by Christine TAPPOLET, Fabrice TERONI, and Anita KONZELMANN ZIV, pp. 132–142. London:

- Routledge. Revised and translated from Tappolet, Teroni and Konzelmann Ziv (2011).
- . 2019. "Is Contempt Redeemable?" *Journal of Philosophy of Emotion* 1(1): 24–43, doi:10.33497/jpe.v1i1.10.
- DESTENO, David A. and SALOVEY, Peter. 1996. "Evolutionary Origins of Sex Differences in Jealousy? Questioning the 'Fitness' of the Model." *Psychological Science* 7(6): 367–372, doi:10.1111/j.1467-9280.1996.tb0039.
- DÖRING, Sabine A. 2015. "What's Wrong with Recalcitrant Emotions? From Irrationality to Challenge of Agential Identity." *Dialectica* 69(3): 381–402. Special issue "Beyond Perceptualism," edited by Sabine A. Döring and Anika Lutz, doi:10.1111/1746-8361.12109.
- FARRELL, Daniel M. 1980. "Jealousy." *The Philosophical Review* 89(4): 527–559, doi:10.2307/2184735.
- FELDMAN, Fred and SKOW, Brad. 2020. "Desert." in *The Stanford Encyclopedia of Philosophy*. Stanford, California: The Metaphysics Research Lab, Center for the Study of Language; Information. Revision, September 22, 2020, of the version of October 9, 2015, <https://plato.stanford.edu/entries/desert/>.
- FRIJDA, Nico H. 1987. "Emotion, Cognitive Structure, and Action Tendency." *Cognition and Emotion* 1(2): 115–143, doi:10.1080/02699938708408043.
- GOLDMAN, Alvin I. 1986. *Epistemology And Cognition*. Cambridge, Massachusetts: Harvard University Press.
- GRAU, Christopher. 2004. "Irreplaceability and Unique Value." *Philosophical Topics* 32(1–2): 111–129, doi:10.5840/philtopics2004321/219.
- GUERRERO, Laura K. and ELOY, Sylvie V. 1992. "Relational Satisfaction and Jealousy Across Marital Types." *Communication Reports* 5(1): 23–31, doi:10.1080/08934219209367540.
- GUERRERO, Laura K., HANNAWA, Annegret F. and BABIN, Elizabeth A. 2011. "The Communicative Responses to Jealousy Scale: Revision, Empirical Validation, and Associations with Relational Satisfaction." *Communication Methods and Measures* 5(3): 223–249, doi:10.1080/19312458.2011.596993.
- HARRIS, Christine R. 2003. "A Review of Sex Differences in Sexual Jealousy, Including Self-Report Data, Psychophysiological Responses, Interpersonal Violence, and Morbid Jealousy." *Personality and Social Psychology Review* 7(2): 102–128, doi:10.1207/S15327957PSPR0702\_102-128.
- HART, Sybil L. 2010. "The Ontogenesis of Jealousy in the First Year of Life: A Theory of Jealousy as a Biologically-Based Dimension of Temperament ." in *Handbook of Jealousy. Theory, Research, and Multidisciplinary Approaches*, edited by Sybil L. HART and Maria LEGERSTEE, pp. 57–82. Oxford: Wiley-Blackwell.
- HENDRICK, Susan S. 1988. "A Generic Measure of Relationship Satisfaction." *Journal of Marriage and Family* 50(1): 93–98, doi:10.2307/352430.



- HERAS GÓMEZ, Roma de las. 2018. "Thinking Relationship Anarchy from a Queer Feminist Approach." *Sociological Research Online* 24(4): 644–660, doi:10.1177/1360780418811965.
- HUPKA, Ralph B. and RYAN, James M. 1990. "The Cultural Contribution to Jealousy: Cross-Cultural Aggression in Sexual Jealousy Situations." *Behavior Science Research* 24(1–4): 51–71, doi:10.1177/106939719002400104.
- JENKINS, Carrie S. I. 2017. *What Love Is: And What It Could Be*. New York: Basic Books.
- JOLLI MORE, Troy. 2011. *Love's Vision*. Princeton, New Jersey: Princeton University Press, doi:10.23943/princeton/9780691148724.001.0001.
- KENNY, Anthony John Patrick. 1963. *Action, Emotion and the Will*. London: Routledge & Kegan Paul. Second edition: Kenny (2003).
- . 2003. *Action, Emotion and the Will*. 2nd ed. London: Routledge. First edition: Kenny (1963), doi:10.4324/9780203711460.
- KOLODNY, Niko. 2003. "Love as Valuing a Relationship." *The Philosophical Review* 112(2): 135–189, doi:10.1215/00318108-112-2-135.
- KRISTJÁNSSON, Kristján. 2002. *Justifying Emotions. Pride and Jealousy*. Routledge Studies in Ethics and Moral Theory n. 3. London: Routledge.
- . 2018. *Virtuous Emotions*. Oxford: Oxford University Press, doi:10.1093/oso/9780198809678.001.0001.
- LANDMAN, Janet. 1993. *Regret: The Persistence of the Possible*. Oxford: Oxford University Press.
- LIAO, S. Matthew. 2015. *The Right To Be Loved*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780190234836.001.0001.
- MARAZZITI, Donatella, DELL'OSSO, Bernardo, BARONIM, Stefano, MUNGAI, Francesco, CATENA, Mario, RUCCI, Paola, ALBANESE, Francesco, et al. 2006. "A Relationship between Oxytocin and Anxiety of Romantic Attachment." *Clinical Practice and Epidemiology in Mental Health* 2(1), doi:10.1186/1745-0179-2-28.
- MATHES, Eugene W. and VERSTRAETE, Christine. 1993. "Jealous Aggression: Who Is the Target, the Beloved or the Rival?." *Psychological Reports* 72(3, suppl.): 1071–1074, doi:10.2466/pro.1993.72.3c.1071.
- MULLEN, Paul E. 1993. "The Crime of Passion and the Changing Cultural Construction of Jealousy." *Criminal Behaviour and Mental Health* 3(1): 1–11, doi:10.1002/cbm.1993.3.1.1.
- MULLEN, Paul E. and MAACK, Lara H. 1985. "Jealousy, Pathological Jealousy, and Aggression." in *Aggression and Dangerousness*, edited by David P. FARRINGTON and John Charles GUNN, pp. 103–126. Current Research in Forensic Psychiatry and Psychology. Hoboken, New Jersey: John Wiley; Sons, Inc.
- NEU, Jérôme. 1980. "Jealous Thoughts." in *Explaining Emotions*, edited by Amélie Oksenberg RORTY, pp. 425–463. Berkeley, California: University of California Press. Reprinted in Neu (2000, 41–67).

- . 2000. *A Tear is an Intellectual Thing: The Meanings of Emotion*. Oxford: Oxford University Press.
- NISBETT, Richard E. and COHEN, Dov. 1996. *Culture of Honor: The Psychology of Violence in the South*. New Directions in Social Psychology. Boulder, Colorado: Westview Press.
- NORDGREN, Andie. 2006. "The Short Instructional Manifesto for Relationship Anarchy." Unpublished manuscript, available from the webpage of the Anarchist Library, <https://acearchive.lgbt/artifacts/nordgren-manifesto-relationship-anarchy/>.
- NOZICK, Robert. 1989. *The Examined Life – Philosophical Meditations*. New York: Touchstone, Simon&Schuster.
- PFEIFFER, Susan M. and WONG, Paul T. P. 1989. "Multidimensional Jealousy." *Journal of Social and Personal Relationships* 6(2): 181–196, doi:10.1177/026540758900600203.
- PISMENNY, Arina. 2021. "The Amorality of Romantic Love." in *Love, Justice, and Autonomy: Philosophical Perspectives*, edited by Rachel FEDOCK, Michael KÜHLER, and Raja ROSENHAGEN, pp. 23–42. New York: Routledge.
- PRICE, Carolyn. 2018. "Grief." in *Shadows of the Soul. Philosophical Perspectives on Negative Emotions*, edited by Christine TAPPOLET, Fabrice TERONI, and Anita KONZELMANN ZIV, pp. 105–112. London: Routledge. Revised and translated from Tappolet, Teroni and Konzelmann Ziv (2011).
- . 2020. "The Many Flavours of Regret." *The Monist* 103(2): 147–162, doi:10.1093/monist/onzo32.
- PRINZ, Jesse J. 2004. *Gut Reactions. A Perceptual Theory of Emotion*. Philosophy of Mind Series. Oxford: Oxford University Press.
- PROTASI, Sara. 2017. "‘I’m Not Envious, I’m Just Jealous!’: On the Difference Between Envy and Jealousy." *The Journal of the American Philosophical Association* 3(3): 316–333, doi:10.1017/apa.2017.18.
- . 2019. "‘Mama, Do You Love Me?’: A Defense of Unloving Parents." in *The Routledge Handbook of Love in Philosophy*, edited by Adrienne M. MARTIN, pp. 35–46. Routledge Handbooks in Philosophy. Milton, Abingdon: Routledge.
- PUNTE, Sylvia and COHEN, Dov. 2003. "Jealousy and the Meaning (or Nonmeaning) of Violence ." *Personality and Social Psychology Bulletin* 29(4): 449–460, doi:10.1177/0146167202250912.
- RYDELL, Robert J. and BRINGLE, Robert G. 2007. "Differentiating Reactive and Suspicious Jealousy ." *Social Behavior and Personality: An International Journal* 35(8): 1099–1114, doi:10.2224/sbp.2007.35.8.1099.
- SALOVEY, Peter, ed. 1991. *The Psychology of Jealousy and Envy*. New York: Guilford Press.
- SCARANTINO, Andrea. 2017. "Do Emotions Cause Actions, and If So How? ." *Emotion Review* 9(4): 326–334, doi:10.1177/1754073916679005.~.

- SHEETS, Virgil L., FREDENDALL, Laura L. and CLAYPOOL, Heather M. 1997. "Jealousy Evocation, Partner Reassurance, and Relationship Stability: An Exploration of the Potential Benefits of Jealousy." *Evolution and Human Behavior* 18(6): 387–402, doi:10.1016/S1090-5138(97)00088-3.
- STEARNS, Peter N. 2010. "Jealousy in Western History: From Past toward Present ." in *Handbook of Jealousy. Theory, Research, and Multidisciplinary Approaches*, edited by Sybil L. HART and Maria LEGERSTEE, pp. 7–26. Oxford: Wiley-Blackwell.
- TAPPOLET, Christine. 2016. *Emotions, Values, and Agency*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199696512.001.0001.
- TAPPOLET, Christine, TERONI, Fabrice and KONZELMANN ZIV, Anita, eds. 2011. *Les Ombres de l'Âme. Penser les émotions négatives*. Genève: Éditions Markus Haller.
- VANDELLO, Joseph A. and COHEN, Dov. 2008. "Culture, Gender, and Men's Intimate Partner Violence." *Social and Personality Psychology Compass* 2(2): 652–667, doi:10.1111/j.1751-9004.2008.00080.x.
- VELLEMAN, David J. 1999. "Love as a Moral Emotion." *Ethics* 109(2): 606–628. Reprinted in Velleman (2006, 70–109), doi:10.1086/233898.
- . 2006. *Self to Self. Selected Essays*. Cambridge: Cambridge University Press, doi:10.1017/CBO9780511498862.
- WALLACE, Richard Jay. 2012. "Duties of Love." *Proceedings of the Aristotelian Society, Supplementary Volume* 86: 175–198, doi:10.1111/j.1467-8349.2012.00213.x.
- WHITE, Gregory L. and MULLEN, Paul E. 1989. *Jealousy: Theory, Research, and Clinical Strategies*. New York: Guilford Press.
- WILSON, Margo I. and DALY, Martin. 1996. "Male Sexual Proprietariness and Violence against Wives." *Current Directions in Psychological Science* 5(1): 2–7, doi:10.1111/1467-8721.ep10772668.
- WREEN, Michael J. 1989. "Jealousy ." *Noûs* 23(5): 635–652, doi:10.2307/2216005.

# The Legend of Hermann the Cognitive Neuroscientist

BRYCE GESSELL

I tell the tale of Hermann, a cognitive neuroscientist with transcendental aspirations. Hermann's story illustrates the fundamental problem of cognitive ontologies, which is the problem of knowing whether the background conceptual scheme for our psychological theories is correct. I show why this problem is so fundamental, how it arises from the nature of psychology as a science, and why various current approaches to solving it are not likely to be successful. The problem, I argue, pushes us toward instrumentalism about mental concepts and categories, in both psychology and cognitive neuroscience.

## 1 The Legend

Once upon a time, there was a cognitive neuroscientist named Hermann. Like his colleagues, Hermann read articles, applied for funding, and was a proficient neuroimager. He taught classes and went to department meetings. But unlike his colleagues, Hermann harbored a dark secret. It was a secret blacker than the coffee he drank while explaining the Libet studies to his undergrads for the fiftieth time. He dared not reveal the secret to anyone, even on his many fake Twitter accounts, lest the information somehow be traced back to him: Hermann was a Kantian.

Graduate school had been difficult for Hermann. A convert to transcendental philosophy at age 17, he didn't share his classmates' enthusiasm for cutting-edge theories of mental processes. He just couldn't see the point of devising or testing newfangled psychological concepts like "attentional control" and "reward-prediction error". After all, hadn't Kant already outlined the *true* psychology back in the 1780s? What more could the world want?

But nearly failing his first psychology courses taught Hermann never to disclose his true convictions, and so he dutifully read his textbooks and reproduced the "correct" answers on tests. He asked questions at department talks

to throw his supervisors off the trail. He gave papers on his lab's work at the APA and SPSP. Those results led to a dissertation on neuroeconomics, which he resented while writing and loathed after it got him a social neuroscience postdoc. Yet once again he did what he was supposed to, and scanned countless fMRI subjects while they watched videos of people talking and laughing. He always wrote his findings up on time and sent papers to well-targeted journals. Many were accepted, some even at prestigious venues.

In reality, though, Hermann was just biding his time. All through graduate school and his postdoc, Hermann pretended to believe in the constructs of contemporary psychology, but deep down, he was just waiting for the moment when he could follow his heart. At last, the years of hypocrisy and dissimulation paid off: his postdoc papers struck a chord with the right committees, his job talk had the perfect jokes ("Based on the work of my very warlike colleague, Sarah Bellum, we..."), and he dazzled the right group of faculty. Hermann landed two big grants and a tenure-track job. To celebrate he took a long walk down the lane near his house at precisely 3:30 pm. On the walk he contemplated his future and looked at sticks; he knew the real work was just beginning.

The next morning Hermann gathered his notes from countless magical nights with Immanuel and got his real research program underway. His goal was simple: find the neural correlates for all the major constructs in Kant's psychology. With his detailed knowledge of the first *Critique* and other texts, Hermann knew that his work would not involve conceptual difficulties. Methodology wasn't a problem either, since his grad-school education was more than sufficient. He was merely doing what every other cognitive neuroscientist did—he just happened to be doing it within a Kantian framework.

It was not hard to find the neural correlates for the faculty of judgment, for example. While inside the fMRI scanner, his subjects read and reflected on propositions. Hermann carefully counterbalanced the stimuli to control for effects like variable propositional content and emotional valence. To the surprise of no one, his statistical analyses showed that certain brain areas activated during the tasks. These crucial areas showed regular patterns both across subjects and across studies. In this way, he identified not only the cortical regions engaged in acts of judgment, but also the sub-regions which process the various logical forms of judgments. He connected judgments of quality to one area and judgments of relation to another, and judgments of quantity and modality to interconnected networks.

Hermann made similar discoveries about spatial representations, thereby illuminating the neural mechanisms of “the form of all appearances of outer sense.” (Kant 1998, A26/B42) His work produced the first map of the Kantian cortex.

No, day-to-day research was not the hard part. The real problem was time. For his arguments to be persuasive, Hermann knew that he needed a lot of data, and needed the computing cluster to analyze it with fancy statistics. But he also knew that he couldn’t let anyone find out what he was doing. So he stonewalled his colleagues when they asked about his results; he ignored emails from his department chair; he kept his unfortunate non-Kantian graduate students in the dark about the true import of their work. The tipping point came at his third-year tenure review. The neuroscience faculty was ready to give him marks for unsatisfactory progress, which would have been grounds for dismissal, but a letter of recommendation from Hermann’s postdoc director saved the day. Her letter assured department members of Hermann’s potential, promised that he would revolutionize the field, and urged them to retain him. Hermann barely survived the vote. He knew he needed to hurry—he had less than three years to go.

In the darkest times, when he doubted his life’s work and his goal appeared most distant, Hermann comforted himself by reading what Kant once wrote to Samuel Thomas Soemmering. In a 1795 letter, Kant spoke to the anatomist Soemmering about the sense organs in the brain. Sensory representations had to be combined, Kant said, and it was incumbent on natural philosophers to “render that unity comprehensible by reference to the structure of the brain.” (Kant 1999, 501) Hermann drew strength from his forebear’s prescient understanding of his own research program, as he attempted to show the neural correlates for *a priori* contributions to cognition. Hermann also knew he carried on Soemmering’s physiological work, to which Kant gave effusive praise, by finding the chemical mechanisms of the mental faculties.<sup>1</sup>

But the next three years passed, and since he released his results only in controlled trickles, Hermann simply did not publish enough to get tenure. It was understood in his department that he would have to leave after his seventh year. Then, at the start of that year, a miracle happened: stacks of finished manuscripts, all on the cognitive neuroscience of Kant’s transcendental philosophy, appeared on the department chair’s desk. All together, the manuscripts told a magnificent story of how the brain realized the posits

---

1 See Kant’s preface to Soemmering’s *On the Organ of the Soul* (Kant 2007, 222–226).

of Kant's psychological theory. Early papers laid the groundwork by finding brain areas for the most fundamental concepts, like the faculty of judgment, the forms of space and time, and the transcendental unity of apperception. Later work described connections among these concepts that even Kant had not noticed. Shorter papers filled in smaller details, and a single flagship paper—Hermann hoped to send it to *Neuron*—assembled the main results into a new, elegant, and powerful theory of the mind and brain. Always and everywhere, Hermann's results met and even exceeded accepted standards of experimental rigor and statistical significance.

At first, the chair was laughing as he leafed through the pile, imagining the pleasure he would feel at firing this Prussian charlatan. But the laughing stopped as he began to see the depth, creativity, and penetrating intuition with which Hermann had carried out his work. He convened a special faculty meeting to discuss the matter. On the one hand, Hermann had published nothing of note during his six years as assistant professor; on the other, he was now sitting on dozens of bold papers, each ready to submit. The chair asked the faculty for their opinions. "It's such a waste!" a recently-tenured associate professor yelled. "Seven years down the drain! This whole thing is a travesty, and a sham, and a mockery! I won't stand for it!" Many others agreed. But Hermann had his defenders, mostly among the older faculty. These full professors, now in the twilight of their careers, had seen countless psychological theories come and go. From their point of view, the conceptual framework of Hermann's research did not differ essentially from so many failed frameworks of the past.

In the end, Hermann's colleagues decided to give him a choice: he could either leave the university or back up his neuroscientific results with behavioral studies. The faculty supporting him were worried that Kant's view was too procrustean to be plausible in the modern age. They wanted to see behavioral results demonstrating that Kantian psychology could account for known complexities of human action. They did not think it could be done, but if Hermann were able to pull it off, they thought, they could not justify forcing him out.

Hermann felt he couldn't abandon his work now—not when he had come so far. So he designed an arc of behavioral studies to support Kant's psychology. Fortunately for him, behavioral results are faster and cheaper to get than neuroimaging, and Hermann Turk'd almost everything. In what became his *annus mirabilis*, Hermann completed his entire suite of studies, performed some requisite follow-ups, and wrote all his results before the end of the spring

semester. He even made original discoveries about the structure of cognition from a Kantian perspective (these he considered submitting to philosophy journals, but seriously, what's the point?). Once again, the chair showed up to work one day to find another pile of papers on his desk, showing how to implement Kantian psychology to describe all aspects of human behavior.

He convened a second meeting, and for a second time, the question divided the members of Hermann's department. Some continued to think that Kantian psychology was unworkable in principle, and that the idea of a "Kantian cognitive neuroscience" was a farce. Others felt that Hermann's body of work was, in many respects, comparable to that of other faculty that the department had tenured. But all agreed on what Hermann had set before them: a coherent, exhaustive, and radical alternative to the contemporary conceptual framework of cognitive neuroscience.

The behavioral work showed Kantian psychological concepts to be much more flexible than anyone had realized, and sufficient to account for human perception, action, and memory. The evidence for the Kantian constructs was every bit as relevant and rigorous as it was for anything else in psychology. The imaging work showed, moreover, that these concepts had clear and reliable neural correlates, and that multivariate analyses could predict their instantiation in a wide variety of tasks. Indeed, it was not a matter of weighing evidence at all, for the evidence was equal on both sides—Hermann's brain data and behavioral studies were beyond reproach. Nor was it that Hermann had shown how to make certain Kantian constructs work *within* contemporary psychology. Rather, he was in fact offering a complete *replacement* for *all* of contemporary psychology. This Hermann's colleagues understood, and it was the root of their complaint. Hermann's work formed a complete science of human behavior which was fundamentally incompatible with competing approaches—that is, with *their* approaches. And he did it all with a brilliance and Teutonic flair that no one had ever noticed in him.

Department members faced a stark choice: dismiss an apparent rising star ("einen aufgehenden Stern", joked an older faculty member no one liked), or tenure a Kantian. They abhorred both options. On the one hand, they could get rid of him. But doing so would be an indictment on their own careers, for Hermann had done everything they had, and just as well, only with a different set of cognitive concepts. They realized that had the history of psychology gone differently, they might have been Kantians too. On the other hand, granting him tenure would sanction Hermann's revival of transcendental psychology—and let's face it, no one wanted that. The empirical evidence was



equal on both sides, and the practical consequences were all bad. How could they decide?

In the end, however, Hermann spared them the trouble. Having been asked to speak in his own defense, he instead offered his resignation. The *annus mirabilis* had ironed out the last wrinkle in his work, he explained, and so he had achieved his goal. There was nothing left for him to do. His results were just as good as theirs or anyone else's—he knew it, they knew it, and he knew they knew it. Hermann rose from his chair, grabbed a few of the big cookies they always had at faculty meetings, and walked out into the sunset. No one ever heard from him again.

Thus the legend of Hermann, the Kantian cognitive neuroscientist, was born.

## 2 The Moral

Hermann's legend makes several important points about cognitive neuroscience. I'll elaborate on some of them here as well as on the philosophical issues involved. I'll also consider some objections to my framing and conclusions.

To begin, I am not the first to tell a tale like Hermann's. Bub (2000) gave a version of it using phrenology, and so did Poldrack (2010). Others have told it as well (Uttal 2001; Anderson 2015).

All these versions involve *cognitive ontologies*. A cognitive ontology is the set of entities, processes, and constructs in one's theory of cognition.<sup>2</sup> We should understand "cognition" broadly here, as including sensation, perception, consciousness, and any other mental process or phenomenon. So if our ability to remember a phone number by silent rehearsal requires a "phonological loop" (Baddeley and Hitch 1974), then the phonological loop belongs in our cognitive ontology.

Most disputes in psychology concern the details of a cognitive ontology: whether this or that entity belongs in it, or whether some entity has this or that property. Memory researchers debate, for example, whether consolidation is distinct from *reconsolidation* (Alberini and LeDoux 2013). Consolidation occurs when a memory becomes insensitive to disruption or change. But each time someone reactivates a memory, it becomes susceptible to interference

---

<sup>2</sup> See Poldrack (2010) and Janssen, Klein and Slors (2017) for similar definitions. Anderson (2015) uses the term "taxonomy".

again. Is this latter event also just consolidation, or is it a separate process with different temporal and mechanistic profiles (Lee, Nader and Schiller 2017)? Our answers to these questions determine part of our cognitive ontology, and we can ask similar questions across psychology.

In turn, most research programs in cognitive neuroscience deal with mappings between a cognitive ontology and brain structures. The mappings involve local questions about processes like reconsolidation, but also global ones about which neural structure types we should map to. Philosophical theories about mechanisms (Piccinini and Craver 2011) and large-scale data projects (Yarkoni et al. 2011) try to solve these problems.

The workflow of a typical research program in cognitive neuroscience begins with whatever constructs the currently accepted cognitive ontology contains. Researchers then design tasks that they believe will involve those constructs. Next, they have study participants perform the tasks while some recording technique, such as fMRI or EEG, measures their neural activity. The hope is to find activity that exceeds a certain threshold or survives some correction for multiple comparisons. Should they find it, researchers map the construct they started with to the area showing the activity. They can then claim that the construct “engages” or “recruits” neural activity in that area. If they are careful, they will condition their claims on the tasks used, for tasks are inescapable mediators of mappings between mind and brain.

The legend of Hermann, however, is not about projects such as these. It is not about local disputes in psychology, nor the details of some mind-brain mapping. Rather, it is about which cognitive ontology we should prefer at the *general level*. It questions why a research program in neuroscience should begin with constructs from the received ontology of contemporary psychology at all. Why not select from an altogether different cognitive ontology? The history of psychology offers many choices. Hermann’s tenure case also raises the possibility of the wholesale *replacement* of one cognitive ontology by another, where the replacing set of concepts is different from and even incompatible with the one replaced.

In short, the primary point behind Hermann’s legend concerns what I call the *fundamental problem of cognitive ontologies*. Most studies in psychology don’t touch this problem, for they work within an accepted ontology in order to refine it or fill in the details. The fundamental problem of cognitive ontologies is whether we should actually accept the received ontology, or prefer some other. Sure, a budding psychologist has practical reasons to reject Franz Joseph Gall’s phrenological concepts as she begins her career. Chief

among them is that she'll never get a job by studying things like "veneration" or "amativeness". However, her practical reasons do not solve the in-principle problem of choosing a cognitive ontology to begin with. She could start her research just as well with the constructs of Aristotle, Galen, Christian Wolff, or anyone else with a theory of mind.

We can also put it this way: the fundamental problem of cognitive ontologies is knowing whether the conceptual scheme structuring your ontology is the right one. The problem is determining whether you have the correct conceptual language *in general*, not just in particular cases.

Kant's psychology is one such conceptual language. So why not be like Hermann and adopt it, instead of contemporary cognitive science, as the scheme to structure our whole ontology? Instead of "consciousness" we could talk about the "transcendental unity of apperception", for example. Kant wrote, "[t]he transcendental unity of apperception is that unity through which all of the manifold given in an intuition is united in a concept of the object" (Kant 1998, B139).

Assuming this is true, we can imagine various ways in which the unity of apperception might break down. People with akinetopsia or motion blindness do not have smooth perceptions of motion—their visual experience of motion is frame-by-frame, as it were, with no perceived connection between the frames. A good Kantian hypothesis would be that akinetopsia results from failing to properly combine the sensible data in the manifold. We could study this phenomenon in many ways: we could get behavioral profiles of people with akinetopsia-like symptoms and correlate our findings with life histories (Ovsiew 2014); we could test lesion patients with similar deficits (Rizzo, Nawrot and Zihl 1995); we could try to induce akinetopsia via transcranial magnetic stimulation and disrupt normal apperception ourselves (Beckers and Hömberg 1992). There would be many other avenues to explore. Some will scoff at this suggestion, but the point is that I have just described a research arc that would carry someone to associate professor and beyond. The published results would look an awful lot like psychology papers now, except Kantian concepts and a Kantian cognitive ontology would structure them.

I could provide more examples to deepen the point, or outline fMRI studies that Hermann could have done to plumb the implementation of Kant's psychology. But the actual history of psychology furnishes us with more and more plausible examples than we could ever hope to invent. The cycle of theory-replacement in the history of psychology *is* the existence proof for an in-principle problem.

There is a temptation to believe that, because psychology is a “science” now, its current cognitive ontology must stand on firmer ground than past ones. Can’t we now draw sharper distinctions between different systems of memory? Don’t we have better information about exact temporal profiles? Aren’t we able to see better how entities in the ontology relate to each other? Yes, psychology does all this now, and it didn’t or even couldn’t do it in Kant’s day. But we should not therefore infer that the accepted ontology has better epistemic credentials. The reason that items in our cognitive ontology have those properties is just that we now do psychology in a way that encourages us to identify those properties. Had we been doing psychology in Germany in 1800, but with modern methods, we could have discovered the same “facts” about the posits of Wolffian and Kantian psychology. That we could identify those properties, however, says little about their reality.

As such, there is no doubt that a real-life Hermann would succeed in finding neural correlates for the faculty of judgment, as described in the legend. He would have no trouble finding consistent, statistically significant patterns. His studies could use classic psychological testing methods like additive factors and subtraction. These methods work regardless of the entities in our cognitive ontology. They are varieties of experimental and task design, and any “justification” they confer on gathered data is irrespective of that ontology.

The problem of cognitive ontologies does not emerge because of modern methods, though two other (independent) methodological issues exacerbate it. The first begins in psychology: it is not difficult to find significant results in human cognitive and behavioral testing. Human behavior is amenable to description by many conceptual languages, which is why the history of psychology is so rich with ideas. A part of the problem stems from current experimental techniques, but another part is more endemic to psychological practice (Meehl 1967). The second methodological problem comes from neuroscience. Brains will show neural activations to anything and everything, so the fact that we have found an activation is not in itself very remarkable. We couldn’t *not* have found an activation.

I will say a bit more about these two problems below, but they are not my primary concern. The legend of Hermann itself just illustrates the fundamental problem of cognitive ontologies and some associated philosophical issues. What, then, should we do about it?

Given the nature of psychology, I think the right move is to be instrumentalist about psychological theories. Earlier I spoke about the “right” ontology,

and finding the “correct” conceptual language. Human behavior and its neural basis may not be the kind of phenomena that allow true theories; it may just be that certain ontologies are better for certain situations. We could do cognitive neuroscience with one of many ontologies, but we pick the one that seems most useful for our purposes, whatever those may be.<sup>3</sup>

Not all practitioners of the mind-brain sciences want to go instrumentalist, however. Other ways to respond seek to carve out more room for realism and a “correct” ontology. Let’s look at some of them.

Adapting Anderson’s (2015) discussion of mind-brain mappings, we can distinguish three realist-motivated approaches to the fundamental problem of cognitive ontologies. The first, taken by the vast majority of psychologists and cognitive neuroscientists, is the *conservative* approach (Price and Friston 2005). This attitude assumes that the correct conceptual scheme is probably a lot like the one we have now, and so our cognitive ontology only requires local tweaking. The second approach is *moderate*. It attempts to let the brain decide which of two cognitive constructs is better. The third is the *radical* approach. It suggests a re-thinking of “the very foundations of psychology in light of evidence from neuroscience and evolutionary biology” (Anderson 2015, 70).

None of these three approaches to the challenge of cognitive ontologies necessitates realist commitments, though all three trend in that direction. All three suggest that there is a “true” ontology and that either we’ve already found most of it, or we at least know the way to get there. I’ll discuss each approach in more detail below, and then explain why I don’t find them very promising.

The first approach is conservative. It suggests that we already have most of the pieces for a true cognitive ontology—they’re just the constructs of contemporary psychology. This approach takes the apparent success of psychological science as evidence of the truth of its claims, and since those claims involve elements in an ontology, the elements must therefore exist.

The problem with the conservative response is that it begs the question against someone like Hermann. Hermann suggests replacing the current ontology with another one; to say we can’t do that, because the one we have now is true, assumes what Hermann denies.

It’s also wrong to think that the “success” of psychological science, or the fact that each published paper finds an effect, creates a problem for Hermann’s

---

3 This view shares something in common with the position outlined in Francken and Slors (2014).

Kantian view. Citing particular successful studies or even batches of them does not support conservatism. This is because the evidence for this or that current psychological theory is not thereby evidence for the background conceptual scheme in which those theories are framed and tested. As noted above, psychology is such that we cannot help but find evidence for virtually any construct we go looking for (Meehl 1967; Open Science Collaboration 2015). Thus finding evidence for some process says very little about the truth of the conceptual language describing that process. In other words, the reason we don't have empirical evidence for Kant's psychology is simply that no one has bothered to gather it yet. If a real-life Hermann ever comes around, he'll find all the evidence he could want, but he'd be no closer to establishing the reality of the Kantian cognitive ontology.

The second approach to the problem of cognitive ontologies is moderate. It uses brain data to adjudicate between competing or incompatible psychological constructs, thus letting the brain "speak for itself". The brain can do this in various ways. One is when competing cognitive categories make different predictions about their neural correlates. We can test these predictions by measuring brain activity during task conditions that involve the categories. Another way is through multivariate analyses, which use patterns of neural activations to predict cognitive constructs or representational categories of stimuli.

The moderate approach faces several challenges. For one, while brain data might be useful for comparisons between constructs, it cannot give an absolute measure of a construct's reality. This point leads to a more serious problem, which is that even brain data cannot adjudicate between entire conceptual schemes or whole cognitive ontologies. Indeed, the brain is a fit counterpart for psychology: it will always give us *some* evidence of whatever we test for. Bub (2000) and Poldrack (2010) used phrenology in their version of Hermann's tale because there is no question that phrenologists, had they used fMRI, would have found copious activations strongly correlated to their phrenological categories, and strongly predictive of those categories in multivariate studies. The same is true for Hermann's transcendental concepts, and for any other set of concepts we care to check: no matter what they are, we will find some neural signature of them—but *it does not follow that they are real*. Brain "data" or "evidence" usually aren't evidence for the reality of the mental construct being tested. This point seems to be either ignored or misunderstood by many philosophers and scientists.

Another way of putting the issue is to say that, while the moderate approach wishes to let the brain speak for itself, our neural organ can really only do so in a language that *we already understand*, where “we” are the designers and interpreters of experiments. If brain data is to shed light on human thought or behavior, we must interpret that data using cognitive concepts. Even the simplest interpretations therefore rely on entities in a cognitive ontology, even when those entities appear to be mere folk-psychological categories like perception, belief, or desire. Those basic categories also inform experiment and task design, as researchers use folk psychology to reach broad (albeit general) agreement on how psychological constructs, tasks, and experimental conditions relate.<sup>4</sup> That whole psychological apparatus forms a conceptual scheme for studying the mind and brain.

But if we bring to the brain a language we already understand—a worked-out cognitive ontology—then the moderate approach begs the question against Hermann no less than the conservative approach does. This criticism also applies to ontology construction if the analysis uses previously existing cognitive constructs to structure the data; such analyses comprise the majority of “data-driven” methods (Poldrack 2010; Yarkoni et al. 2011; Yeo et al. 2015; Tamar et al. 2016; Eisenberg et al. 2019; Genon et al. 2018; Bolt et al. 2020).

The third and final approach is the radical one. My objections to the first two approaches suggest that Hermann himself begs the question against current cognitive science since he brought a worked-out cognitive ontology of his own to studying the brain. But there are even more radical approaches that try to avoid begging the question. One example is Cisek (2019), who synthesizes a new cognitive ontology by analyzing the evolutionary history of simple behavioral systems. Another attempt is Pessoa, Medina and Desfilis (2021), who reject “standard mental terms” and instead found a new cognitive ontology with “complex, naturalistic behaviors”.

It’s too early to know whether projects like these will succeed. If a “true” cognitive ontology exists, these are our best bets to find it, because they throw out our current conceptual language and start with the evolutionary environment. There are other radical approaches that I think we can object to, however, so I will focus on those.

Other examples of the radical approach to cognitive ontologies use large data sets to find non-obvious dimensions or axes in brain activations. Call this the “latent structure” strategy (Yarkoni et al. 2011). I’ll discuss the strategy

---

4 I thank a reviewer for making this connection clear.

a bit and then present a problem for it, which applies in varying degrees to other radical approaches.

The latent structure strategy uses computational techniques to find structure in neural data. The assumption is that the data's latent dimensions may trace the contours of categories the brain itself uses to organize cognition. In this approach, the brain goes beyond playing arbiter for competing constructs to reveal a brand-new set of categories. For example, Chen et al. (2017) use independent component analysis (ICA) with resting-state fMRI data from hundreds of scans to identify four previously hidden brain networks. The authors dub them the “auditory”, “control”, “default mode”, and “visual” networks. Biswal, Mennes and Xi-Nian Zuo (2010) perform a similar analysis on resting state data, and Schaefer et al. (2018) use functional connectivity to produce a new cortical parcellation.

These analyses outdo Hermann's because they are based purely on brain measurements. You apply a technique like ICA and a robust structure emerges that may have been impossible to detect otherwise. Unlike every other approach, you need not bring anything to the table other than the data. Prior to identifying the structural contours, no part of any background conceptual scheme plays a role. This is another radical way of tackling the fundamental problem of cognitive ontologies, and perhaps another hope to avoid begging the question.

The challenge for latent structure strategies is *interpreting* what they find. Sure, Chen et al. (2017) find four separable networks. But where do the “auditory”, “control”, “default mode”, and “visual” labels come from? Why interpret the networks with that conceptual language, instead of some other?

Now, the source for the labels is, of course, the authors' prior knowledge of similar networks. Chen et al. (2017) know that, in previous studies, participants who engaged in tasks requiring cognitive control showed activation patterns matching one of the networks they discovered. The authors then import those labels—those entities in the background cognitive ontology—into their own study, and use them to interpret the data. So even though the data's structure is *discovered* ontology-free, it can only be *interpreted* by some existing ontology or conceptual scheme. Just as we saw with the moderate approach, the brain can only speak in a language we already understand. The lesson is that big data may help introduce new neural categories, but it doesn't and can't provide the psychological labels for those categories.

Jerry Fodor and Ernie Lepore (1992, 1996) once developed a similar objection to Paul Churchland's semantic theory. Churchland (1989, 1998) devel-



oped a theory of meaning in which different aspects of conceptual content were represented by different dimensions in a high-dimensional neuronal activation space. So, to use a simplified example, the concept “dog” might be represented by neural activations along dimensions like “furriness”, “barking-ness”, “four-footed”, and so on. Various ranges of those dimensions define a high-dimensional solid that constitutes the concept “dog”.

The crux of Fodor and Lepore’s objection is that Churchland begs the question about the labels on the dimensions. Why does the first dimension in the activation space represent “furriness” instead of “barking-ness”, or something else entirely? By taking the labels for granted, Churchland smuggles semantic terms into a theory that is supposed to explain how there could be semantics in the first place.

Latent structure strategies make the same mistake. Why is this particular structure the “control” network, and that structure the “default mode” network? Labeling the networks requires interpreting the data, but interpretation only happens through cognitive concepts we already have. In trying to discover the brain’s categories for cognition, we smuggle in the psychological labels, and so accomplish nothing other than putting old wine into new bottles.

In sum, I see the fundamental problem of cognitive ontologies as leading us toward instrumentalism about psychology. Although there are realist-friendly responses to this problem, most of them take the items in their cognitive ontology for granted, and we can’t yet evaluate the ones that don’t.

The moral of Hermann’s legend is the problem I’ve been discussing, which connects to many issues in the philosophy of mind and of various sciences. Other than inertia and the vicissitudes of history, we have much less reason than we like to believe to prefer current cognitive ontologies over possible alternatives. And, as Bub (2000) notes, without some resolution for this problem,

[we cannot] differentiate what is currently undertaken [in cognitive neuroscience] from a pointless activity in which inevitable differences between experimental and baseline conditions are falsely attributed specific cognitive interpretations that do not in fact correspond to reality (Bub 2000, 470).

I conclude by considering some objections to my arguments and the way I’ve set them up. First, you might say that this is all just a problem of reverse

inference. Suppose my neuroimaging study discovers activation in brain area *X*. From previous studies, I know that *X* is associated with emotion, and so I infer that my subjects used emotional processing in my task, even though the task didn't explicitly involve emotion. This pattern of reasoning is called a reverse inference (Poldrack 2006). Reverse inferences require caution because area *X* could be involved in many other cognitive processes, not just emotion.

The problem of cognitive ontologies is not one of reverse inference, however. Reverse inferences have to do with evidence, and gathering more evidence alone does nothing to solve the problem. We have an enormous amount of papers published in cognitive psychology, but the sheer number does not resolve the in-principle problem of ontology selection.

A second objection could be that we could solve the problem with multivariate analyses in neuroscience. Both philosophers and neuroscientists sometimes believe that multivariate pattern analysis (MVPA), representational similarity analysis (RSA), and other multivariate techniques yield some special insight into brain function that ordinary univariate imaging analyses cannot (Nathan and Del Pinal 2017). I am skeptical of that view, but even if it were true, it would be irrelevant to my arguments. The problem of cognitive ontologies is not a methodological one—at least, not one internal to psychology or cognitive neuroscience as they are currently constituted. As I said above, certain methodological issues do exacerbate the problem, such as the ease with which we find publishable results in the mind-brain sciences. But it is not the current methods of psychology and cognitive neuroscience that give rise to the problem. It goes beyond the conceptual boundaries of either field and so we cannot solve it with more sophisticated statistics.

Someone might also object that the problem of cognitive ontologies is really an issue of underdetermination of theory by data (Aktunc 2021). According to this objection, alternative ontologies only look like live options because we don't yet have enough evidence for our current one. But this objection also says that psychological theories are *theories*, and as such, they will always go beyond the data. Every theory in every science outstrips the available observations, and it's unfair to expect a cognitive ontology to be an exception. This objection can therefore say that the problem of cognitive ontologies is not an issue of principle; it's just the expected result of humans doing psychological science.

This objection is a sophisticated one. To lay out and respond to all the issues involved would take another paper. Here I will just give some reasons to think

that the problem of cognitive ontologies goes beyond the underdetermination of theory by data.

As we've seen, Hermann wasn't going to convince anyone of Kant's psychology, no matter how much his evidence "determined" his theory. While Hermann's work isn't real, the cycle of theory replacement in the history of psychology is, and we have no reason to think that the cycle will stop with something like our current cognitive ontology. Superficial similarities between psychology and other sciences, such as that they are practiced in universities and use quantified measurements and mathematical analyses, give the impression that psychology, like physics or chemistry, trods a monotonic path up the mountain of truth. But those similarities belie deep conceptual and interpretational problems which may be inevitable not only in psychology but also in the phenomena it studies.

In describing human behavior and mentality, we face a situation in which many distinct but mutually incompatible conceptual schemes could do the job. It isn't just the history of psychology that shows this; current cross-cultural psychology does too. Take "indigenous" or "local" psychological theories, which describe human thought and behavior in specific cultural contexts (Allwood and Berry 2006). Rather than fitting received psychological categories to non-Western peoples, indigenous psychologies develop new categories tailored to their environment. Inputs to this development include literature, observations of behavior, self-reports, and past scientific evidence (Cheung et al. 1996). The results are psychological theories that may account for patterns of thought and behavior better than traditional (Western) theories.

One of the most empirically successful indigenous psychologies is the Chinese Personality Assessment Inventory, now known as the Cross-cultural Personality Assessment Inventory (CPAI). In addition to categories from the standard five-factor personality model, the CPAI includes psychological constructs like "Harmony", "Ren Qing" (relationship orientation), "Ah-Q Mentality" (defensiveness), and "Face" (Cheung et al. 2001). These constructs constitute a personality factor, "Interpersonal Relatedness", which is not reducible to other personality theories (Cheung et al. 2003).


If "Interpersonal Relatedness" and associated constructs like "Ren Qing" and "Ah-Q Mentality" are incompatible with other psychological theories, then what do we say about the state of the science? Underdetermination suggests that we're just lacking the evidence to decide between them, whether or not psychology is capable of providing it. But it's not a leap to think there may be some real indeterminacy here, and that there simply is no fact about

whether “Ren Qing” is real. We can study it, we can use it, and we can endorse it, but we don’t need to conclude it must exist.

There are indefinitely many conceptual schemes for psychology, limited only by our imagination. Whatever they are like, the brain will oblige with consistent profiles of activation. If the data underdetermines all the available theories to the same degree, then maybe the problem lies not in our ability to gather evidence but in the *Dinge an sich*.

One final objection. In a “no-miracles” spirit, one may say that our current ontology can’t be *that* wrong, since psychology and neuroscience are so successful. To those with the courage to make this response: I envy your faith, but see no reason to share it.\*

Bryce Gessell

 0000-0003-4424-5627

Southern Virginia University

bryce.gessell@svu.edu

## References

- AKTUNC, M. Emrah. 2021. “Productive Theory-Ladenness in fMRI.” *Synthese* 198(3): 7987–8003, doi:[10.1007/s11229-019-02125-9](https://doi.org/10.1007/s11229-019-02125-9).
- ALBERINI, Cristina M. and LEDOUX, Joseph E. 2013. “Memory Reconsolidation.” *Current Biology* 23(17): R746–R750, doi:[10.1016/j.cub.2013.06.046](https://doi.org/10.1016/j.cub.2013.06.046).
- ALLWOOD, Carl Martin and BERRY, John W. 2006. “Origins and Development of Indigenous Psychologies: An International Analysis.” *International Journal of Psychology* 41(4): 243–268, doi:[10.1080/00207590544000013](https://doi.org/10.1080/00207590544000013).
- ANDERSON, Michael L. 2015. “Mining the Brain for a New Taxonomy of the Mind.” *Philosophy Compass* 10(1): 68–77, doi:[10.1111/phc3.12155](https://doi.org/10.1111/phc3.12155).
- BADDELEY, Alan D. and HITCH, Graham. 1974. “Working Memory.” *Psychology of Learning and Motivation* 8: 47–89, doi:[10.1016/S0079-7421\(08\)60452-1](https://doi.org/10.1016/S0079-7421(08)60452-1).

---

\* This paper began as a meme and has turned into something much more serious, mostly thanks to the members of the Imagination and Modal Cognition Lab and the SPACE lab at Duke. In particular, I thank the PIs of those two labs, Felipe De Brigard and Jenni Groh, for their support over so many years. Thank you to Derek Haderlie for his help. I would also like to thank an anonymous reviewer from another journal, and the various anonymous reviewers at *Dialectica*, who helped me make major improvements to the paper (and were willing to review such a bonkers paper in the first place). Most of all, I want to thank all my students at SVU—and especially those in my Minds, Brains, and Neuroscience course for spring semester 2022. To Nate, Nick, Hannah, Caroline, Drake, Allie, Elizabeth, Madelyn, Brett, Rachel, Katie, Caleb, Sabra, Gunner, Isaiah, Carter, Tay, Clayton, and Tyson: you’re the best—and if this doesn’t convince you, nothing will!

- BECKERS, Gabriel J. L. and HÖMBERG, Volker. 1992. "Cerebral Visual Motion Blindness: Transitory Akinetopsia Induced by Transcranial Magnetic Stimulation of Human Area v5." *Proceedings of the Royal Society of London. Series B: Biological Sciences* 249(1325): 173–178, doi:10.1098/rspb.1992.0100.
- BISWAL, Bharat B., MENNES, Maarten and XI-NIAN ZUO, Clare Kelly, Suril Gohel. 2010. "Toward Discovery Science of Human Brain Function." *Proceedings of the National Academy of Sciences of the U.S.A.* 107(10): 4734–4739, doi:10.1073/pnas.0911855107.
- BOLT, Taylor, NOMI, Jason S., ARENS, Rachel, VIJ, Shruti G., RIEDEL, Michael, SALO, Taylor, LAIRD, Angela R., EICKHOFF, Simon B. and UDDIN, Lucina Q. 2020. "Ontological Dimensions of Cognitive-Neural Mappings." *Neuroinformatics* 18(3): 451–463, doi:10.1007/s12021-020-09454-y.
- BUB, Daniel N. 2000. "Methodological Issues Confronting PET and fMRI Studies of Cognitive Function." *Cognitive Neuropsychology* 17(5): 467–484, doi:10.1080/026432900410793.
- CHEN, Shaojie, HUANG, Lei, QIU, Huitong, NEBEL, Mary Beth, MOSTOFKY, Steward H., PEKAR, James J., LINDQUIST, Martin A., EOLOYAN, Ani and CALFO, Brian S. 2017. "Parallel Group Independent Component Analysis for Massive fMRI Data Sets." *PLoS One* 12(3), doi:10.1371/journal.pone.0173496.
- CHEUNG, Fanny M., CHEUNG, Shu Fai, WADA, Sayuri and ZHANG, Jianxin. 2003. "Indigenous Measures of Personality Assessment in Asian Countries: A Review." *Psychological Assessment* 15(3): 280–289, doi:10.1037/1040-3590.15.3.280.
- CHEUNG, Fanny M., LEUNG, Kwok, FAN, Ruth M., SONG, Wei-Zheng, ZHANG, Jian-Xin and ZHANG, Jian-Ping. 1996. "Development of the Chinese Personality Assessment Inventory." *Journal of Cross-Cultural Psychology* 27(2): 181–199, doi:10.1177/0022022196272003.
- CHEUNG, Fanny M., LEUNG, Kwok, ZHANG, Jian-Xin, SUN, Hai-Fa, GAN, Yi-Qun, SONG, Wei-Zhen and XIE, Dong. 2001. "Indigenous Chinese Personality Constructs: Is the Five-Factor Model Complete?" *Journal of Cross-Cultural Psychology* 32(4): 407–433, doi:10.1177/0022022101032004003.
- CHURCHLAND, Paul M. 1989. *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*. Cambridge, Massachusetts: The MIT Press.
- . 1998. "Conceptual Similarity Across Sensory and Neural Diversity: The Fodor/Lepore Challenge Answered." *The Journal of Philosophy* 95(1): 5–32, doi:10.2307/2564566.
- CISEK, Paul. 2019. "Resynthesizing Behavior Through Phylogenetic Refinement." *Attention, Perception, & Psychophysics* 81(7): 2265–2287, doi:10.3758/s13414-019-01760-1.
- EISENBERG, Ian W., BISSETT, Patrick G., ENKAVI, A. Zeynep, LI, Jamie, MACKINNON, David P., MARSCH, Lisa A. and POLDRACK, Russell A. 2019. "Uncovering the

- Structure of Self-Regulation through Data-Driven Ontology Discovery.” *Nature Communications* 10(1): 2319, doi:10.1038/s41467-019-10301-1.
- FODOR, Jerry A. and LEPORE, Ernest. 1992. *Holism: A Shopper's Guide*. Oxford: Basil Blackwell Publishers.
- . 1996. “All at Sea in Semantic Space: Churchland on Meaning Similarity.” *The Journal of Philosophy* 93(8): 381–403. Reprinted in Fodor and LePore (2002, 174–199), doi:10.2307/2564628.
- . 2002. *The Compositionality Papers*. Oxford: Oxford University Press.
- FRANCKEN, Jolien C. and SLORS, Marc. 2014. “From Commonsense to Science and Back: The Use of Cognitive Concepts in Neuroscience.” *Consciousness and Cognition* 29: 248–258, doi:10.1016/j.concog.2014.08.019.
- GENON, Sarah, REID, Andrew, LANGNER, Robert, AMUNTS, Katrin and EICKHOFF, Simon B. 2018. “How to Characterize the Function of a Brain Region.” *Trends in Cognitive Science* 22(4): 350–364, doi:10.1016/j.tics.2018.01.010.
- JANSSEN, Anelli, KLEIN, Colin and SLORS, Marc. 2017. “What is a Cognitive Ontology, Anyway?” *Philosophical Explorations: An International Journal for the Philosophy of Mind and Action* 20(2): 123–128, doi:10.1080/13869795.2017.1312496.
- KANT, Immanuel. 1998. *Critique of Pure Reason*. The Cambridge Edition of the Works of Immanuel Kant. Cambridge: Cambridge University Press. Translated and edited by Paul Guyer and Allen W. Wood.
- . 1999. *Correspondence*. The Cambridge Edition of the Works of Immanuel Kant. Cambridge: Cambridge University Press. Edited by Arnulf Zweig.
- . 2007. *Anthropology, History, and Education*. The Cambridge Edition of the Works of Immanuel Kant. Cambridge: Cambridge University Press. Edited by Günter Zöller and Robert B. Louden, doi:10.1017/CBO9780511791925.
- LEE, Jonathan L. C., NADER, Karim and SCHILLER, Daniela. 2017. “An Update on Memory Reconsolidation Updating.” *Trends in Cognitive Science* 21(7): 531–545, doi:10.1016/j.tics.2017.04.006.
- MEEHL, Paul E. 1967. “Theory-Testing in Psychology and Physics: A Methodological Paradox.” *Philosophy of Science* 34(2): 103–115, doi:10.1086/288135.
- NATHAN, Marco J. and DEL PINAL, Guillermo. 2017. “The Future of Cognitive Neuroscience? Reverse Inference in Focus.” *Philosophy Compass* 12(7), doi:10.1111/phc3.e12427.
- OPEN SCIENCE COLLABORATION. 2015. “Estimating the Reproducibility of Psychological Science.” *Science* 349(6251), doi:10.1126/science.aac4716.
- OVSEW, Fred. 2014. “The Zeitraffer Phenomenon, Akinetopsia, and the Visual Perception of Speed of Motion: A Case Report.” *Neurocase* 20(3): 269–272, doi:10.1080/13554794.2013.770877.

- PESSOA, Luiz, MEDINA, Loreta and DESFILIS, Ester. 2021. "Mental Categories and the Vertebrate Brain: the Neural Basis of Behavior." OSF preprint, doi:10.31219/osf.io/8cmhg.
- PICCININI, Gualtiero and CRAVER, Carl F. 2011. "Integrating Psychology and Neuroscience: Functional Analyses as Mechanism Sketches." *Synthese* 183(3): 283–311, doi:10.1007/s11229-011-9898-4.
- POLDRACK, Russell A. 2006. "Can Cognitive Processes be Inferred from Neuroimaging Data? ." *Trends in Cognitive Science* 10(2): 59–63, doi:10.1016/j.tics.2005.12.004.
- . 2010. "Mapping Mental Function to Brain Structure: How Can Cognitive Neuroimaging Succeed? ." *Perspectives on Psychological Science* 5(6): 753–761, doi:10.1177/1745691610388777.
- PRICE, Cathy J. and FRISTON, Karl J. 2005. "Functional Ontologies for Cognition: The Systematic Definition of Structure and Function." *Cognitive Neuropsychology* 22(3): 262–275, doi:10.1080/02643290442000095.
- RIZZO, Matthew, NAWROT, Mark and ZIHL, Josef. 1995. "Motion and Shape Perception in Cerebral Akinetopsia." *Brain: A Journal of Neurology* 118(5): 1105–1127, doi:10.1093/brain/118.5.1105.
- SCHAEFER, Alexander, KONG, Ru, GORDON, Evan M., LAUMANN, Timothy O., ZUO, Xi-Nian, HOLMES, Avram J., EICKHOFF, Simon B. and YEO, B. T. Thomas. 2018. "Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI." *Cerebral Cortex* 28(9): 3095–3114, doi:10.1093/cercor/bhx179.
- TAMAR, Diana I., THORNTON, Mark A., CONTRERAS, Juan Manuel and MITCHELL, Jason P. 2016. "Neural Evidence that Three Dimensions Organize Mental State Representation: Rationality, Social Impact, and Valence." *Proceedings of the National Academy of Sciences of the U.S.A.* 113(1): 194–199, doi:10.1073/pnas.1511905112.
- UTTAL, William R. 2001. *The Limits of Phenology. The Limits of Localizing Cognitive Processes in the Brain*. Cambridge, Massachusetts: The MIT Press.
- YARKONI, Tal, POLDRACK, Russell A., NICHOLAS, Thomas E., VAN ESSEN, David C. and WAGER, Tor D. 2011. "Large-Scale Automated Synthesis of Human Functional Neuroimaging Data." *Nature Methods* 8(8): 665–670, doi:10.1038/nmeth.1635.
- YEO, B. T. Thomas, KRIENEN, Fenna M., EICKHOFF, Simon B., YAAKUB, Siti N., FOX, Peter T., BUCKNER, Randy L., ASPLUND, Christopher L. and CHEE, Michael W. L. 2015. "Functional Specialization and Flexibility in Human Association Cortex." *Cerebral Cortex* 25(10): 3654–3672, doi:10.1093/cercor/bhu217.

# Alternative Possibilities and the Meaning of ‘Can’

MARIA SEKATSKAYA & GERHARD SCHURZ

Our account of free will integrates a counterfactual conditional analysis of abilities with a Frankfurt-style sourcehood psychological approach and is compatible with both determinism and indeterminism. It effectively addresses criticisms of the conditional analysis of “can” by demanding coherence between agents’ free actions and their personality frames. The paper begins by discussing conditional analyses of abilities, followed by an exploration of three strategies to counter the consequence argument: first, by assuming determinism with a backtracking analysis of counterfactuals; second, by assuming determinism with a local miracle analysis of counterfactuals; and third, by assuming indeterminism. We further demonstrate that the first two strategies we propose are immune to the criticisms faced by other conditional accounts. Moreover, we show that the third strategy effectively solves the luck problem. The paper concludes by affirming the reality of free will and its consistency with a naturalistic worldview.

## 1 Alternative Possibilities and Conditional Analysis of Abilities

There is a wide agreement in the free will debate that having free will implies possessing the capacity to choose one’s course of action. The natural reading of “choosing” seems to demand that an agent choose between alternative possibilities. The contested question, however, is how to interpret these alternative possibilities. Are there alternative possibilities in a deterministic world? Incompatibilists argue that determinism precludes alternative possibilities, and is, therefore, incompatible with free will. This reasoning can be shortly summarized as follows. An agent’s act is free only if it is in the agent’s power (up to the agent) to choose to act in one way or another, and to act in this way:

- (1)  $x$  acted freely only if  $x$  could have done otherwise.



Henceforth, we abbreviate the thesis “ $x$  could have done otherwise” as CDO. CDO implies that agent  $x$  has alternative possibilities of the right sort (at some time not later than the time of the agent’s action) so that he can choose and perform alternative possible actions. Can CDO be true in a deterministic world? According to incompatibilists,

(2) Physical determinism rules out any alternative possibilities

because determinism (D) is defined as the thesis that “there is at any instant exactly one physically possible future” (van Inwagen 1983, 3). Hence, if D is true, CDO is false: in a deterministic world no one acts freely.

Compatibilists can reject either thesis (1) or thesis (2). Some compatibilists have rejected (1) (cf. Dennett 1984; Frankfurt 1969). However, this move is rather radical, because denying our capacity to choose otherwise looks suspiciously close to denying free will outright. The classical compatibilist move is to reject (2), which can be done by reading CDO as a conditional statement: “ $x$  could have done otherwise” means “ $x$  would have done otherwise if a certain condition  $C$  obtained”.

The conditional analysis of freedom of will was initially proposed by David Hume (1748), and later developed by G.E. Moore (1912), Dickinson Miller (Miller 1934), and Alfred Ayer (1954). It enjoyed wide acceptance among naturalistically inclined analytic philosophers until John Austin’s (1961) and Keith Lehrer’s (1968) criticisms showed that the versions of the conditional analysis that had been provided so far were flawed. However, if one wants to demonstrate that incompatibilism is right, it is not enough to show that some versions of conditional analysis are wrong. Although thesis (2) might seem intuitively true, if some version of conditional analysis succeeds, (2) will turn out false. Incompatibilists must show that there are strong reasons to believe that physical determinism and alternative possibilities are incompatible. This is the aim of the so-called “consequence argument” (CA), first published by Carl Ginet (1966) and Peter van Inwagen (1975).

Before criticizing the CA, which we do in sections 2 and 3 of our paper, in this first section we give a brief review of the recent theories of a classical compatibilist style. These theories propose a conditional analysis of CDO along the following lines: an agent could have done otherwise if he had an ability such that, if condition  $C$  obtained, and he tried to use this ability, he would have succeeded. After that, we will clarify the notion of ability we rely on. In section 3 we propose three ways to reject the CA: by assuming (i)

determinism with backtracking analysis of counterfactuals, (ii) determinism with local miracle analysis of counterfactuals, and (iii) indeterminism. In section 4 we present our backtracking compatibilist analysis of abilities to do otherwise. In sections 5 and 6 we explain in more detail why our backtracking compatibilist account does not have the problems that some other conditional accounts have. In the rest of the paper, we present our local miracle and our indeterministic compatibilist analyses of abilities to do otherwise, and show that they effectively solve the randomness objection and the luck problem.

“New dispositionalist” compatibilists explain agents’ abilities in terms of dispositions to give a certain response to the stimulus of their own trying (Vihvelin 2004, 2013; Fara 2008). While we do agree that agents’ abilities can be analyzed in terms of dispositions to give certain responses to particular stimuli that are partly constituted by some relevant psychological state of the agent, we don’t assert that it is necessarily a stimulus of the agent’s own trying since it has been shown that in some cases the analysis in terms of trying is problematic (Franklin 2011; Kittle 2015b). In our explanation of abilities, we will follow David Lewis (1997), who connected dispositions to give responses to certain stimuli with intrinsic properties of the bearers of these dispositions. In order to avoid the problem with Finkish dispositions or Finkish lack of dispositions, Lewis introduced a time interval during which the intrinsic properties of the bearers of these dispositions should not change. Although Lewis himself did not explicitly use his analysis of dispositions to explain agents’ abilities, it can, in our opinion, quite naturally be extended in this way. We assert that an agent’s abilities are a specific class of the agent’s dispositions to act in particular ways in particular circumstances, where these acts are partly caused by the agent’s intrinsic psychological and physical properties, such as the agent’s skills, beliefs, desires, etc. Thus, we propose the following definition of having an ability at a time:

**ABILITY.** An agent  $x$  has at time  $t$  the ability to do  $A$  iff

- (a)  $x$  has an intrinsic property  $B$  between  $t$  and some later time point  $t'$ , and
- (b) if certain conditions  $C_i$  ( $i = 1, 2, \dots$ ) would obtain between  $t$  and later times  $t_i \leq t'$ , then  $C_i$  and  $x$ ’s having of  $B$  would jointly be an  $x$ -complete

cause of  $x$ 's doing  $A$  (where "an  $x$ -complete cause" is "a cause complete in so far as properties intrinsic to  $x$  are concerned").<sup>1,2</sup>

What the conditions  $C_i$  and property  $B$  are depends on the ability in question. If we consider the ability to play the violin, then the conditions  $C_i$  consist of proper external circumstances (e.g., having a violin at hand, etc.) plus a proper internal psychological stimulus on the part of the agent, for example, his decision to play or his desire to play the violin; note that different conditions have to endure for different time spans after time  $t$ . Property  $B$ , on the other hand, consists of the agent's skills, beliefs, etc. Our account will show how, given any suitable understanding of abilities along the lines above, one can explain the abilities to do otherwise, within the framework of either physical determinism or indeterminism. We will use this definition of ability in our own account of CDO in sections 4–8 of this paper, where we will explain what the conditions  $C_i$  are for the abilities to do otherwise, by using the framework of possible world analysis of counterfactual conditionals. There it will be clear that our account offers the kind of ability to do otherwise that many theories of free will are after, the one called "all-in ability" (Austin 1961), "wide ability" (Vihvelin 2013), "ability with an opportunity" (Franklin 2011), or "maximally specific ability" (Kittle 2015b).

Abilities, understood in a new dispositionalist way, are compatible with determinism. However, if the CA is sound, then physical determinism implies that no one could have ever done otherwise, and therefore, in a deterministic world no one has the ability to do otherwise. In sections 3–8 we will show how compatibilists can secure their position against the destructive effect of the CA without being vulnerable to standard objections against compatibilism. We will do so by combining conditional analysis with a suitable version of a

- 
- 1 This definition is based on a modified version of Lewis' definition in Lewis (1997, 157); the characterization of an " $x$ -complete cause" is found in Lewis (1997, 156). The main difference between our account of abilities and Lewis' account of dispositions is that Lewis is interested in dispositions of any kind of entities to respond to relevant stimuli. Neither the disposition nor the stimulus must have something to do with agency or the psychological circumstances of the act, which are essential for questions about free will. In our account, the condition  $C_i$  has to include the agent's first-order desires, and the intrinsic property  $B$  has to include the agent's personality frame, as will be shown in section 4.
- 2 Further problems of Lewis' (1997) account of dispositions can be fixed but cannot be discussed here. For example, in order to admit a probabilistic or gradual notion of disposition one could follow Vihvelin's proposal (2013, 187) and weaken *Ability* so that the condition following "then" must hold only in a suitable proportion of cases.

sourcehood account (drawing on Frankfurt's (1971) notion of second-order desires as well as Fischer and Ravizza's (1998) condition of reason-reactivity).

## 2 The Review of the Consequence Argument

The "third" version of the CA, published in van Inwagen's (1983) book, has attracted the most attention in the free will debate. It contains three propositions:

$P_0$ . A proposition that describes the total state of the world at some moment in the distant past ( $t_0$ ).

$L$ . A proposition that is the conjunction of all the laws of nature.

$P$ . A true proposition about time  $t_1$  after time  $t_0$ .

$N$ , a sentential modal operator defined:

$Np$ .  $p$ , and no one has, or ever had, any choice about whether  $p$ .

Two modal principles, or rules of inference:

RULE ALPHA. If  $p$  is a necessary truth, then  $p$  is true and no one has, or ever had, any choice about  $p$ . ( $\Box p \vdash Np$ )

RULE BETA. If  $p$  and no one has or had any choice about  $p$ , and if  $p \supset q$  and no one has or had any choice about  $p \supset q$ , then  $q$  and no one has or had any choice about  $q$ . ( $Np, N(p \supset q) \vdash Nq$ ).

Using these notations, the argument has the following logical structure:

1.	$\Box((P_0 \wedge L) \supset P)$	Symbolic definition of Determinism
2.	$NP_0$	Principle of the Fixity of the Past
3.	$NL$	Principle of the Fixity of the Laws
$\therefore$	$NP$	Conclusion, contradicts CDO

The proof:

---

4. $\Box(P_0 \supset (L \supset P))$	1, Exp within $\Box$ -scope
5. $N(P_0 \supset (L \supset P))$	4, Rule Alpha
6. $N(L \supset P)$	2, 5, Rule Beta
$\therefore NP$	3, 6, Rule Beta

---

Since  $P$  can be any true proposition about what someone does,  $NP$  asserts that no one has any choice about any of her actions. If a compatibilist wants to reject the CA, she has to reject either one of the inference principles or one of the premises.

Rule Alpha is very plausible and has been widely accepted in this discussion.

Rule Beta is known as the Principle of Transfer of Powerlessness and, according to the proponents of the CA, is also very plausible (cf. [Ginet 1980, 182](#); [van Inwagen 1983, 99](#)). However, Rule Beta is the most disputed part of the CA, usually criticized by means of counterexamples ([Widerker 1987](#); [McKay and Johnson 1996](#); [Carlson 2003](#)). McKay and Johnson argue that Alpha and Beta together entail the Principle of Agglomeration:  $Np, Nq \vdash N(p \wedge q)$ .<sup>3</sup> However, this principle can be shown as invalid by applying the condition “ $N(-)$ ”, that no one can do anything about, to the outcome of a random process, viz. the tossing of a fair coin. After criticizing van Inwagen’s formulation of Beta they propose four different modal principles closely resembling Beta, which are immune to this counterexample but still can be used in deriving the conclusion of the CA. It has been argued that these different principles are less intuitive than the original Beta and have unwelcome consequences ([Blum 2003](#)). Other versions of Beta have been proposed ([Carlson 2000, 2003](#); [Crisp and Warfield 2000](#)) and currently the discussion is very much alive ([Gustafsson 2017](#)).

Van Inwagen himself reacted to McKay and Johnson’s (1996) counterexample by conceding that his version of Beta was invalid, and by modifying the  $N$ -operator as follows:

---

3 McKay and Johnson give the proof (1996, 115):

1.  $Np$  (premise)
2.  $Nq$  (premise)
3.  $\Box[p \supset (q \supset (p \wedge q))]$  (necessity of a logical truth)
4.  $N[p \supset (q \supset (p \wedge q))]$  (from 3 and  $\alpha$ )
5.  $N[q \supset (p \wedge q)]$  (from 1, 4 and  $\beta$ )
6.  $N(p \wedge q)$  (from 2, 5 and  $\beta$ )

$p$  and every region to which anyone has, or ever had, *exact* access is a subregion of  $p$ . One has *exact* access to a region if one has access to it *and to none of its proper subregions*. Intuitively, one has exact access to  $p$  if one can ensure the truth of  $p$  but of nothing “more definite”. (van Inwagen 2000, 8)

So, according to this definition, McKay and Johnson’s (1996) case is not a counterexample to Rule Beta anymore. An agent can have exact access to the region of logical space in which “The coin is tossed and it lands either heads or tails” holds, but not to its subregions where specifically “The coin is tossed and it lands heads” holds, or where specifically “The coin is tossed and it lands tails” holds.

Lynn Baker has argued that the new  $N$ -operator leads to the conclusion that “every region of logical space to which anyone has, or ever had, exact access is the region containing only the actual world” (2008, 16). If this is correct, then the conclusion of the CA will follow quite independently from the assumption of determinism. However, we will not explore the implications of the new  $N$ -operator, because in the following sections we will show that compatibilists can deny  $NR_0$  and  $NL$  on both readings of  $N$ .

Moreover, we claim that supporters of agents’ abilities to do otherwise do not need to withdraw (suitable versions of) Rule Beta if a version of Beta that withstands objections can be formulated. Rather, they can and should reject one of the three premises:

- If determinism is accepted, then either 2. or 3. is to be rejected, as in classical compatibilist and new dispositionalist positions,
- If indeterminism is accepted, then 1. must be rejected, as in libertarian positions.

In the next sections, we show how these premises can be rejected by developing a new kind of conditional analysis of freedom using the formal tool of the analysis of counterfactual conditional statements in terms of possible worlds. We argue that our proposal has three advantages compared to traditional accounts:

1. It answers the standard objection against conditional explications of CDO that they are too weak to remove counterexamples, by combining them with a suitable version of a sourcehood account based on the

condition of coherence with one's personality frame, abbreviated as (CPF).

2. It is flexible enough to be compatible with two versions of determinism, (i) a backtracking variant and (ii) a local-miracle variant, as well as with (iii) indeterminism.
3. Moreover, the condition CPF may also solve specific problems of the three versions, e.g., the "anything possible" objection against (ii) and the problem of luck objections against (ii) and (iii).

Some compatibilists have provided arguments against the CA (cf. [Vihvelin 1988](#); [Taylor and Dennett 2002](#)) and developed their own compatibilist theories of free will ([Vihvelin 2013](#); [Dennett 2003](#)). However, an analysis of the precise connection between a refutation of the CA and a formulation of a conditional compatibilist analysis of CDO is missing so far. In this paper, we are going to address this issue. We will incorporate the results of the previous critics of the CA into one unified framework that is based on a counterfactual analysis and shows how both a backtracking and a local miracle analysis of counterfactuals can be used to refute the CA and to provide a positive account of CDO. Moreover, we will show that the same counterfactual framework can be used to explicate an indeterministic account of CDO that drops the assumption of determinism instead of employing backtracking or local miracles.

We chose the version of the CA that we did because it is arguably one of the strongest arguments against compatibilism (cf. [Capes 2019](#)). Our proposed refutation of this version of the CA also works against the "Basic Version" of the CA. The Basic Version depends on the acceptance of the "Extension Principle": "An agent can do  $X$  only if his doing  $X$  can be an extension of the actual past, holding the laws fixed" ([Fischer 1994, 88](#)). The "Extension Principle" is a straightforward affirmation of the Fixity of the Past and the Fixity of the Laws, and in sections 3–7 we show how both Fixity Principles can be consistently denied.

### 3 Compatibilist Rejection of the Fixity Principles

In the following sections 3–7 we assume determinism and propose our new explication of compatibilist conditional freedom within a deterministic framework. Premise 2., the Principle of the Fixity of the Past, states that  $B_0$ , a proposition that describes the total state of the world at some moment in the

distant past, is true and no one can make  $P_0$  false where “ $x$  makes  $P_0$  false” is understood in the following weak sense: if  $x$  had acted otherwise, then the distant past would have been different.<sup>4</sup> This formulation does not assume that there is a direct causal relationship between the agent’s actions and the change of the events in the remote past.

John Saunders (1968) was the first to reject the Fixity of the Past with this kind of strategy, which later became known as backtracking (Fischer 1988). In our viewpoint, the main advantage of the backtracking strategy in the context of the freedom debate is that it applies to the conditional analysis of “can”. According to the backtracking strategy, “ $P_0$  and no one has, or ever had, any choice about whether  $P_0$ ” is wrong, because the agent,  $x$ , can perform not- $P$  now, and if  $x$  performs not- $P$  now then it would have been false that  $P_0$ .<sup>5</sup> So  $x$  has the power to change the past  $P_0$  in the weak sense explained above. The connection between the backtracking strategy and the conditional analysis of “can” is that the causal chain leading from the counterfactual alteration of  $P_0$  to the counterfactual alteration of  $P$  involves a counterfactual alteration of the agent’s will (decision) at some intermediate time.

The backtracking strategy as applied to the abilities of agents is just a particular case of the backtracking analysis of the truth of counterfactual statements of the form ( $P > Q$ ). According to Jonathan Bennett, “( $P > Q$ ) is true iff  $Q$  is true at all the  $P$ -worlds which are closest to the actual world” (1984, 57), and since we want, even in deterministic worlds, some counterfactuals to be true and some false, we have two options to choose from:

[...] if  $P$  is false (at the actual world), then every causally possible  $P$ -world is unlike the actual world in respect of its whole history up to the time ( $T$ ) to which  $P$  pertains. Any good statement of the determinist thesis will tell you that much, making it clear that any two worlds which are strictly determined by the same laws are unlike at time  $T$  only if they are unlike at every earlier time. So, if we want to evaluate ( $P > Q$ ) where  $P$  is false, we must either

- 
- 4 On van Inwagen’s new formulation of  $N$ , this should be read as “ $P_0$  and every region to which anyone has, or ever had, exact access is a sub-region of  $P_0$ ”. The rest of our argument applies equally well to the old and the new formulation of the  $N$ -operator. The difference is that where we say “ $x$  has a choice about whether  $P_0$ ” on the old formulation of  $N$ , we substitute “ $x$  has exact access to a region where  $P_0$  is false” on the new formulation of  $N$ .
- 5 By “performing  $P$ ” we mean performing an act such that “ $P$ ” is a proposition describing this act, and by “performing not- $P$ ” we mean performing an act such that “not- $P$ ” is a proposition describing this act.



accept as “closest” some worlds which are unlike ours at all times earlier than  $T$ , or deem to be “closest” some worlds which are just like ours up to about  $T$  and are then pushed off our course by a miracle—an event breaking some actual causal law. (Bennett 1984, 59)

Bennett chooses the first option, Lewis (1979) chooses the second. The backtracking analysis rejects Premise 2., i.e., assumes a different past globally, i.e., in many instances. In contrast, Lewis’ local miracle strategy rejects Premise 3., i.e., requires a violation of a law but only locally, i.e., in only one instance.

In section 4 we will elaborate on how the backtracking analysis of counterfactuals together with the conditional analysis of abilities yields a compatibilist analysis of free will. In a nutshell, the idea is the following: we say that  $x$  could have done otherwise than  $P$  at  $t_n$  if there are possible worlds close to the actual world at  $t_{n-1}$  in some respect to be clarified in what follows, such that in these worlds  $x$  does otherwise than  $P$  at  $t_n$ .

The local miracle account is also a possible way to go for a compatibilist, viz. to reject Premise 3., the Principle of the Fixity of the Laws. It states that no one can change the laws of nature (has a choice about what the laws of nature are), where “ $x$  changes the laws of nature” is understood in the following weak sense: if  $x$  had acted otherwise then the laws of nature would have been different. This formulation is a slight reformulation of Lewis’ (1981) weak compatibilist thesis in a way that avoids van Inwagen’s (2004) criticism.

David Lewis (1981) distinguished between two senses in which a law of nature can be broken in connection with what an agent does. In a strong sense, it can be broken by an action that an agent performs or by a direct consequence of an action that an agent performs. For example, a law is broken in a strong sense if an agent moves his hand faster than the speed of light or throws a stone that flies faster than the speed of light. Crediting an agent with this kind of ability is implausible, so, read in the strong sense, Premise 3.,  $NL$ , is true. However, a law of nature can be broken in a weak sense: it is possible that somewhere in the past a local miracle happened. In this case, it is possible that an agent does otherwise than  $P$  as a consequence of this prior miracle, which could have happened at any time between  $P_0$  and  $P$ .

Lewis made his objection against an earlier version of the CA (van Inwagen 1975) in the following way:

(Weak Thesis) I am able to do something such that, if I did it, a law would be broken.

(Strong Thesis) I am able to break a law. (Lewis 1981, 115)

The Weak Thesis, which as a soft determinist I accept, is the thesis that I could have rendered a law false in the weak sense. The Strong Thesis, which I reject, is the thesis that I could have rendered a law false in the strong sense. (Lewis 1981, 120)

According to Lewis, it is the strong sense that is incredible, and it is the weak sense that follows from the CA, so the CA is not a problem for a compatibilist (soft determinist) position. Van Inwagen objected that even the Weak Thesis is incredible because it ascribes to an agent the power to perform miracles, where a miracle is defined as “an event or state of affairs whose occurrence would be inconsistent with the whole truth about the past and the laws of nature” (van Inwagen 2004, 349). But this incredibility is arguably due to the inappropriateness of the phrase “a law of nature is broken”. We think that what Lewis means with this is nothing more than what we said above, namely, that the laws of nature are different in the actual world (where I perform  $P$ ) and the counterfactual world where I act differently. Tognazzini (2016) argues in more detail why this is what Lewis must have meant with a law being “broken”, and what “miracles” are according to Lewis. Helen Beebe and Alfred Mele (2002) argue that Humeanism about laws of nature supports not only the Weak, but also the Strong Thesis, and this is a problem for Lewis’ local miracle compatibilism. However, in our local miracle compatibilist proposal, presented in section 7, we do not endorse Humeanism about laws of nature. Therefore we are free from the problems discussed in Beebe and Mele (2002). In section 7 we also show how our account solves the problems discussed in Beebe (2003).

Both the backtracking strategy and the local miracle strategy are legitimate ways for a compatibilist to reject the CA. Indeed, the compatibilist accepts *ex hypothesi* that  $(P_0 \wedge L) \supset P$  and that  $P_0$ ,  $L$ , and  $P$  are true in the actual world. The decisive question is: what is involved in making the conditions  $C_i$  true in a possible world sufficiently close to the actual world? The compatibilist who accepts determinism can give two answers according to our analysis, one based on the backtracking strategy and the other one on the local miracle strategy. The indeterminist, in contrast, can explicate the conditions  $C_i$  in a way that neither implies a global change of the past nor a local miracle. In the next five sections, we elaborate on these three options.

## 4 Backtracking Compatibilist Proposal

If the actual world  $a$  is deterministic, the backtracking strategy allows us to formulate the following conditions that have to obtain for CDO to be true about a person  $x$  in  $a$ .

$CDO_B$ .  $x$  could have done otherwise than  $P$  at  $t_n$  if  $x$  does not- $P$  at  $t_n$  in some possible world  $w$  that satisfies the following conditions:

- a.  $w$  and  $a$  are governed by deterministic laws that are identical.
- b. The pasts of  $w$  and  $a$  are different at all past times.
- c.  $x$ 's personality frame agrees in  $a$  and  $w$  at all times until  $t_{n-1}$  and it does not change between  $t_{n-1}$  and  $t_n$ .
- d. (1)  $x$ 's internal state at  $t_{n-1}$  in  $w$  differs from the corresponding internal state of  $x$  in  $a$  in regard to some FODs of  $x$ , in coherence with  $x$ 's personality frame, where
  - (2)  $w$  and  $a$  agree in all agent-external facts at  $t_{n-1}$  that were causally relevant to  $x$ 's actual action at  $t_n$ .

Characteristics required by  $CDO_B$  (a) and (b) were discussed in previous sections. Characteristics  $CDO_B$  (c) and (d) need explanation.

Concerning  $CDO_B$  (c): According to our account, in order to have free will an agent must have a personality frame (F), which, in turn, includes reasoning and volitional abilities meeting minimal rationality conditions. Minimal rationality conditions demand that an agent meet the criteria of moderate reasons-responsiveness, i.e., have a certain level of reasons-receptivity and reasons-reactivity, as discussed in Fischer and Ravizza (1998). Necessary volitional abilities include an agent's capacity to form first- and second-order desires and volitions. Following Harry Frankfurt, we define a first-order desire (FOD) as a desire "to do or not to do one thing or another" (1971, 7), a second-order desire (SOD) as a desire that a certain FOD become causally efficient (1971, 10), a first-order volition as an effective FOD that causally contributes to an agent's act (1971, 8), and a second-order volition as a SOD that is a part of the cause of the agent's first-order volition (1971, 10). A person having abilities at a time should be understood as explained in ABILITY. A person  $x$ 's having the ability to do  $A$  at  $t_n$  is a necessary condition for the truth of the claim that " $x$  could have done  $A$  at  $t_n$ ". For example, if we want to know whether the claim " $x$  could have played the piano at  $t_n$  instead of playing the violin" is true, we should consider the possible worlds where  $x$  has the same

intrinsic property  $B$  (causal basis of  $x$ 's ability to play the piano) and where conditions  $C_i$  vary without violating  $\text{CDO}_B$  (d).

In addition to the necessary abilities listed above, common to all persons with free will, each  $F$  has a particular set of characteristics, including this person's SODs, essential FODs such as the desire to live, stable character traits, and general and specific abilities and skills, such as an ability to play the violin, an ability to play the violin in front of a big audience, an ability to play the violin in front of a big audience while being tired, and all the rest of this person's abilities in all the ranges of specificity.<sup>6</sup>

These characteristics of a personality frame stay fixed across periods of time under consideration in  $\text{CDO}_B$  (c). The qualification that  $x$ 's personality frame "does not change between  $t_{n-1}$  and  $t_n$ " is needed because although we allow for changes of personality in a more distant past we need to exclude that the agent undergoes changes of her personality between times  $t_{n-1}$  and  $t_n$ . To find out if  $\text{CDO}_B$  is true about  $x$ , we only consider the possible worlds where  $x$ 's personality frame is the same at  $t_{n-1}$  as  $x$ 's personality frame in  $a$  at  $t_{n-1}$ . We don't want to say that " $x$  could have done otherwise" is true about  $x$  if, in other possible worlds where  $x$  does otherwise, she has different skills or significantly different character, values, and beliefs, in particular, where  $x$  has different SODs. As we shall see, this removes some important problems.

Concerning  $\text{CDO}_B$  (d): Coherence of the FODs with the personality frame is the key step in the conditional analysis of CDO. We claim that  $x$  could have performed not- $P$  out of her own free will if there is a possible world where  $x$  has different FODs that are coherent with her personality frame at  $t_{n-1}$  and  $x$  performs not- $P$  at  $t_n$ , whereas all agent-external facts at  $t_{n-1}$  that were causally relevant to  $x$ 's actual action at  $t_n$  are the same as in  $a$ . Coherence with one's personality frame (CPF) is characterized as follows:

CPF. An action  $A$  of agent  $x$  is *coherent with  $x$ 's personality frame  $F$*  iff performing the action  $A$  does not imply consequences that  $x$  can draw (using her instrumental reasoning abilities which are part of  $F$ ) that contradict certain elements of  $F$ .

Condition  $\text{CDO}_B$  (d1) is a formal explication of what kind of changes internal to agent  $x$  are allowed in the possible worlds under consideration: namely, the

6 Since our account fixes all the abilities of a person in all ranges of specificity, it is consistent with both Whittle's (2010) and Kittle's (2015b) competing claims regarding what level of specificity of abilities is most relevant to free will.

changes in  $x$ 's FODs that don't contradict CPF. Taken together with the fixity of F explicated in  $CDO_B$  (c), this restricts the counterfactual FODs to such FODs that are neither essential to F nor lead to actions that imply consequences that  $x$  can draw (using her instrumental reasoning abilities which are part of F) that contradict certain elements of F. Differences in world  $w$  that are mentioned in condition  $CDO_B$  (d1) reach all the way back to the Big Bang. These past differences are causally relevant to counterfactual FODs:  $x$  has a different FOD in  $w$  because  $a$  and  $w$  have different past histories that cause differences in the present states of these worlds, including the differences in  $x$ 's FODs.

Condition  $CDO_B$  (d2) specifies that these differences must not affect those agent-external facts in the actual world at time  $t_{n-1}$  that were causally relevant to  $x$ 's action at  $t_n$ . Otherwise, the counterfactual analysis would be trivialized and obviously unfree actions such as ones resulting from being forced by physical violence would come out as "free". However, these differences may affect those agent-external facts in the actual world at time  $t_{n-1}$  that were not causally relevant to  $x$ 's action at  $t_n$  and may have further causal consequences in  $w$  at times later than  $t_n$ .

## 5 Discussion of the Backtracking Compatibilist Proposal

In this section, we will show how our account meets the intuitive desiderata by analyzing some pertinent examples. In the next section, we will discuss in detail the differences between our account and some other conditional accounts, and demonstrate that our account solves the problems that those accounts face.

First of all, we note that our account agrees with the classical conditional analysis on those examples that the latter gets right: those in which an external force prevents an agent from doing otherwise. For example, it follows from our account that if an agent is physically chained, he cannot move his arms even if he wanted to move them, because the causally relevant external conditions stay fixed in the counterfactual analysis of  $CDO_B$ .

Second, our account explains an important pre-theoretical intuition, according to which not everything that a person can physically do she can do in a free will sense of "can". Consider Jones: when a robber points a gun to his head and demands that Jones hand over his wallet, we do not want to say that Jones is free to do otherwise than obey the robber, even if there is a possible world where Jones refuses. If Jones, like most of us, values his life more than

his wallet, then his personality frame contains the essential FOD to preserve his life. According to our account, there is no possible world where Jones' personality frame  $F$  does not change and he refuses to hand over the wallet because we excluded all other possible differences, such as Jones' mishearing the threat, or having some form of hallucination, or being manipulated by neuroscientists, by the condition that everything except  $x$ 's FODs that are not part of  $F$  in these possible worlds at  $t_{n-1}$  is the same as it is in  $a$ .

Third, our account answers an important objection to a standard conditional analysis of abilities raised by van Inwagen (1983). Consider Smith, who is in a coma in a hospital. Van Inwagen observes that:

The two propositions

Smith cannot get out of bed

If Smith wanted to get out of bed, he would

would seem both to be true, the former because he is in a coma, and the latter because, if he *did* want to get out of bed he wouldn't be in a coma. (van Inwagen 1983, 119)

This objection is a problem for many other versions of classical compatibilism, but we think that there is a straightforward way to avoid this problem on our account, because being in a state of coma violates condition  $CDO_B$  (c), according to which  $x$ 's personality frame is fixed in the possible worlds under consideration. But a person in a state of coma doesn't have a personality frame in the sense in which we understand this notion, because, while being comatose, the person is not moderately reasons-responsive, and at least temporarily lacks a capacity to form first- and second-order desires and volitions.

Fourth, our account solves the notorious red candy problem, dating back to Lehrer's (1968) example. The example is as follows:

Suppose that I am offered a bowl of candy and in the bowl are small round red sugar balls. I do not choose to take one of the red sugar balls because I have a pathological aversion to such candy. [...] It is logically consistent to suppose that if I had chosen to take the red sugar ball, I would have taken one, but, not so choosing, I am utterly unable to touch one. (Lehrer 1968, 32)

The conclusion from Lehrer's thought experiment seems to be that a conditional analysis of abilities is bound to fail because in this case, it will give the implausible result that I can take the red candy if I decide/choose/want to, whereas intuitively I cannot choose a red candy because of my phobia. Our account, however, gives the conclusion that the person with a phobia cannot take the candy precisely because of his pathological aversion which is a part of his personality frame, so it has to be fixed in all of the possible worlds that we consider. It could be objected that we should also consider the possible worlds where something distracts the person with the phobia so that he forgets about his phobia at the critical moment. However, such a counterfactual distraction would be caused by a change in the agent-external facts at time  $t_{n-1}$  that were causally relevant to  $x$ 's action at  $t_n$ , and that is excluded by condition (d2) of  $CDO_B$ . It could also be objected that some competing FOD can ultimately outweigh the aversion, so that a person takes the red candy after all, even if the phobia is included in  $F$ . However, the point of mentioning phobias in these kinds of counterexamples is precisely because they entail the inability of agents to form certain kinds of desires. Phobias distinguish what individuals cannot do from what they can do. Our account secures this intuition by including phobias and other kinds of irresistible psychological impulses into  $F$ . Consequently, any counterfactual FOD incompatible with the phobia is excluded by  $CPF$ . What is an irresistible psychological impulse and what is not (like weakness of will) is an empirical question, and the answer to this question determines whether some desires and fears should be fixed as elements of  $F$ .

Fifth, our account captures the pre-theoretical intuition about the cases in which agents freely perform some actions that are not very consequential for them. For example, we have a strong pre-theoretical intuition that we could have put on different clothes in the morning or ordered a different meal at the restaurant. The extreme case of choosing among inconsequential options is the so-called freedom of indifference: when an agent has to choose between two (or more) options and has no reason whatsoever to prefer one option over the other. This situation seemed problematic to those philosophers who thought that every choice must happen for a reason, i.e., be caused by a prior decision of the intellect (cf. [Kenny 1973](#)), but it is not problematic on our account, because any variation in the internal life of the agent, including the slightest unconscious biases or simply differences in neuronal activity will be enough for the agent to act otherwise.

Sixth, our account explains why we do not say that all animals have free will. According to a Humean style simple conditional analysis which says that *x* acted freely if he would have acted differently given different desires, all animals that have desires would also be free in their actions. It would follow, e.g., that a mosquito is freely stinging, because if it were not hungry but, rather, sleepy, it would do otherwise. However, a mosquito has no personality frame and therefore its actions are not free according to our analysis. For the same reason, primitive robots are not free according to our explication.

Finally, our account explains why human beings who have a rudimentary form of free will but have not yet developed a personality frame, such as young children, or who have a defective personality frame (for example, due to severe psychological disorders), do not qualify as free agents.

It may be objected that a satisfactory explication of freedom should also apply to situations in which a person changes her personality frame, but it is hard to say what freedom means in this case. Typically, an action that is involved in such a change violates some elements of the person's old frame but is in line with the person's new but not yet fully developed personality frame. So, what counts for the evaluation of an action as free or unfree in such a situation, the old (past) or the new (future) personality frame? If a person is manipulated by another person in a way that changes her personality, but after the change she considers herself free and her action is compatible with the new frame, then in which sense was this change unfree or free? We do not intend to develop a solution to this difficult but distinct problem in this paper; we postpone it to future work.

A final remark: Normally an agent's personality frame is not so strong as to determine her actions or first-order desires. In the exceptional case, however, in which someone does something as an immediate consequence of her personality frame, for example, regularly breathes, eats, and drinks (because the personality frame includes her desire not to die), then our present analysis implies that the person *is indeed not free* in regard to these actions. This sounds reasonable in the case of our example, but there are other cases where it does not seem so reasonable. Dennett (1984) draws our attention to cases where an agent's deeply held convictions make any alternative course of action inconceivable to the agent. According to Dennett, when Luther claimed "Here I stand; I cannot do otherwise" he might have been telling the truth, while still being free and responsible for the choice that he made. Regardless of whether this diagnosis really does apply to Luther on this occasion, it does seem plausible that sometimes there is only one way a person can act. There



are many deeply held convictions that make some courses of action inconceivable for certain agents. It might be this intuition of fixity of everything that is subjectively important for a person that brings some compatibilists to deny that having alternative possibilities is at all relevant to having free will. If one intends an analysis that does not make CDO a part of the explication in cases of free action to require a change of the personality frame, one has to change our defining condition by adding the following disjunct:

CDO<sub>B</sub>\*. “[...] or the action *P* follows already from the content of *x*’s personality frame.”

In this case, the modified definition of free action would be: *x* acted freely if either CDO<sub>B</sub> or CDO<sub>B</sub>\* obtain. Whether CDO<sub>B</sub> or CDO<sub>B</sub>\* (or something in between) is the better analysis of free action in a deterministic world is left here as an open question to be treated in future work.

## 6 The Advantages of the Backtracking Compatibilist Proposal

In the current free will debate one sometimes sees the contrast being drawn between the conditional analysis of abilities and the counterfactual possible world analysis, as if these two ways of analysis were mutually exclusive (Kittle 2015b, 101). We think that this understanding is mistaken since counterfactual possible world analysis is a way to provide a conditional analysis, as our paper illustrates. The contrast itself dates back to Lehrer (1976), who rejected conditional analysis and proposed his possible world analysis instead. However, it is important to note that what Lehrer rejected were the then available versions of a simple conditional analysis, which failed due to objections similar to those we considered in the previous section, including Lehrer’s own red candy counterexample (1968), but not the very idea of finding a suitable conditional definition of free will. We think that our backtracking compatibilist proposal is a step towards such a definition, and we will now highlight how it differs from some influential versions of conditional analyses proposed by other authors.

Lehrer’s (1976, 1990) possible worlds analysis states that a person is able to do otherwise if there is an accessible minimally different possible world where he does otherwise, and there is “no advantage” he has in that possible world as compared to the actual world. Lehrer’s account of “advantage” was

persuasively criticized by Horgan (1977) and Kittle (2015b). These criticisms do not apply to our account, because  $CDO_B$  does not use the notion of advantage, but instead specifies in detail what is and what is not allowed to be different in the possible worlds under consideration. For this reason, our account is free from the difficulties that face the possible world analysis by John Campbell (1997), who develops Lehrer's notion of advantage.

Our account is also free from the problems that the new dispositionalist analyses of abilities face. Randolph Clarke (2009) has argued that the new dispositionalist accounts are vulnerable to objections similar to the red candy case, where an agent is unable to *A* because he is unable to try to *A*. He provides the following example:

Suppose that on a certain occasion Bob formed an intention to wave to Cathy, but a momentary neural glitch made it impossible for Bob, on that occasion, to try to wave – he could not even begin to implement his intention – though he would have waved if he had managed to try. (Clarke 2009, 335–336)

Clarke argues that the new dispositionalism gives the implausible result that Bob was able to wave because he had a disposition to wave, which would have manifested itself if he had tried. We, however, answer that Bob was not able to wave, because by the condition below we fix all abilities of the agent at  $t_n$ :

$CDO_B$  (c). *x*'s personality frame agrees in *a* and *w* at all times until  $t_{n-1}$  and it does not change between  $t_{n-1}$  and  $t_n$ .

Due to the glitch, Bob lacks the ability to wave at  $t_n$ , because he temporarily lacks the proper causal basis *B* of this ability, namely, the normal functioning of his neural pathways. So he could not have waved at  $t_n$ .

Franklin (2011) argues that both Vihvelin (2004) and Fara (2008) succeed in providing dispositional accounts of narrow, or general abilities, but not of wide, specific abilities, or, as Franklin calls them, abilities with opportunities. This leads to an implausible claim that even externally constrained agents possess abilities to do otherwise:

According to Vihvelin's analysis, free will is just a set of abilities, abilities are just (bundles of) dispositions, and dispositions are solely grounded in an agent's intrinsic properties. These claims prevent her from being able to appeal to the extrinsic features of

an agent (such as being tied to a chair) in order to explain why the agent is not free. (Franklin 2011, 97)

Our account does not have this problem, because the condition below excludes such changes in the agent's environment that prevent the agent from exercising her abilities:

CDO<sub>B</sub> (d2).  $w$  and  $a$  agree in all agent-external facts at  $t_{n-1}$  that were causally relevant to  $x$ 's actual action at  $t_n$ .

Vihvelin (2013) proposes a modified account of narrow abilities that attempts to solve the problems raised by Clarke (2009) and Franklin (2011) by introducing a proportion of success cases:

(LCA-PROP-Ability)  $S$  has the narrow ability at time  $t$  to do  $R$  in response to the stimulus of  $S$ 's trying to do  $R$  iff, for some intrinsic property  $B$  that  $S$  has at  $t$ , and for some time  $t'$  after  $t$ , if  $S$  were in a test-case at  $t$  and  $S$  tried to do  $R$  and  $S$  retained property  $B$  until time  $t'$ , then in a suitable proportion of these cases,  $S$ 's trying to do  $R$  and  $S$ 's having of  $B$  would be an  $S$ -complete cause of  $S$ 's doing  $R$ . (Vihvelin 2013, 187)

Kittle (2015a) argues that this modified account fails because it attributes to an agent abilities not relevant to free will. According to Kittle, Vihvelin's account has the following result: "When stood on the road miles from any water, I am such that *if I were in a test-case for my swimming abilities* and I tried to swim, then I would swim" (Kittle 2015a, 3031), but it would be wrong to conclude that I was free to swim there and then.

Our account does not face this problem, because it specifies precisely which situations are the test-cases: those that fit the conditions of CDO<sub>B</sub>.

## 7 Local Miracle Compatibilist Proposal

If the actual world  $a$  is deterministic, the local miracle strategy allows us to formulate the following conditions that have to obtain for CDO to be true about a person  $x$  in the actual world.

CDO<sub>M</sub>.  $x$  could have done otherwise than  $P$  at  $t_n$  if  $x$  does not- $P$  at  $t_n$  in some possible world  $w$  that satisfies the following conditions:

- a.  $w$  is governed by deterministic laws that are identical to the laws of  $a$  except for the one local miracle mentioned in (b).
- b. The pasts of  $w$  and  $a$  are identical until some past time  $t_m$  ( $m < n$ ) at which a local miracle happens.
- c.  $x$ 's personality frame agrees in  $a$  and  $w$  at all times until  $t_{n-1}$  and it does not change between  $t_{n-1}$  and  $t_n$ .
- d. (1)  $x$ 's internal state at  $t_{n-1}$  in  $w$  differs from the corresponding internal state of  $x$  in  $a$  in regard to some FODs of  $x$ , in coherence with  $x$ 's personality frame, where
  - (2)  $w$  and  $a$  agree in all agent-external facts at  $t_{n-1}$  that were causally relevant to  $x$ 's actual action at  $t_n$ .

Explications  $CDO_B$  and  $CDO_M$  differ in conditions (a) and (b), but are the same in (c) and (d).

Conditions  $CDO_M$  (c) and (d) provide forward-looking restrictions on what kind of miracles are allowed that are analogous to the backward-looking restrictions of the backtracking analysis and are needed for the same reasons. We have to exclude miracles that affect the agent-external facts in the actual world at time  $t_{n-1}$  (condition (d2)), since otherwise the counterfactual analysis would be trivialized. Moreover, also within the miracle account, we need to avoid an implausible conclusion that  $x$  could have done something that contradicts her personality frame. Indeed, imagine that Ann is sitting beside an open window in a high building and thinking about the fine day that awaits her. Can she freely jump out of the window for no particular reason? Of course, there is a possible world where she does precisely that due to a prior local miracle. But we would call such a situation a fluke, a random and unhappy incident, and not a free action of Ann's. The conclusion is that not only should the miracles leave  $x$ 's actual personality frame intact, but also they should not bring about any consequences that are inconsistent with  $x$ 's personality frame. This requirement is captured by condition  $CDO_M$  (d).

Condition  $CDO_M$  (d) also captures rationality requirements. Consider Jane: she is offered an apple and a pear and takes the apple. What has to be the case for the sentence "Jane could have taken the pear" to come out true? Presumably, a local-miracle compatibilist would not want to say that Jane could have taken the pear if there is a possible world where Jane decides to take the apple, but takes the pear instead, because of a prior miracle. This analysis would show that Jane could have done otherwise only if she had been irrational. It would not help much if a local-miracle compatibilist says

that Jane could have taken the pear if there is a possible world where Jane wants to take the apple, but decides to take the pear instead, because of a prior miracle. This would also be irrational. Condition  $CDO_M$  (d) provides us with the analysis that states that Jane could have taken the pear if there is a possible world where she forms a desire to take the pear, and takes it.

As the foregoing discussion shows, the conditions specifying which miracles are acceptable for  $CDO_M$  to be true about  $x$  are very similar to the conditions specifying which differences in the past states of the world are acceptable for  $CDO_B$  to be true about  $x$ .

Finally, our  $CDO_M$  analysis of abilities solves the problem for the local miracle compatibilism raised in Beebe (2003). Beebe argues that given the interpretation of abilities that can be reconstructed from Lewis (1981), there is no justification for the claim that the Weak Thesis is true, whereas the Strong Thesis is false because the possible world closest to the actual world where  $x$  does otherwise might be the possible world where  $x$ 's act itself is a divergence miracle. Our local miracle compatibilist proposal does not have this problem, because condition  $CDO_M$  (b) specifies that divergence of  $w$  and  $a$  happens at some past time  $t_m$  earlier than  $t_n$ , so the local miracle cannot be an action of  $x$  at  $t_n$ , whereas conditions  $CDO_M$  (c) and  $CDO_M$  (d2) ensure that this divergence could not have been an action of  $x$  at some earlier time.

## 8 Indeterministic Compatibilist Proposal

In the previous sections we have shown how a philosopher can have a theory of free will compatible with physical determinism. Thereby we have defended free will against the objection based on the putative incompatibility asserted by the CA. However, there are strong (though not decisive) arguments, based on contemporary quantum physics, that physical determinism is probably false. Prima facie, indeterminism seems to be a much easier way to refute the CA and establish freedom in the sense of CDO, simply by denying premise 1. of the CA. However, free will sceptics argue that physical indeterminism poses other threats to free will, namely, the problem of irrationality and the problem of luck.

These problems arise for those libertarians who accept both the Fixity of the Past and the Fixity of the Laws, and deny Determinism. The problem of irrationality can be expressed as follows: in order to have free will at  $t_n$ ,  $x$  must be able to do otherwise than  $P$  at  $t_n$ .  $x$  is able to do otherwise at  $t_n$  if there are worlds with the same laws of nature and the same past up to  $t_n$  where  $x$  does

otherwise at  $t_n$ . However, the same past up to  $t_n$  contains all of  $x$ 's mental states and dispositions, including all of her first- and second-order desires, beliefs, deliberations and intentions (for short: deliberations) up to  $t_n$ . But if  $x$  is in fact justified in doing  $P$  at  $t_n$  as a result of her prior deliberations, then in other possible worlds she is acting irrationally when she performs not- $P$  at  $t_n$ , as in our example with Jane who forms the desire to take the apple, decides to take the apple, but still takes the pear.

The standard libertarian response to this kind of worry consists in placing indeterminism not between the decision and the act, but between the desire and the decision, at some moment of deliberation when different motives and desires are being considered by the agent (Kane 1999, 2011b; Mele 2006; Ekstrom 2003).

This answer is good against the problem of irrationality. But it is not good enough to solve the problem of luck. Libertarians insist that a radically free act is never entirely determined by the past and the laws. No matter how firmly an agent decides to do something, how good her reasons are, and how strongly she wants it, she is free to do otherwise. This libertarian intuition has a troubling consequence when formulated in terms of possible worlds. Imagine that Mary, a libertarian agent, is considering whether she should cheat. She weighs pros and cons, thinks carefully, decides not to cheat, and acts in accordance with her decision. But she could have done otherwise, given precisely the same past up to  $t_n$ . Since everything about Mary is fixed right up to  $t_n$  after which either the situation  $w_1$  where she doesn't cheat or  $w_2$  where she cheats becomes actualized, it seems that if indeterminism obtains, then it is simply a matter of luck whether Mary cheats or not. And if some outcome is a matter of luck, it seems natural to say that the agent lacks control over this outcome, and therefore lacks free will in performing it. Some libertarian philosophers have devoted considerable efforts to address this problem (Kane 1999; Mele 2006). We claim that our indeterministic compatibilist account provides a solution to it, based on the conditional analysis of "can".<sup>7</sup>

If the world is indeterministic, the following conditions have to obtain for CDO to be true about an agent in this world:

---

<sup>7</sup> We call this account "indeterministic compatibilist", and not libertarian, because, while it says that indeterminism is compatible with free will, it doesn't say that it is necessary for free will, whereas libertarian accounts do so. All three accounts we propose in this paper are versions of what Vihvelin calls "commonsense compatibilism", the position which maintains that "we actually have free will and that this is so regardless of the truth or falsity of determinism" (2013, 34).

$CDO_1$ .  $x$  could have done otherwise than  $P$  at  $t_n$  if  $x$  does not- $P$  at  $t_n$  in some possible world  $w$  that satisfies the following conditions:

- a.  $w$  and  $a$  are governed by indeterministic laws that are identical.
- b. The pasts of  $w$  and  $a$  are identical until some past time  $t_i$  during  $x$ 's life span ( $i < n$ ) at which  $x$  has spontaneously generated some counterfactual FOD in world  $w$ .
- c.  $x$ 's personality frame agrees in  $a$  and  $w$  at all times until  $t_{n-1}$  and it does not change between the time  $t_i$  and  $t_n$ .
- d.
  1.  $x$ 's internal state at  $t_{n-1}$  in  $w$  differs from the corresponding internal state of  $x$  in  $a$  in regard to some FODs of  $x$ , in coherence with  $x$ 's personality frame, where
  2.  $w$  and  $a$  agree in all agent-external facts at  $t_{n-1}$  that were causally relevant to  $x$ 's actual action at  $t_n$ .

Explication  $CDO_1$  differs from  $CDO_B$  and  $CDO_M$  in conditions (a), (b), and (c), but is the same in (d).  $CDO_1$  (b) resembles  $CDO_M$  (b) because they both hold the past fixed until a divergence happens. There are two important differences between them: first,  $CDO_1$  (b) allows divergence of worlds' paths without miracles. Second, time  $t_i$  mentioned in  $CDO_1$  (b) is restricted to  $x$ 's lifespan:  $x$  could not have generated a FOD before he came into existence, whereas time  $t_m$  mentioned in  $CDO_M$  (b) could be a time point before  $x$  is born. However, time  $t_i$  is not restricted to a short period between  $x$  forming a desire and  $x$  making a decision, as some libertarians argue in their solutions of the luck problem (Kane 1999, 2011b; Mele 2006; Ekstrom 2003). While  $CDO_1$  (b) does allow  $x$ 's counterfactual FODs to be generated precisely in that time period (between  $x$ 's desire and  $x$ 's decision to act), it also allows for  $x$ 's counterfactual FODs to be generated earlier. This provides a weak indeterministic position on an agent's free will, which does not require that in order for an agent to act freely an agent's choice must not be determined right up to the moment of the agent's making a decision. An agent will also be free even if he spontaneously generates a FOD sometime in the past, makes a plan in accordance with the FOD, and sticks to the plan. Thus,  $CDO_1$  seems to be a formal analysis capable of incorporating the intuition that sometimes we are really determined to do what we are doing because of the FODs we had some time ago, but we are nevertheless free because these FODs could have been different. However, what has to be required is that  $x$ 's personality frame does not change in both worlds between the time  $t_i$  at which  $x$  spontaneously

formed the FOD causally relevant for his counterfactual FOD in world  $w$  at time  $t_{n-1}$  and time  $t_n$ .  $CDO_I$  provides a solution to the problem of luck similar to that of  $CDO_M$  without presupposing determinism. According to  $CDO_I$ , how  $x$  acts is not entirely determined by  $x$ 's past and laws of nature. But it is not a matter of luck, because not every nomologically possible action could have happened with a corresponding probability, e.g., that the agent, instead of visiting his mother, could have killed his mother or ignored her for the next few months. Exactly that is afforded by our condition  $CDO_I$  (d) since it excludes all actions incoherent with  $x$ 's personality frame. In other words, only actions coherent with  $x$ 's personality frame are allowed. Therefore, it is no longer a matter of luck how  $x$  acts, although it is not determined either, because we are assuming indeterminism.

In conclusion, if physical indeterminism obtains, and spontaneous will-forming processes do indeed occur in our brains, then we claim that  $CDO_I$  is the correct analysis of alternative possibilities necessary for free will. On the other hand, if physical determinism obtains and spontaneous will-forming processes do not occur in our brains, then  $CDO_B$  or  $CDO_M$  can do the job. Either way, there is no reason to think that we need to know the truth about fundamental laws of physics before we can assert that some agents could have done otherwise.

## 9 Conclusion

We have provided a new account of free will, based on a conditional analysis of agents' abilities to do otherwise combined with sourcehood components. It allows alternative possibilities whether determinism or indeterminism obtains, and makes use of Frankfurt's psychological approach. Our proposal has three advantages:


1. It answers the objections against other versions of conditional analysis of "can" by demanding coherence of what one can freely do with one's personality frame,  $CPF$ . This allows us to analyze situations of coerced or irrational actions in an intuitively plausible way.
2. It is compatible with three metaphysical background assumptions:
  - (i) determinism with backtracking
  - (ii) determinism with local miracles and
  - (iii) indeterminism.



3. It is immune to the consequence argument and also solves the luck problem.


Our account meets the intuitions behind the classical compatibilist approach, the sourcehood compatibilist approach, and the leeway libertarian approach. It is also not vulnerable to either the CA, which, according to a received opinion in the contemporary free will debate, is one of the most pressing worries for the compatibilists, or to the luck problem, which, according to another received opinion in the contemporary free will debate, is one of the most pressing worries for the libertarians. Therefore, it has the merits of both of these positions without having their drawbacks. Finally, our account of free will is naturalistic, because it is compatible with any answer that the fundamental physical theory can give to the question of determinism. Free will is real, and some agents have it, whether our world is fundamentally deterministic or not.\*

Maria Sekatskaya

 0000-0002-5381-2913

University of Düsseldorf  
maria.sekatskaya@hhu.de

Gerhard Schurz

 0000-0002-4107-9240

University of Düsseldorf  
gerhard.schurz@phil.hhu.de

## References

- AUSTIN, John Langshaw. 1961. *Philosophical Papers*. Oxford: Oxford University Press.  
Edited by J.O. Urmson and G.J. Warnock.
- AYER, Alfred Jules. 1954. *Philosophical Essays*. London: MacMillan Publishing Co.
- BAKER, Lynne Rudder. 2008. "The Irrelevance of the Consequence Argument."  
*Analysis* 68(1): 13–22, doi:10.1093/analys/68.1.13.
- BEEBEE, Helen. 2003. "Local Miracle Compatibilism." *Noûs* 37(2): 258–277,  
doi:10.1111/1468-0068.00438.

---

\* This work was supported by Deutsche Forschungsgemeinschaft, Research unit "Inductive Meta-physics" (FOR 2495), research grant SCHU 1566/11-2. We would like to thank Hans J. Briegel, Christian J. Feldbacher-Escamilla, Vera Hoffmann-Kolss, Andreas Hüttemann, Kristin M. Mickelson, Thomas Müller, Corina Strössner, Kadri Vihvelin, and Verena Wagner for helpful comments on an earlier version of this paper.

- BEEBEE, Helen and MELE, Alfred R. 2002. "Humean Compatibilism." *Mind* 111(442): 201–224, doi:10.1093/mind/111.442.201.
- BENNETT, Jonathan. 1984. "Counterfactuals and Temporal Direction." *The Philosophical Review* 93(1): 57–91, doi:10.2307/2184413.
- BLUM, Alex. 2003. "The Core of the Consequence Argument." *Dialectica* 57(4): 423–430, doi:10.1111/j.1746-8361.2003.tb00281.x.
- CAMPBELL, Joseph Keim. 1997. "A Compatibilist Theory of Alternative Possibilities." *Philosophical Studies* 88(3): 319–330, doi:10.1023/a:1004280421383.
- CAPEL, Justin A. 2019. "What the Consequence Argument Is an Argument for." *Thought* 8(1): 50–56, doi:10.1002/tht3.404.
- CARLSON, Erik. 2000. "Incompatibilism and the Transfer of Power Necessity." *Noûs* 34(2): 277–290, doi:10.1111/0029-4624.00211.
- . 2003. "Counterexamples to Principle Beta: A Response to Crisp and Warfield (2000)." *Philosophy and Phenomenological Research* 66(3): 730–737, doi:10.1111/j.1933-1592.2003.tb00287.x.
- CLARKE, Randolph. 2009. "Dispositions, Abilities to Act, and Free Will: The New Dispositionalism." *Mind* 118(470): 323–351, doi:10.1093/mind/fzpo34.
- CRISP, Thomas M. and WARFIELD, Ted A. 2000. "The Irrelevance of Indeterministic Counterexamples to Principle Beta." *Philosophy and Phenomenological Research* 61(1): 173–184, doi:10.2307/2653408.
- DENNETT, Daniel C. 1984. "I Could Not Have Done Otherwise – So What?" *The Journal of Philosophy* 81(10): 553–565, doi:10.5840/jphil1984811022.
- . 2003. *Freedom Evolves*. New York: Viking Press / Penguin Books.
- EKSTROM, Laura Waddell. 2003. "Free Will, Chance, and Mystery." *Philosophical Studies* 113(2): 153–180, doi:10.1023/a:1023940209581.
- FARA, Michael. 2008. "Masked Abilities and Compatibilism." *Mind* 117(468): 843–865, doi:10.1093/mind/fzn078.
- FISCHER, John Martin. 1988. "Freedom and Miracles." *Noûs* 22(2): 235–252, doi:10.2307/2215861.
- . 1994. *The Metaphysics of Free Will*. Oxford: Basil Blackwell Publishers.
- FISCHER, John Martin and RAVIZZA, Mark. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.
- FRANKFURT, Harry G. 1969. "Alternate Possibilities and Moral Responsibility." *The Journal of Philosophy* 66(23): 828–839, doi:10.4324/9781315248660-2.
- . 1971. "Freedom of the Will and the Concept of a Person." *The Journal of Philosophy* 68(1): 5–20, doi:10.2307/2024717.
- FRANKLIN, Christopher Evan. 2011. "Masks, Abilities, and Opportunities: Why the New Dispositionalism Cannot Succeed." *The Modern Schoolman* 88(1-2): 89–103, doi:10.5840/schoolman2011881/26.
- GINET, Carl. 1966. "Might We Have No Choice?" in *Freedom and Determinism*, edited by Keith LEHRER, pp. 87–106. New York: Random House.

- . 1980. "The Conditional Analysis of Freedom." in *Time and Cause: Essays Presented to Richard Taylor*, edited by Peter VAN INWAGEN, pp. 171–186. Philosophical Studies Series n. 19. Dordrecht: D. Reidel Publishing Co., doi:10.1007/978-94-017-3528-5.
- GUSTAFSSON, Johan E. 2017. "A Strengthening of the Consequence Argument for Incompatibilism." *Analysis* 77(4): 705–715, doi:10.1093/analys/anx103.
- HORGAN, Terence E. 1977. "Lehrer on 'Could'-Statements." *Philosophical Studies* 32(4): 403–411, doi:10.1007/bf00368695.
- HUME, David. 1748. *Philosophical Essays Concerning Human Understanding*. London: Andrew Millar of the Strand.
- KANE, Robert H. 1999. "Responsibility, Luck, and Chance: Reflections on Free Will and Indeterminism." *The Journal of Philosophy* 96(5): 217–240, doi:10.2307/2564666.
- , ed. 2002. *The Oxford Handbook of Free Will*. Oxford Handbooks. New York: Oxford University Press. Second edition: Kane (2011a), doi:10.1093/oxfordhb/9780195178548.001.0001.
- , ed. 2011a. *The Oxford Handbook of Free Will*. 2nd ed. Oxford Handbooks. New York: Oxford University Press. First edition: Kane (2002), doi:10.1093/oxfordhb/9780195399691.001.0001.
- . 2011b. "Rethinking Free Will: New Perspectives on an Ancient Problem." in *The Oxford Handbook of Free Will*, edited by Robert H. KANE, 2nd ed., pp. 381–405. Oxford Handbooks. New York: Oxford University Press. First edition: Kane (2002), doi:10.1093/oxfordhb/9780195399691.001.0001.
- KENNY, Anthony John Patrick. 1973. "Freedom, Spontaneity, and Indifference." in *Essays on Freedom of Action*, edited by Ted HONDERICH, pp. 89–104. London: Routledge & Kegan Paul.
- KITTLE, Simon. 2015a. "Abilities to Do Otherwise." *Philosophical Studies* 172(11): 3017–3035, doi:10.1007/s11098-015-0455-8.
- . 2015b. "Free Will and the Ability to do Otherwise." PhD dissertation, Sheffield: Department of Philosophy, <https://etheses.whiterose.ac.uk/9523/>.
- LEHRER, Keith. 1968. "Cans Without Ifs." *Analysis* 29(1): 29–32, doi:10.1093/analys/29.1.29.
- . 1976. "'Can' in Theory and Practice: A Possible Worlds Analysis." in *Action Theory: Proceedings of the Winnipeg Conference On Human Action, held at Winnipeg, Manitoba, Canada, 9-11 May 1975*, edited by Myles BRAND and Douglas N. WALTON, pp. 241–270. Synthese Library n. 97. Dordrecht: D. Reidel Publishing Co.
- . 1990. *Theory of Knowledge*. 1st ed. Dimensions of Philosophy Series. Boulder, Colorado: Westview Press. Second edition: Lehrer (2000).
- . 2000. *Theory of Knowledge*. 2nd ed. Dimensions of Philosophy Series. Boulder, Colorado: Westview Press. First edition: Lehrer (1990).

- LEWIS, David. 1979. "Counterfactual Dependence and Time's Arrow." *Noûs* 13(4): 455–476. Reprinted, with a postscript (Lewis 1986b), in Lewis (1986a, 32–51), doi:10.2307/2215339.
- . 1981. "Are We Free to Break the Laws?" *Theoria* 47(3): 113–121. Reprinted in Lewis (1986a, 291–298), doi:10.1111/j.1755-2567.1981.tb00473.x.
- . 1986a. *Philosophical Papers, Volume 2*. Oxford: Oxford University Press, doi:10.1093/0195036468.001.0001.
- . 1986b. "Postscript to Lewis (1979)." in *Philosophical Papers, Volume 2*, pp. 52–66. Oxford: Oxford University Press, doi:10.1093/0195036468.001.0001.
- . 1997. "Finkish Dispositions." *The Philosophical Quarterly* 47(187): 143–158. Reprinted in Lewis (1999, 133–151), doi:10.1111/1467-9213.00052.
- . 1999. *Papers in Metaphysics and Epistemology*. Cambridge: Cambridge University Press, doi:10.1017/CBO9780511625343.
- MCKAY, Thomas J. and JOHNSON, David. 1996. "A Reconsideration of an Argument against Compatibilism." *Philosophical Topics* 24(2): 113–122, doi:10.5840/philtopics199624219.
- MELE, Alfred R. 2006. *Free Will and Luck*. Oxford: Oxford University Press, doi:10.1093/0195305043.001.0001.
- MILLER, Dickinson S. 1934. "Free Will as Involving Determination and Inconceivable Without It." *Mind* 43(169): 1–27. Published under the name "R.E. Hobart" , doi:10.1093/mind/XLIII.169.1.
- MOORE, George Edward. 1912. *Ethics*. Oxford: Oxford University Press.
- SAUNDERS, John Turk. 1968. "The Temptations of 'Powerlessness' ." *American Philosophical Quarterly* 5(2): 100–108.
- TAYLOR, Christopher and DENNETT, Daniel C. 2002. "Who's Afraid of Determinism? Rethinking Causes and Possibilities." in *The Oxford Handbook of Free Will*, edited by Robert H. KANE, pp. 257–280. Oxford Handbooks. New York: Oxford University Press. Second edition: Kane (2011a), doi:10.1093/oxfordhb/9780195178548.001.0001.
- TOGNAZZINI, Neal A. 2016. "Free Will and Miracles." *Thought* 5(4): 236–238, doi:10.1002/tht3.224.
- VAN INWAGEN, Peter. 1975. "The incompatibility of Free Will and Determinism." *Philosophical Studies* 27(3): 185–199, doi:10.1007/bf01624156.
- . 1983. *An Essay on Free Will*. Oxford: Oxford University Press.
- . 2000. "Free Will Remains a Mystery." in *Philosophical Perspectives 14: Action and Freedom*, edited by James E. TOMBERLIN, pp. 1–19. Oxford: Basil Blackwell Publishers. The Eight Philosophical Perspectives Lecture, doi:10.1111/0029-4624.34.S14.1.
- . 2004. "Freedom to Break the Laws." in *Midwest Studies in Philosophy 28: The American Philosophers*, edited by Peter A. FRENCH and Howard K. WETTSTEIN,

pp. 334–350. Boston, Massachusetts: Basil Blackwell Publishers,  
doi:10.1111/j.1475-4975.2004.00099.x.

- VIHVELIN, Kadri. 1988. “The Modal Argument for Incompatibilism.” *Philosophical Studies* 53(2): 227–244, doi:10.1007/bf00354642.
- . 2004. “Free Will Demystified: A Dispositional Account.” *Philosophical Topics* 32(1–2): 427–450, doi:10.5840/philtopics2004321/211.
- . 2013. *Causes, Laws, and Free Will. Why Determinism Doesn’t Matter*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199795185.001.0001.
- WHITTLE, Ann. 2010. “Dispositional Abilities.” *Philosophers’ Imprint* 10(12).
- WIDERKER, David. 1987. “On an Argument for Incompatibilism.” *Analysis* 47(1): 37–41, doi:10.1093/analys/47.1.37.

# Neopragmatist Inferentialism and the Meaning of Derogatory Terms —A Defence

DEBORAH RAIKA MÜHLEBACH

Inferentialism seems to be an unpopular theory where derogatory terms are concerned. Contrary to most theorists in the debate on the meaning of derogatory terms, I think that inferentialism constitutes a promising theory to account for a broad range of aspects of derogatory language use. In order to make good on that promise, however, inferentialism must overcome four main objections that are usually raised against Michael Dummett's and Robert Brandom's inferentialist explanations of derogatory terms. This paper aims at debunking these objections and thereby further developing the inferentialist interpretation of derogatory terms. I shall first discuss and reject three of the objections by pointing to the core assumptions of Brandomian inferentialism. Overcoming the fourth objection requires adjusting Dummett's and Brandom's explanation of the meaning of derogatory terms. In order to do so, I shall elaborate on the role that the explication of implicit material inferences plays with regard to different kinds of derogatory terms. The inferentialist account I am proposing fares better in terms of its explanatory power and broadness of application than Dummett's and Brandom's sketchy and oft-criticised views.

## 1 The Dismissal of Inferentialism for Derogatory Terms

Within philosophy of language, slurs have recently entered the limelight. Slurs are commonly seen as a highly offensive form of derogatory terms. They derogate their targets by virtue of some perceived membership in a social category pertaining to nationality, gender, sexual orientation, race, ethnicity, or religion which is valued negatively. Slurs are both pejorative and derogatory terms. The more general notion of pejorative terms additionally includes words that target people based on individual characteristics or single actions,

for example, “jerk” and “arsehole”. These terms are pejorative in the sense that in our discursive practices, we treat them as offensive. The notion of derogatory terms as I use it, refers to terms that structurally derogate their targets, i.e. they are part of and contribute to oppressive social structures.<sup>1</sup> In this paper I aim to focus on structurally derogating terms in the first place. I shall also talk about the offensiveness of slurs and pejoratives. This broad scope allows me to look into mechanisms of change whenever implicitly derogatory terms turn into slurs.

Inferentialism is unpopular where derogatory terms are concerned. Few advocate it to explain the derogatory force of such terms, the variation of force among them, their potential change of meaning over time, or why their derogatory force can be independent of the speaker’s intentions. Inferentialist accounts of these phenomena tend to be quickly dismissed, and consequently, there has not yet been much discussion about the extent to which inferentialism is capable of explaining other important aspects of derogatory language use. I think that the rejection of inferentialism is too quick; better than any other theory of derogatory terms, it explains the direct connection between the meaning of our terms and the broader practices in which they are used without losing sight of the semantic differences among various similar terms, such as “sl-t”<sup>2</sup> and “wh-re”.<sup>3</sup> An inferentialist account of derogatory terms needs to prove its worth on two fronts: on the one hand, of course, it needs to live up to the constructive task of accounting for various features of such expressions; but it also needs to be able to withstand pressure from a variety of objections that have been levelled against inferentialism—objections that

---

1 For a detailed discussion of this distinction see Mühlebach (2023a).

2 In what follows, I introduce “.” whenever I use a derogatory term or concept as an example which I do not consider to be part of my vocabulary. Even though certain uses and certain mentions of derogatory terms function non-derogatorily if the context of the utterance makes it sufficiently clear that the speaker does not utter the term in an endorsing way, the mere appearance of derogatory terms may trigger memories of violence. Thus, not semantic but broader political reasons lead me to not spell out the derogatory terms whose content I do not endorse.

3 For example, like perspectivalism (Camp 2013), it accounts for the systematicity of whole perspectives in terms of practical commitments. However, Camp only takes into account whole perspectives without allowing for commitments to specific assertions. Inferentialism, by contrast, preserves the difference between terms such as “sl-t” and “wh-re” which involve the same perspective, but slightly differ in meaning. Whilst both signal the allegiance to a sexist perspective, or even more specifically, the commitment to the claim that the target enjoys sex with too many people, only the latter additionally includes the commitment to a claim about the venality of the target. Explaining the meaning of these two terms only by pointing to a sexist perspective misses the difference in meaning.

have so far proven influential in putting people off inferentialist accounts of slurs.

It is this second, defensive task that I want to tackle in this paper. I leave the constructive task to another paper (Mühlebach 2023b), where I show how inferentialism can account for various distinctive features of derogatory terms, including notably their derogatory (as distinct from their offensive) force which is relatively autonomous from the speaker's intentions and the target's actual feeling of being hurt, the variation of this force among different derogatory terms, the semantic dimension of appropriated uses of derogatory terms by the target group, the fact that in some specific occurrences, the derogatory force of a term scopes out of its embedding, as well as that there are non-appropriated, non-derogatory uses of derogatory terms. Among other things, I argue there that structural derogation involved in derogatory language use is politically more pressing than its offensive potential, that the inferentialist framework does not impose one specific linguistic mechanism onto every type of derogatory expression but allows for various mechanisms to serve as the explanation of different phenomena of derogatory language use,<sup>4</sup> that appropriation by the target group involves semantic change, and that non-appropriated, non-derogatory terms, as well as the scoping out of embeddings, is best explained by making use of the inferentialist distinction between practical endorsement/non-endorsement of inferentially structured commitments, rather the mention/use distinction. But such a positive account, however elaborate, is only as robust as the inferentialist foundation it is built on. One also needs to do the negative work of showing that this foundation does not crumble under the weight of the major objections to it. It is therefore to such a defence of inferentialism that I now turn.

The aim of this paper is to deflect the four main objections raised against an inferentialist theory of derogatory terms. This involves adjusting and further developing Dummett's and Brandom's explanation of the meaning of derogatory terms so that it applies to a broad range of such terms. Accounting for the meaning of derogatory terms in this new way has three advantages: firstly, it enables us to understand using explicitly derogatory and implicitly derogatory terms as two distinct practices, and it explains the workings of

---

<sup>4</sup> For example, it highlights the importance of stereotypical ascriptions as a whole set of (practical) commitments if an expression makes use of a stereotype (such as in "French shower"). Within a broad inferentialist framework we can also make sense of changes in meaning and force over time and explain how formerly pragmatic mechanisms such as derogation through metaphorical force may turn the semantics of a derogatory expression.



each. This explanation amounts to situating different kinds of terms on a continuum between explicitly derogatory and purely descriptive terms, which, as I will argue, is crucial to understanding how and why the meaning and force of derogatory terms may change over time. Secondly, and contrary to most theories of derogatory terms, including Dummett's and Brandom's, it provides a general explanation of all kinds of derogatory terms, including gendered, racial, religious, ableist, homo- or transphobic derogatory terms, as well as terms for individual behaviour ("arsehole" or "jerk"). And thirdly, it suggests why criticising derogatory language use by merely making explicit the pernicious inferences that it licenses, often misses the point of the practice of using explicitly derogatory terms.<sup>5</sup>

My negative account complements Lynne Tirrell's (1999) inferentialist view of broader communicative situations in that it shows how such a framework affects how we need to model the semantics of derogatory terms more narrowly. Moreover, it differs significantly from Daniel Whiting's (2008, 2013) view which combines inferentialism with conjunctive non-cognitivism, since my paper clears the path for a full-blown account of the meaning of derogatory terms that does not resort to an additional theory in order to account for derogation through language use. Finally, Esa Díaz-León (2020) aims to defend a semantic strategy by presenting an inferentialist version of Christopher Hom's semantic externalism (Hom 2008, 2012). I take her view to fare better with regard to several objections that have been raised against truth-conditional content views (e.g. Cepollaro and Thommen 2019). However, since her view imports inferentialist ideas into the truth-conditional paradigm, I do not consider it to be a neopragmatist inferentialist account of derogatory terms as I defend it here.<sup>6</sup>

---

5 The same holds for merely challenging or blocking conventional or conversational implicatures and presuppositions in a specific speech situation.

6 Inferentialism turns the truth-conditional programme upside down by not starting off from reference, but rather from inferentially structured social practices and ultimately explaining reference through inference. Current truth-conditional accounts generally base their theories on the assumption of there being a neutral counterpart to every slur. According to these views, the counterpart and the slur share the same reference. As I have argued elsewhere (Mühlebach 2021), this leads them to ignore a broad range of derogatory terms that lack such a counterpart. Moreover, where *prima facie* applicable, these counterparts are themselves often so complex that the assumption of them being neutral is not true at best and morally highly problematic at worst.

## 2 Common Objections to the Inferentialist View

Inferentialism receives its name from the assumption that the content of a concept is determined by its inferential relations to other concepts. Concepts are expressed by terms, but a specific term may be ambiguous so that it expresses more than one concept (e.g. “light”) and a specific concept may be expressed by different terms (e.g. “red”, “rot”, “rouge”). Thus, two tokens of the same term differ in meaning if they express a different concept.<sup>7</sup> In Brandom’s pragmatist understanding of semantics, the picture of the inferentialist theory of meaning looks roughly as follows: the meaning of a sentence is determined by the role this sentence plays in the practice of making assertions and giving and asking for reasons. Furthermore, the meaning of a word or an expression is determined by the roles it can play in the assertions of this practice. For a broader understanding of linguistic meaning, we must of course consider a variety of pragmatic moves we can make in and with our speech acts, but as far as semantics are concerned, inferentialism confines itself to the practice of making assertions.<sup>8</sup>

Making assertions and giving and asking for reasons is a social practice that is modelled as a deontic scorekeeping game with different parties. It involves ascribing commitments to other parties as well as undertaking commitments oneself. The most important rule governing such a practice or game is that by making an assertion, you as the speaker undertake the responsibility to provide reasons for your utterance in case you are challenged by a hearer. For the whole practice to work, the different parties involved in this game keep score of both their own commitments and the commitments of the other speakers.

The different sentential roles that an assertion<sup>9</sup> can play depend on the inferential relations of this sentence to other assertions according to the con-

---

7 In what follows, I shall often use the expression “conceptual norms of discursive communities” in the sense in which conceptual and discursive norms are strongly connected with each other. Linguistic norms govern the use of terms which, in turn, express concepts that are expressed by these terms and are governed by the conceptual norms of the discursive community.

8 Jennifer Hornsby’s (2001, 138) objection that the inferentialist model cannot capture how individual speakers use the derogatory term in a particular occasion is misguided. Inferentialists do not hold that particular instances of (derogatory) language use can be explained on the grounds of semantics alone. But if we want to understand why certain words are so apt to be used as weapons while others are not, it is helpful to have a look at their semantics, too.

9 For Brandom, commitments come in the form of assertions, but there are other accounts of inferentialism such as Daniel Whiting’s (2007, 2013), which takes non-propositional attitudes to play a similar role in the game of giving and asking for reasons, at least in some cases.

ceptual norms of the discursive community (for the following, see [Brandom 1994, 157ff.](#)).<sup>10</sup> These relations come in two different forms—in the form of *commitment-preserving* inferential relations, and *entitlement-preserving* inferential relations. As far as the commitment-preserving inferential relations are concerned, as an utterer of an assertion I am committed to further assertions, the so-called *concomitant commitments*. Take the following assertions:

- (1) New York is to the East of Paris.
- (2) Paris is to the West of New York.
- (3) New York is not to the East of Paris.

If I claim (1), I am concomitantly committed to assertion (2), because according to the conceptual rules of the English-speaking community, these two assertions stand in a commitment-preserving relation to each other.<sup>11</sup>

Entitlement-preserving inferential relations, by contrast, do not compel commitments. Whenever I am entitled to make an assertion, i.e. my assertion is taken to be true by other scorekeepers, I am thereby also entitled to commit myself to the claims that materially follow from my assertion.<sup>12</sup> My entitlement to the claim “This is a dry, well-made match” entitles me to be committed to the further claim “It will light if struck.” I do not have to be so committed, however, because my first assertion is compatible with the claim “The match is at such a low temperature that friction will not succeed in igniting it” (see [Brandom 1994, 169](#)). The interplay of commitment-preserving and entitlement-preserving relations can be put in terms of material incompatibility: being committed to assertion (3) from above, which is incompatible with (1), precludes me from being entitled to claim (2).

- 
- 10 For reasons of simplicity, I only talk about *the* discursive community and its practical and conceptual norms here. However, our societies consist of several discursive (sub-)communities with somewhat different social practices and, hence, different practical and conceptual norms. Elsewhere ([Mühlebach 2021](#)), I model them as communities of practice (cf. [McConnell-Ginet 2011](#); [Anderson 2018](#)) and, by the example of the race term “black”, I argue that some of the political disputes among different communities of practice amount to semantic contestations of specific terms.
- 11 Lynne Tirrell ([1999, 146ff.](#)) distinguishes between three kinds of commitments: assertional, referential, and expressive commitments. Although the referential and the expressive commitments are semantically reducible to the basic assertional commitments, distinguishing them becomes relevant as soon as we establish a broader inferentialist theory of communication which involves the possibilities of criticising or challenging language use.
- 12 Brandom’s inferentialism concerns material rules (material inferences, material incompatibilities), and not just formal rules. These material rules depend on the content of the concepts involved.

The description of the inferential relations I have given so far highlights the intrapersonal or, according to Brandom, the *intercontent* dimension of these relations because the change of normative statuses (commitment/entitlement) produced by an assertion has consequences for the persons already committed or entitled to the claim in question. If I make an assertion, I present it as a reason for other assertions. By putting it forward as a premise from which other assertions can be materially inferred, I change the landscape of further claims I am committed to. Similarly, an assertion that is uttered changes the landscape of additional claims that everyone who is entitled to the assertion in question is now also entitled to.

However, the inferential relations are *interpersonally* significant, too. This is where entitlement-preserving relations become especially relevant. If I assert claim (1), I am thereby committed to sentences such as (2) which, according to the conceptual norms of my discursive community, stand in a commitment-preserving relation to (1). In undertaking this commitment, I license or entitle others to attribute that commitment to me. I license them to assume that I endorse both my assertion and the further claims that can be inferred from it. Moreover, I entitle my listeners and others to reassert my assertion, as well as the claims that follow from it.

Inferentialism is often taken to model linguistic exchanges as a rational, cooperative enterprise that is confined to propositional content. This view, however, misses a major part of the inferentialist theory. Inferentially structured propositions are an important part of the inferential web of commitments, but so are non-verbal perceptions and actions. The latter are a part of this web as so-called *language-entry* and *language-exit* transitions (see Brandom 1994, 233–234). According to the first, verbal claims cannot only be made in response to other verbal claims but also in response to non-linguistic, perceptible circumstances. With regard to the second, non-linguistically acting in response to some verbal claim is as much a possible move in the game of giving and asking for reasons as responding by making another claim. For example, cordially hugging somebody right after calling her a b-che are two moves that are as incompatible with each other, just as calling someone a b-che and saying what a lovely person they are would be.<sup>13</sup>

In a more general vein, inferentialists think of semantics in pragmatic terms: sentences are meaningful in the practice of making assertions and giving and

---

<sup>13</sup> This does not apply to reclaimed uses by the target group. According to the inferentialist view, the reclaimed term “b-che” would express a concept different from the derogatory “b-che” because it would license significantly different inferences.

asking for reasons. Making assertions includes much more than producing meaningful sound. It involves speakers and hearers both perceptively reacting to the world in accordance with the assertions that are made in this practice and acting upon them correspondingly. This whole practice of making assertions, in turn, is only meaningful as a practice in a broader communicative context. Our broader social norms enable and constrain which conceptual norms our discursive community abides by. The social embeddedness of terms becomes especially important in the case of verbal derogation. A whole set of propositional and practical commitments and, in consequence, the whole social practice surrounding the use of a specific term produce the derogatory force of this term.

Derogatory terms are based on, contribute to, and reinforce social structures of oppression. Social oppression occurs if, through everyday habits and practices, certain people are being systematically disadvantaged and treated unjustly (see Young 1990, chap. 2). Oppressive social structures facilitate or even call for derogatory expressions to name the vertical differences produced by these structures. Such terms, in turn, reinforce these structures: because our discursive practices are inferentially structured, the use of such terms licenses further oppressive speech and action if it is not effectively counteracted.<sup>14</sup> Note that negative evaluations do not per se contribute to oppressive social structures. Some of them are legitimate and harmless, and some are still treated as offensive. Terms that involve legitimate negative evaluations that are treated as offensive are pejoratives. In an egalitarian society there could be pejoratives, but no derogatory terms. However, given that social oppression is so pervasive and multidirectional, it is more than unlikely that egalitarian societies, and thus societies without derogatory language, will ever exist.

Inferentialist approaches to the meaning of derogatory terms are neither new nor undisputed.<sup>15</sup> Many contemporary contributions to the discussion on the semantics and pragmatics of derogatory terms take inferentialism into account, but most of them limit themselves to criticising some of its alleged core assumptions. I take the main objections to be the following four:

- (i) *Understanding without endorsement*: Listeners are able to understand sentences that contain derogatory terms even if they do not (morally)

---

<sup>14</sup> For a compelling example, see Tirrell's (2012) discussion of language in the Rwandan genocide.

<sup>15</sup> Dummett (1973) uses one example of derogatory terms, Brandom (1994, 2000) adopts it and explains its most basic semantic workings, and Tirrell (1999) develops an account of derogatory speech situations that draws on an inferentialist vocabulary but is not itself semantic.

endorse the inferences that have to be drawn from the utterance of these sentences. Inferentialism cannot explain this condition.

- (ii) *Overdemandingness*: Inferentialism seems to overlook the fact that bigoted as well as non-bigoted speakers understand a derogatory term before or even without knowing most of the inferences that arguably need to be drawn from its use.
- (iii) *Inability to explain variation in force*: There are subtle differences in pejorative force between different derogatory terms. Inferentialism fails to explain this complex variation of pejorative force, because we cannot assume that competent speakers have knowledge of all the properties that inferentialism takes to be semantically relevant.
- (iv) *Indeterminate reference*: Inferentialism faces difficulties in explaining the route from inference to reference when it comes to sentences containing derogatory terms.

As I shall show, the first three objections all call for similar explanations. They can be met by pointing to the inferentialist’s core assumptions. The fourth objection, however, requires some extra theoretical work. Whilst objection (iii) applies to slurs or pejoratives specifically, objections (i), (ii), and (iv) concern the inferentialist picture in general. By deflecting these objections, I aim to defend inferentialist semantics in general, and I show that this is helpful in order to understand how the derogatoriness of terms may change over time.

### 3 Meeting the Objections

#### 3.1 *Understanding Without Endorsement*

Timothy Williamson observes that “we find racist and xenophobic abuse offensive because we understand it, not because we fail to do so” (2009, 141). According to him, inferentialists are not able to explain this. According to Williamson, Dummett believes that speakers and listeners understand the concept B-CHE iff they are disposed to draw inferences according to the following introduction and elimination rules:

<i>B-che-Introduction:</i>	<i>B-che-Elimination:</i>
<i>x</i> is a German	<i>x</i> is a b-che
<i>x</i> is a b-che	<i>x</i> is cruel

These rules specify the term's inferential role in a specific language by using the vocabulary from this language without the term in question. The introduction rule states under which circumstances it is appropriate to apply the term in question. The elimination rule, in turn, states the consequences of applying the term. Hence the rules show how the term in question can be "introduced" to a language and then "eliminated" from its vocabulary so that we get a picture of the term's inferential role without resorting to the term itself.<sup>16</sup> Since, according to Williamson, non-xenophobic people who are not willing to draw the inferences from the "b-che" rules nevertheless do understand the term "b-che" perfectly well, he doubts that it is the disposition to draw these inferences that makes speakers and listeners understand the term.

This objection rests on two mistakes. First, inferentialists would not agree with Williamson's claim that non-bigoted speakers are not willing to draw the relevant inferences. Being disposed to draw inferences is not a matter of will. Rather, I fully understand a concept iff I know what inferences are licensed by its use, i.e. iff I know which inferences are correct according to the conceptual norms of the discursive community. The crucial question with regard to derogatory terms is whether I think that the inferences which are correct according to the conceptual norms of the discursive community are also morally and epistemically correct. Williamson's objection conflates the two standards of correctness, the semantic and the moral and epistemic standard. We evaluate somebody's understanding of a term with reference to the semantic standard of correctness, whereas the moral and epistemic standard leads people to criticise the use of certain terms and to refrain from using them. Understanding whilst refraining from using a term amounts to both knowing the inferential role of the term in question and rejecting a whole set of practices in which these inferences are treated as correct.

Thus, secondly, Williamson's objection conflates understanding a concept with using this concept. People can smoothly communicate even if they have different moral and epistemic standards because *understanding* a term does not necessarily imply endorsing its concomitant commitments, whereas *using* it necessarily implies endorsement.<sup>17</sup> Listeners may well know the

<sup>16</sup> Dummett's prime example is the introduction of the logical connective "&": the introduction rule for "&" is the transition from "p, q" to "p&q" and the elimination rule is the transition from "p&q" to "p, q".

<sup>17</sup> This difference can also be stated in terms of *attributing a commitment* to another person vs. *undertaking or adopting the commitment* oneself. See Brandom (2000, 169).

inferences that can be drawn from assertions that involve the term “b.che”, whilst refraining from endorsing some or even all of these inferences. If a listener understands but does not endorse the assertion in question, the utterer of the assertion is committed to the inference “x is more likely to be cruel than other Europeans” while the listener is not. However, the listener knows that the utterer is so committed. Yet the listener does not think that the utterer is entitled to the inference, since the listener herself does not endorse the inference. Thus, by using certain terms, the speaker is committed to the inferences that must be drawn from the utterance in question and he entitles his listener to draw these inferences, too. The listener, however, needs neither to be committed to these inferences, nor must she think that the speaker is entitled to draw these inferences if she does not endorse them.<sup>18</sup> If the listener does not endorse a certain inference, it is because she thinks that the inference in question is not sound, i.e. the assertion that is supposed to follow is not true and, in the case of derogatory terms, that it is morally and/or politically problematic.

### 3.2 *Overdemandingness*

Critiques of inferentialism hold that it is far too demanding with regard to the cognitive work speakers and listeners would have to undertake if they were really committed to all the inferences ascribed to them. Take the British boy Sebastian telling his friends in the aftermath of WWI:

(4) There are too many b-ches in town these days.

Some listeners of (4) may easily understand that Sebastian’s utterance is derogatory without exactly knowing what claims he is thereby committed to.<sup>19</sup> The problem is best captured by Hornsby who already leads us partially out of it:

---

18 As will become clear in my response to the objection of indeterminate reference, my version of an inferentialist account views the introduction rules as they are stated here as incorrect. According to the introduction rules that I will propose, my response to the understanding without endorsement objection amounts to the following view: if a listener understands a term, they know under which circumstances the bigoted speaker applies the term even if the listener thinks that the conditions of application do not obtain. And the listener also knows what consequences the use of the term will have even though the listener does not endorse these consequences. With regard to many strongly derogatory terms, this view amounts to the null-extension view as it has been proposed by Horn and May (2013, 2018).

19 See Jeshion (2013, 245) for a similar view.



With enough hard work, a historically minded social theorist might [...] provide the inferentialist with materials to make explicit that which is to be treated as implicit in each racist word. The question now is whether speakers themselves undertake commitments, which the historian uncovers. The answer appears to be *No*. [...] We are trying to account for something readily picked up by speakers of a certain social formation; and we have to allow for the fact that racist and other derogatory words can be passed on quite easily. If speakers' involvement with the ideology went as deep as it would need to in order to be implicit in their very use of words, then common understandings would be difficult to preserve. (Hornsby 2001, 137)

In short, the worry is this: there is a basic understanding of terms prior to many inferences we might draw from their use.

Two things have to be noted here. First, Hornsby confounds the inferentialist's use of "commitment" with an internalist understanding of "speaker meaning". Which claims the utterer of (4) is committed to, is not a question of conscious inferences made by the speaker in the first place, but a question of the term's inferential role in a discursive community. According to the inferentialist view regarding derogatory terms, it is indeed the case that historians, social scientists, and people with a sense of social mechanisms are the ones most likely to uncover the commitments of speakers in a discursive community. Only by making the commitments that are implicit in our discursive practices explicit do we see whether our commitments correspond to what we think we are saying. Sometimes our conceptual commitments are not so much in agreement with what we mean to say.

This, however, might give us an awkward picture of concept use in everyday communication. How should people understand each other if they do not know any of the important inferences to which they are committed? Thus, the second point is that conceptual understanding comes in degrees. If Sebastian's listeners do not take him to be committed to any of the assertions he is actually committed to, we would simply say that they do not understand the concept B-CHE at all. As soon as the listeners have some grasp on the basic inferences, they understand that the utterance is derogatory. In that case, however, they have a very poor understanding of the concept of b-che. They are not able to see that the alleged despicableness is explained by the likeliness to be cruel and not, for example, by some ascribed special visual appearance. A simple

distinction between understanding a concept and not understanding it is untenable. Various people are in different states of conceptual competence with regard to different concepts, but the difference between those states is not a distinction of kinds, but of degrees (Brandom 1994, 120).<sup>20</sup> The more inferential roles of a concept I know, the richer my understanding of the concept in question is. And the richer my understanding of the concept I use is, the more I move away from merely deploying a randomly picked up concept towards using it with the explicit knowledge about my linguistic and extra-linguistic commitments that come along with it.

It is certainly a helpful first move to attend to degrees of understanding in order to account for different understandings of derogatory terms by different people. However, we need not forget that slurs are part of politically relevant language use. Since politics are about collectively negotiating (social) realities, the contents of words that name these realities are often part of these negotiations, too. It is a constitutive part of such language use that people might disagree about correct uses in general or in different contexts. Thus the political complexity of such terms sometimes runs contrary to the interest of philosophers of language in always exactly determining the truth conditions of a sentence, or the exact set of inferences that come along with an assertion.

Moreover, the underlying complex and dynamic social structures of derogatory language use make it likely that many derogatory concepts function as cluster concepts, as has been suggested by Croom (2011, 353ff). With regard to the term “ch-nk”, for example, a full understanding of the term does not involve knowing about the long list of all the stereotypical ascriptions for Chinese people, such as being slanty-eyed, devious, good at laundry, etc., but contextual norms make it clear which parts of the whole set are relevant to a specific speech situation. However, stereotypes of a social group often systematically hang together and are part of a whole set of social meanings that affect the meaning of a term in a specific context (see Mühlebach 2022). Language users who are aware of the connections between these social meanings might well be taken to possess a deeper understanding of the terms they use than those who are not.

I have argued elsewhere that inferentialism is in a good position to account for the social complexities of such words (Mühlebach 2021). For example, it allows for different communities to have different uses of a specific term, as

---

<sup>20</sup> Cf. Higginbotham (1998, 149ff.) for a helpful distinction of different states of conceptual competence.

Luvell Anderson (2018) has argued regarding the n-word. Moreover, inferentialism provides the resources to theorise about the power relations among these communities and how these affect the meaning of such terms. As I have shown, that is true even if we take many slurring concepts to function as cluster concepts.

### 3.3 *Inability to Explain Variation in Derogatory Force*

Hom worries that inferentialism cannot explain why “speakers have a pretty good understanding of the lexical, negative ordering of slurs (e.g. the n-word is worse than ‘ch·nk’ is worse than ‘l·mey’, etc.)” (2010, 175, “.” added by the author). Since the list of licensed inferences would need to distinguish properties in very subtle ways and we cannot take competent speakers to have knowledge of all these properties, inferentialists are not able to explain the difference in force between, say, the n-word and “ch·nk”.

Hom’s insights into the variation of force are important, but, as I shall argue here, the variation can be equally accommodated by inferentialist semantics. In Hom’s view,

the derogatory content of an epithet is semantically determined by an external source. The plausible candidates for the relevant external social practices that ground the meanings of racial epithets are social institutions of racism. [...] An institution of racism can be modelled as the composition of two entities: an ideology, and a set of practices. (2008, 430–431)

If we state Hom’s claims in inferentialist terms, his explanation of derogatory force fits well into the broad inferentialist picture provided by Brandom or Tirrell. In a pragmatist understanding of inferentialism, the meaning of a term is determined externally through the inferences that are treated as valid by the discursive community.<sup>21</sup> Recall my sketch of the inferentialist framework called *deontic scorekeeping* from above: the content of a term is determined by the term’s inferential role in the game of giving and asking for reasons. The parties involved in this game keep score of what the other parties and they themselves are committed and entitled to. Every assertion

<sup>21</sup> Note that treating inferences as valid ones is not exhausted by those that are consciously treated as such. Inferentialism leaves room for the fact that people, individually or even collectively, are sometimes not aware of what they are doing. Explicating their implicit inferences amounts to bringing these inferences to the realm of reasoning.

that is made in this game alters the scores of the scorekeepers. In other terms, every assertion changes the set of claims to which the parties involved in the game are inferentially committed and entitled.

Scorekeeping is a social game. Which inferences count as correct depends on the social norms of the discursive community. The stronger the racism against a social group within a discursive community, the more numerous the racist inferences that are licensed by this community, and the more pernicious. The same holds true for sexism, xenophobia, and other discriminatory institutions. While the war-experienced French community directly after WWI would have been likely to treat the inference from “*x* is of German nationality” to “*x* is likely to be cruel” as correct, this is not what French communities would treat as a correct inference today.

In Hom’s explanation of the force of derogatory terms, a set of specific social practices and a web of negative ideological beliefs add up to the social institution of racism. Since inferentialism not only includes propositional commitments but extends to language-entry and language-exit transitions, non-verbal perception and action, which are important for Hom’s view, are not missing in the inferentialist account. Inferentialism captures the finely-grained ordering of different derogatory terms by the *set of commitments* that come along with these terms, and not by single inferences such as moving from “*x* is a b-che” to “*x* is likely to be cruel.” The variation of derogatory force is dependent upon the actual devaluation of the target in the practices in which the expression is used.<sup>22</sup> How strongly a target can be derogated by the content of a concept depends on the power relations within a given discursive community, i.e. on how much a target can be devalued according to their social, economic, and cultural capital within the discursive community.

In cases of strongly derogatory terms such as the contemporary meaning of the n-word, the set of commitments includes claims such as that black people are inferior to white people, as well as practical commitments to treat black people in dehumanising ways. Commitments are not necessarily consciously held beliefs, but rather commitments that are operative in a given practice

---

<sup>22</sup> Note that the offensive force and the derogatory of an expression are not necessarily the same. Offensiveness tells us something about which utterances a discursive community treats as appropriate, the derogatory force varies according to the actual devaluation of the target which is operative in the economic, social, and cultural practices in which the use of the expression is embedded. Elsewhere (Mühlebach 2021), I argue that criticism of language use often consists in that a sub-group of a discursive community tries to make other members of the community aware of the actual derogation involved in the use of a specific expression.

and that can be made explicit if we try to rationalise the practice. As we know from critical race theorists such as Linda Alcoff (2015), contemporary race relations in Anglo-American countries are such that whiteness functions as the unmarked norm whereas other races function as categories that deviate from it.<sup>23</sup> Thus, to date, the alleged inferiority of black people eventually amounts to their alleged inferiority to the white race. As Alcoff shows, it is a complicated question as to whether race relations can be transformed in a way such that in a distant future, whiteness could simply be one category among many others.

Given the role that the n-word historically played in practices of slavery, lynching, and other types of inferiorisation and dehumanisation, and given that the power relations between whites and blacks are still highly asymmetrical, the current use of the n-word still carries the weight of inferiorising black people and licensing to treat black people cruelly. By contrast, the US-American use of the term “l-mey”, which targets British people, licenses inferences that are much weaker in their derogatory force. In the US, there is no deeply rooted xenophobia against British people and there are no clear hierarchical power relations in play on which a web of deeply pernicious inferences could be based. If it is used at all, it licenses mocking behaviour of the British stereotype.

With regard to two or more expressions that target the same group of people, their derogatory force may still vary. For example, even though the n-word and “d-rkie” are both highly derogatory, the practices in which they have typically been used differ. The former involves dehumanising, inferiorising, and especially violent and cruel behaviour, whereas the latter also involves inferiorising, but rather patronising than cruel behaviour. At least, that is true for their historical use. The more they are used in the same practices and in the same way, the more their derogatory force, as well as their meaning, converges.

### 3.4 *Indeterminate Reference*

Both Williamson and Hom worry that the inferentialist faces the problem of unfixed references when explaining the meaning of derogatory terms. By

---

<sup>23</sup> Note, however, that this does not imply that whiteness is also unmarked for the ones who are dominated by whites. Sara Ahmed rightly observes that “whiteness is only invisible for those who inhabit it. For those who don’t, it is hard not to see whiteness; it even seems everywhere” (2004)

responding to their worry, I shall address their misunderstanding of Brandom and Dummett: in the case of derogatory concepts they assume a symmetrical relationship between introduction and elimination rules of concepts, whereas Brandom and Dummett take this relationship to be asymmetrical. My response in this section will be Brandomian in order to show what we are able to explain with the help of classical inferentialism. In section 4, however, I shall point to the limits of this Brandomian view, and in section 5 I shall show that the distinction between explicitly and implicitly derogatory terms is crucial to an explanation of meaning in derogatory language use.

Recall the explanation of a term's inferential role in a game of giving and asking for reasons. This role is best explained using a vocabulary from the same language that does not contain the term itself. We introduce the term into this language by stating the circumstances under which its application is appropriate (the introduction rule). And we state the consequences of applying the term (the elimination rule). According to Dummett, the “b-che”-introduction rule is the transition from “ $x$  is German” to “ $x$  is a b-che” and the elimination rule is the transition from “ $x$  is a b-che” to “ $x$  is cruel.” Williamson's and Hom's objection rests on the principle according to which expression  $E$  refers to  $X$  if “the hypothetical assignment of  $X$  as the reference of  $E$  makes  $R(E)$ , [i.e. the rules for the use of  $E$ ] truth-preserving” (Williamson 2009, 143). Williamson and Hom convincingly argue that there “is no determinate route from inference to reference” (Hom 2010, 175) because, according to the introduction and elimination rules of “b-che”, there is more than one set of objects to which “b-che” can refer.

Instead of spelling out in what ways the reference might be indeterminate, let me note that the objection rests on the assumption that the relationship between the introduction and the elimination rule of derogatory terms is symmetrical. However, both Dummett and Brandom treat derogatory sentences as clear examples of an asymmetry between those two rules. The difference between a symmetrical and an asymmetrical relation between introduction and elimination rules lies in that only with regard to the latter, the elimination rule introduces new inferences which were not yet part of the web of inferences that are relevant for the circumstances of application. Terms that are stably institutionalised and not just about to change do generally have introduction and elimination rules which are *in harmony* with each other, i.e. they are *symmetrical*. If we introduce a new term into an existing language and if its elimination rules are part of the web of inferences which already determines the introduction rules, this is called a *conservative extension* (see

Williamson 2009, 138–139). A *non-conservative extension*, by contrast, is built on the *asymmetrical* relation between these rules. In this case, the elimination rule involves further inferences which have not been part of the web of inferences relevant for the introduction rule yet. As we will see, conceptual changes in everyday language use and conceptual progress in science are phenomena in which we are confronted with asymmetry between the introduction and elimination rules.

The wide-spread confusion about symmetrical and asymmetrical introduction and elimination rules in the case of derogatory terms makes the indeterminate reference objection *prima facie* pointless because it targets a view that neither Dummett nor Brandom adhere to. However, there is something peculiar about the fact that all critics of the inferentialist position who are concerned with the reference problem take the introduction and elimination rules in the case of “b-che” to be symmetrical, even though Dummett and Brandom clearly argue for their asymmetrical relationship. In the remaining part of this paper, I shall explore in what ways this confusion can be made fruitful to adjust Dummett’s and Brandom’s insufficient account of a broad range of derogatory terms. Dummett and Brandom assign different roles to the asymmetry of introduction and elimination rules with regard to “b-che”. Take Dummett’s remarks on these rules first:

It [the distinction between introduction and elimination rules] remains, nevertheless, a distinction of great importance, which is crucial to many forms of linguistic change, of the kind we should characterize as involving the rejection or revision of concepts. Such change is motivated by the desire to attain or preserve a harmony between the two aspects of an expression’s meaning. A simple case would be that of a pejorative term, e.g. ‘B-che’. (1973, 454, “.” added by the author)

Claiming an asymmetrical relation of introduction and elimination rules with regard to ‘b-che’ is to say that “*x* is likely to be cruel” stands in no inferential relation to “*x* is German,” or put differently, “*x* is likely to be cruel” is not part of the inferential web of assertions which is relevant for utterances of the form “*x* is German.” In order to attain harmony between the introduction and elimination rules, one of them, or both, have to be revised or the concept needs to be rejected. Dummett thinks of B-CHE as an example of concepts which need to be rejected for reasons of harmony: there is no way in which

non-xenophobes would and should be licensed in drawing the inference from “*x* is German” to “*x* is disposed to be cruel.”<sup>24</sup>

Brandom reminds us, however, that this lack of conceptual harmony can be productive and does not lead to the rejection of concepts *per se*. According to him, the lack of harmony between the introduction and elimination of B-CHE is analogous to the lack of harmony in cases of conceptual change. A prime example of conceptual change is GRAVITY before and after Einstein’s theory of relativity. The Newtonian introduction rules were not in harmony with the elimination rules anymore, as soon as Einstein put forward his theory of relativity. The inferences of the elimination rule were not yet part of the inferential web of the introduction rule. According to Brandom, B-CHE works similarly: “the problem with ‘B-che’ or ‘n-gger’ is not that once we explicitly confront the material inferential commitment that gives them their content, it turns out to be *novel*, but that it can then be seen to be indefensible and inappropriate” (1994, 127, “.” added by the author).

Unfortunately, Brandom does not mention that in the case of GRAVITY, the introduction rules *do* change once we accept the new consequences (or elimination rules) that the application or use of this concept has, so that introduction and elimination rules ultimately come into harmony again. The concept of gravity could not be introduced with Newton’s criteria anymore as soon as Einstein’s theory of relativity, which made explicit the new inferences that had to be drawn from the use of this concept, was widely accepted. The introduction rules of GRAVITY had to be adjusted once the conceptual change brought about by scientific progress had been made explicit.

The same applies to conceptual changes and conceptual clarification in everyday language use. Take the example of RAPE. For a long time, there was no way in which, according to the predominant linguistic culture, rape could have happened between a married couple.<sup>25</sup> It was only with the rising awareness of women having sexual rights independently of their marital status that non-consensual sexual activity towards one’s wife could be socially and legally treated with the same consequences as non-marital rape. In order to attain harmony between the introduction and elimination rules of RAPE,

---

24 Or, as Brandom’s (2000, 70) remarks on Oscar Wilde’s trials where he was accused of blasphemy by the cross-examining Mr. Carson suggest, Wilde gave *the only answer he could give* by saying “‘Blasphemous’ is not a word of mine.” For a more detailed description of the events of Wilde’s trials, see Montgomery Hyde (1962, 121ff.).

25 Legal and dictionary definitions were such that “rape” applied only to a man’s penetrating a woman who was not his wife. See McConnell-Ginet (2006).



the former introduction rules had to be revised so that now, it is appropriate to apply the term “rape” whenever sexual activity and usually sexual intercourse without the other person’s consent is concerned, regardless of the relationship between the persons involved.

If we take Dummett as seeing the difference between endorsing and rejecting inferences once they are made explicit, and if we take Brandom to agree that in cases such as the conceptual change of GRAVITY, the introduction rules had to be adjusted once the changed material inferences were made explicit, their views on introduction and elimination rules do not differ significantly. It can be formulated as follows: if the set of claims that are the consequence of applying or using a concept is not yet part of the set of claims that license the use of this concept, this is a non-conservative extension. According to this picture, GRAVITY and RAPE on the one hand, and B-CHE on the other are examples of non-conservative extension with different moral consequences. In the first case, making the (new) commitments explicit leads to the revision of these concepts by adjusting their introduction rules. In the case of the latter, making the commitments explicit and, as a consequence, adjusting the introduction rules leads non-bigoted speakers to reject its use altogether. In both cases Williamson’s and Hom’s worry does not apply anymore. Once the introduction and elimination rules are in harmony again, the reference is determinate.

The idea behind making explicit the material inferences that are implicit in our discursive practices is the following: our language use is governed by the conceptual norms that are embedded in our broader social practices. The inferences that are licensed by our everyday use of concepts are material, not formal inferences. With regard to many of our concepts it is not obvious what the circumstances of their appropriate application and the consequences of their use are. The point of explicating the inferences hitherto implicit in using GRAVITY or RAPE, i.e. the claims to which we are committed according to the conceptual norms of our discursive community, is to bring these implicit inferences into the realm of reasoning. Once these commitments are on the table, they can be criticised or justified, and the concepts can be rejected or revised. In cases of revision, we adjust the introduction or elimination rules so that they are again in harmony with each other, at least as long as further relevant material inferences are made explicit.

There are several historical cases of derogatory language use in which the explication of the formerly implicit inferences has led to a widespread rejection of the concepts in question. Take CH-NAMAN and the US American

use of OR-ENTAL as two examples of such implicitly derogatory concepts. For a long time, their introduction rules were considered to be the moves from “Chinese man” to “Ch-naman” and “person from a Near or Middle Eastern country” to “Or-ental”. But their underlying social practices were formed in such a way that most of their uses committed the speakers to devaluating and exoticising inferences.<sup>26</sup> Only with the rising awareness of these underlying social structures was it possible to make explicit the formerly implicit inferences. As many English speakers do not openly approve of the devaluating and exoticising conceptualisation of people from China, Asia, or the Near and Middle Eastern countries, the use of these terms has diminished remarkably and so made room for alternatives such as “Chinese person”.<sup>27</sup>

#### 4 Broadness of Application

Brandom holds that his remarks on the concept of *b-che* “should go over *mutatis mutandis* for pejoratives in current circulation” (2001, 86). Moreover, he thinks that what makes this French epithet from WWI especially suitable for semantic investigations is that “we are sufficiently removed from its practical effect to be able to get a theoretical grip on how it works” (2001, 89). I doubt that “*b-che*” does the work Brandom expects it to do. The explanation does not generally hold true for all kinds of derogatory or pejorative terms such as “*sl-t*” or “*arsehole*”, nor does this explanation help us understand the phenomenon of explicitly derogatory terms, including “*b-che*” itself. Our temporal and linguistic-cultural detachment from its usage in WWI, I contend, is exactly the reason why it is difficult to understand the basic workings of this term’s use.

A closer look at the practice of expressing explicitly derogatory concepts suggests that not every derogatory concept is defective in the way Brandom assumes and thus it is questionable whether we know enough about the phenomenon of derogatory or pejorative terms in general if we understand the way in which the concept *B-CHE* is defective. Our English vocabulary makes

---

<sup>26</sup> See Stavroula Glezakos (2013) for an illuminating discussion of “Ch-naman”.

<sup>27</sup> Contemporary examples of terms that are widely seen as purely descriptive, categorising social terms are “woman” or “black person”. As feminist and critical race scholars show, however, their use is still governed by more or less hidden sexist and racist practices. This does not mean that they have to be rejected altogether, but rather that both a reconceptualisation and a change of the underlying social structures are necessary—talking about persons-gendered-as-woman and persons-racialised-as-Black is an example of the former; feminist and anti-racist political contestations are examples of the latter.

use of various kinds of derogatory terms targeting gender, race, ethnicity, nationality, religion, sexual orientation, ability, physical appearance, single actions, or patterns of behaviour. As sexism, racism, xenophobia, ableism, homo- and transphobia are institutionalised differently, they come in different forms and in different degrees of derogatoriness. In light of their variety, the conclusion that derogatory or pejorative terms always express defective concepts is too hasty.

Take, for example, the term “arsehole”. It is undisputed that this term is a pejorative. Nevertheless, I am sceptical of both its derogatoriness and defectiveness. If we take an utterer of “*x* is an asshole” to be committed to claims along the lines of “*x* arrogantly allows himself to enjoy special advantages,” there is nothing defective to be found here. For if we follow Aaron James (2012, 205ff.) in his analysis of “arsehole”, we see that the use of the term is mostly gendered—hence the “allows *himself*” from above—in that our unjust social arrangements make it possible mostly for men to arrogantly allow themselves to enjoy special advantages. Thus, even though “arsehole” is pejorative, i.e. we treat it as an offensive term, it is not derogatory. It is not based on nor contributes to oppressive social structures. On the contrary, it might even be a suitable means to point to behaviour that takes advantage of unjust social structures. Moreover, although we might always be able to put our message in a less forceful and more diplomatic way than by using the pejorative term “arsehole”, it does not express a defective concept. All of the inferences involved are valid.

However, many of the pejorative and derogatory concepts that are currently in circulation are indeed defective, and thus Brandom’s analysis of B-CHE might still prove valuable to understanding those. But even among the concepts that license flawed inferences, we find a considerable number that do not fit the model of “descriptive” circumstances of application and “evaluative” consequences of use proposed by Brandom’s discussion of the B-CHE example.<sup>28</sup> Take the case of gendered derogatory terms such as “sl-t”, “b-tch”, or “S-ssy”. Lauren Ashwell (2016) has argued that gendered slurs do not have “purely descriptive” and unproblematic circumstances of application. The circumstances of application of “sl-t”, for example, already contain the evaluative description “*x* has sex with too many partners.” If we follow Ashwell,

<sup>28</sup> Brandom notes that “[a]lthough they are perhaps among the most dangerous, highly charged words—words that couple ‘descriptive’ circumstances of application with ‘evaluative’ consequences of application—they are not alone in incorporating inferences we may need to criticize” (2001, 87).

our everyday language does not seem to provide any non-evaluative way of picking out who the target of the assertion “*x* is a sl-t” is.<sup>29</sup> Thus, Brandom’s explanation of the “b-che” case does not seem to apply to gendered derogatory terms, either.

The non-applicability of Brandom’s explanation of “b-che” to other common derogatory terms such as “arsehole” and “sl-t” calls for a re-examination of his account of derogatory terms. Its fruitfulness for understanding cases such as “Ch-naman” and “Or-ental” notwithstanding, it does not fully apply to the phenomenon of explicitly derogatory terms such as the n-word, “sl-t,” “ch-nk,” “k-ke,” or “arsehole”. Moreover, it even does not rightly capture the case of “b-che” because it conflates two distinct practices—the use of explicitly derogatory and the use of implicitly derogatory terms.

## 5 Implicitly vs. Explicitly Derogatory Terms

As discussed above, terms such as our historical examples “Ch-naman” and “Or-ental” have not always been considered pejorative. They were institutionalised within the English-speaking discursive community, which has been dominated by white people, in a way that has treated them as purely descriptive terms. However, the dominant use of these terms functioned derogatorily, and with the rising awareness of the devaluating and exoticising conceptualisation of Ch-namen and Or-entals, the formerly implicit inferences could be made explicit. By bringing them into the realm of reasoning, people had to take a stance by either openly committing themselves to pernicious inferences or by refraining from using the terms and engaging in practices of using alternative terms. The terms that philosophers of language are usually concerned with, by contrast, are explicitly derogatory terms such as the n-word, “sl-t,” “ch-nk,” or “k-ke”. To say that these are explicitly derogatory terms is to say that their use is put under strict social constraints and that more or less competent English speakers know that by using them, they are strongly derogating their target. The way in which these terms are used among bigots, or to hurt somebody, and the way their use is sanctioned suggest that users of such terms know both that and in what sense these terms are derogatory.

If people roughly know about the *that* and *how* of the derogation involved, the main inferences that have to be drawn from the use of derogatory terms

---

<sup>29</sup> Note that if we find a non-evaluative description, such as in a sociological or meta-language vocabulary, for some gendered slurs, then the objection of indeterminate reference, as it has been raised in the case of “b-che”, could be raised against the case of the respective gendered slur.

are explicitly available in the realm of reasoning. Whilst in the case of other politically significant terms such as our historical example “Ch·naman” and contemporary uses of “woman”, “disabled”, or “immigrant”, we still need to work out in what relation to derogatoriness these stand, the use of the n-word, for example, is institutionalised in a way in which people who use it both know that it is not used synonymously with “black person”, and that to call people that term is to ascribe racial inferiority to them and not, for example, the predisposition to cruelty. Similarly, competent French speakers in the period surrounding WWI knew that with the use of “b·che” they were derogating their target because of the (alleged) predisposition to cruelty and not because of a physical appearance held to be different from a specific norm. And competent English speakers know that with the term “sl·t”, they are derogating their target because of (a behaviour that indicates) having sex with too many partners and not because of being too assertive.

But if we take the main inferences that are licensed by the use of explicitly derogatory terms to already be more or less explicit, this affects the introduction rules of the terms in question. There is no reason as to why the cases of the n-word, “ch·nk”, or “b·che” should be different from making explicit the inferences with regard to the concepts of gravity, temperature, and rape. Recall the case in which the changed consequences of the use of GRAVITY have been made explicit once Einstein’s theory of relativity was fully established. By accepting these new consequences of use, the circumstances of application had to be adjusted, too. To continue with introducing the concept of gravity with Newton’s old criteria, which are not in harmony with Einstein’s consequences of use, is pointless if there are new criteria available according to which the circumstances of application that are in harmony with Einstein’s consequences of use can be framed.

Applied to the explicitly derogatory n-word, the circumstances in which for a speaker, the application of this term is appropriate are those in which the speaker is already committed to the claims “x is a black person” and “x is inferior to white people” (among others). Similarly, the circumstances of application in the case of “b·che” would not only involve “x is German,” but also “x is likely to be cruel.” These “new” commitments do not merely add the speaker’s evaluative attitude towards the target to the descriptive content of the term, rather they shape the content of what the speaker is talking about. This interpretation brings the workings of racist and xenophobic terms closer to those of sexist terms (“sl·t”) and terms for individual behaviour

(“arsehole”) in that all of them require evaluative descriptions to be part of their circumstances of appropriate application.

If we introduce evaluative descriptions into the circumstances of application, this has at least two benefits, besides enabling a better understanding of the bigoted practice of using derogatory terms. Firstly, we can make sense of the inferentialist claim that assertions function as premises and consequences of inferential moves in the game of giving and asking for reasons. Scorekeepers draw the inferences that are licensed by the speaker’s assertion. Thus, the assertion serves as a premise for further moves in the scorekeeping game. But it is also a consequence of former claims made in the game of giving and asking for reasons. By requiring the conditions of application of, say, the n-word to include “x is black” and “x is inferior to white people,” we can explain why the assertion “x is a n-gger” leads to puzzled reactions on the part of a listener both in the case in which the speaker is talking about a Mexican person and in the case in which her record of assertions commits her to claims such as that race is socially constructed and that people are to be treated equally regardless of their race.<sup>30</sup> In both cases the circumstances of application are not given so that making sense of the speaker’s assertion becomes difficult.

Note that we can judge whether the circumstances of application are given according to two different standards: the semantic and the epistemic and moral standard. Given my own set of commitments, a bigoted speaker may correctly apply the n-word in a specific situation in the sense in which they correctly use the term according to the conceptual norms of N-GGER. They meet the semantic standard. However, since I take most of their inferentially organised web of commitments to be both epistemically and morally false, there is, according to my view, no situation in which the circumstances of application of the n-word are given. In this regard, the inferentialist position is the same as Hom and May’s (2018) null-extension view: most derogatory terms do not refer to anything in our world because they involve commitments to epistemically false claims.

The two standards allow us to distinguish between different types of language critique. On the one hand, we can criticise uses of terms if they violate the conceptual norms of the concepts they aim to express. Just as in my example from above, we can point to semantically false uses of the n-word because

---

<sup>30</sup> Just as in the incompatibility case of using “b-che” and hugging the target before, this does not apply to the reclaimed use of the n-word by some members of the target group.

we know that the speaker must not be committed to the claims “*x* is Mexican” or “*x* is racialised-as-Black” if they want to apply the n-word according to the norms of the bigoted discursive community. These claims are not part of the elimination rule determining the content of the n-word. On the other hand, we can criticise the use of the n-word not only in specific situations, but in general. Its use always commits us to epistemically false and morally problematic claims, so that we should refrain from using it altogether if we want to make epistemically correct and morally unproblematic claims about the world.

The second benefit of including evaluative descriptions into the circumstances of application is that the indeterminate reference objection that has been raised against the inferentialist position becomes obsolete. Williamson and Hom have worried that the term “b-che” does not definitely refer if we take its introduction rule to be the transition from “*x* is German” to “*x* is a b-che” and its elimination rule the transition from “*x* is a b-che” to “*x* is likely to be cruel.” By treating explicitly derogatory terms as expressing concepts whose main inferences are already made explicit in the realm of reasoning—which is why they are treated as derogatory terms—we must change their introduction rules so that they are in accordance with their elimination rules again, at least as long as the meaning of these terms is not about to change, or as long as we do not learn about new implicitly operative inferences. Above I have shown what this looks like regarding the examples of “b-che” and the n-word. If the leap between the circumstances of application and the consequence of use is bridged this way, the elimination rule of a specific derogatory term does not add any new claim that is not yet part of the web of inferences relevant to the introduction rule. Hence, the objection of unfixed reference does not apply anymore.

If we acknowledge that the use of explicitly derogatory terms is stably institutionalised in our discursive practices, we need to explain not only why we should refrain from using many of them, but also how their usage does not create communicational impasses. Many people still use the n-word and as wrong as they are in doing so, they do it quite successfully. Most importantly, their use of this term does not necessarily lead them to materially incompatible commitments. Bigoted speakers do not use the n-word *despite* the fact that it commits them to the claim that black people are inferior to whites. They use it *because* they are so committed. If we seek to understand how the n-word, “ch-nk”, or “sl-t” are correctly used, we need to turn to the commitments of their users, and not to the commitments of people who refrain from using

these terms, because it is the former who uphold the linguistic practice of consistently using such derogatory terms.

The fact that many derogatory terms are consistently used, however, suggests that making explicit pernicious inferences is but a start in the enterprise of criticising derogatory language use. Merely making pernicious inferences explicit in the cases of “n-gger”, “f-ggot”, or “k-ke” does not bring us far if our opponent disagrees about them being pernicious. In this case, the criticism has to be more sophisticated and amounts to criticising a whole set of social practices. Other cases such as the use of “arsehole” call for yet another kind of criticism. Making explicit the inferences involved in its use, I contend, does not tell us anything about why we should refrain from using it on certain occasions (or altogether?).

Thus, on the one hand, the explication of implicit inferences is crucial to criticising derogatory language use, but on the other, it does not do all the work that is required for different kinds of derogatory terms. By taking into account a variety of implicitly and explicitly derogatory terms and finding common mechanisms for their basic workings, the version of inferentialism about derogatory terms that I have proposed enables us to understand that explicitly derogatory terms differ from implicitly derogatory terms only in degree. Most theorists who work on the meaning of derogatory terms set out to account for derogatory terms based on their divergence from their allegedly purely descriptive correlates. Thereby, they neglect to explain the continuum between highly derogatory and purely descriptive terms. Understanding this continuum as depending both on differences in inferential commitments and on different degrees of their explicitness, however, is crucial to see that the meaning of terms and their derogatory force may change over time if their underlying social practices change.

## 6 Conclusion

In this paper, I defended an inferentialist explanation of derogatory terms against the main objections that are usually raised against the inferentialist view. I rejected three of the four main objections by showing that they are based on misunderstandings of the core assumptions of inferentialist semantics. These criticisms can be accommodated by pointing to the role that conceptual norms play in inferentialist semantics, the degrees of conceptual competence, and to the social embeddedness of linguistic and non-linguistic moves that can be made in the inferentialist scorekeeping practice.




I further argued both that the indeterminate reference objection rests on a misunderstanding of Dummett's and Brandom's rules for the application and use of derogatory terms, and that this misunderstanding is productive in that these rules are indeed misguided for a broad range of derogatory terms. Dummett and Brandom discuss their "b-che" example as a case in which the inferences that are licensed by the use of the term are not yet part of the web of inferences that were relevant to determine whether the application of the term was appropriate. This usually happens whenever the meaning of a term changes. If the meaning of a term changes, either the rules of application or the consequences of using the term change first. Initially, these changes are only implicit in the practices. By making these changes explicit, we bring the new inferences into the realm of reasoning. Brandom thinks that this mechanism not only underlies meaning change, but also the use of derogatory terms.

However, this is not the case. At most it explains the mechanism involved whenever members of a discursive community who thought that a term was purely descriptive become aware of its derogatory function. Historical examples are "Ch-naman" and the US American use of "Or-ental". In such cases, these members thought that the introduction rule for " $x$  is a Ch-naman" were " $x$  is a Chinese man" and they learn that the elimination rule is " $x$  is less civilised than Europeans". As in the case of meaning change, making explicit the derogatory consequences of a term's use amounts to a change of the application rules, too. In a case of meaning change, for instance, it was no longer possible to have Newton's application rules for "gravity" once Einstein's theory of relativity had been widely adopted. The application rules had to be adjusted to be in accordance with Einstein's theory. Analogously, in the case of explicitly derogatory terms, it is no longer possible to turn to the past ignorant discursive community—which took "Ch-naman" to be a neutrally descriptive term while implicitly treating it as a derogatory one—if we look for the current application rules. Once the pernicious consequences of the term's use are made explicit, its application rules change, too.

Accounting for explicitly derogatory terms thus amounts to including a whole set of inferences into the rules of application, some of which might involve evaluative components. If we do not allow devaluating claims to enter the set of inferences that is relevant to determine whether the application of a term is appropriate, we not only have to face Williamson's and Hom's indeterminate reference objection, but we also misdescribe the practice of derogatory language use. Instead of following Dummett and Brandom in

looking at the commitments of language users who refrain from using a specific term, we need to turn our gaze toward the commitments of those who engage in the practice of using the explicitly derogatory term in question. Highlighting the commitments of bigoted language users, in turn, suggests that a successful critique of derogatory language use does not merely consist in problematising the use of a specific term, but rather taking issue with a whole set of practices in which the use of this term is embedded. These practices may involve a broad range of sets of commitments, from highly pernicious commitments to sets of commitments that are not morally pernicious. The view I have put forward is thus an inferentialist account of conceptual content in general which treats explicitly derogatory concepts as part of a continuum with implicitly derogatory terms and terms that are not derogatory at all.\*

Deborah Raika Mühlebach

 0000-0002-5720-3437

Freie Universität Berlin

d.muehlebach@fu-berlin.de

## References

- AHMED, Sara. 2004. “Declarations of Whiteness: The Non-Performativity of Anti-Racism.” *Borderlands* 3(2). Republished as Ahmed (2006).
- . 2006. “The Nonperformativity of Antiracism.” *Meridians* 7(1): 104–126. Republication of Ahmed (2004), doi:[10.2979/mer.2006.7.1.104](https://doi.org/10.2979/mer.2006.7.1.104).
- ALCOFF, Linda Martin. 2015. *The Future of Whiteness*. Cambridge: Polity Press.
- ANDERSON, Luvell. 2018. “Calling, Addressing, and Appropriation.” in *Bad Words. Philosophical Perspectives on Slurs*, edited by David SOSA, pp. 6–28. Oxford: Oxford University Press, doi:[10.1093/oso/9780198758655.001.0001](https://doi.org/10.1093/oso/9780198758655.001.0001).
- ASHWELL, Lauren. 2016. “Gendered Slurs.” *Social Theory and Practice* 42(2): 228–391, doi:[10.5840/soctheorpract201642213](https://doi.org/10.5840/soctheorpract201642213).
- BRANDOM, Robert B. 1994. *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Cambridge, Massachusetts: Harvard University Press.
- . 2000. *Articulating Reasons: An Introduction to Inferentialism*. Cambridge, Massachusetts: Harvard University Press.

---

\* Thanks to Sally Haslanger, Samia Hesni, Rebekka Hufendiek, Justin Khoo, Dominique Kuenzle, Michael O’Leary, Melanie Sarzano, Jennifer Saul, Christine Sievers, Pietro Snider, Marie van Loon, and Markus Wild for helpful comments on earlier drafts of this paper. Special thanks to Matthieu Queloz for inspiring discussions about issues raised here. I am grateful to the anonymous reviewers of this and other journals whose comments helped to significantly improve the content of this paper.

- . 2001. "Reason, Expression, and the Philosophic Enterprise." in *What Is Philosophy?*, edited by C. P. RAGLAND and Sarah HEIDT, pp. 74–95. New Haven, Connecticut: Yale University Press.
- CAMP, Elisabeth. 2013. "Slurring Perspectives." *Analytic Philosophy* 54(3): 330–349, doi:10.1111/phib.12022.
- CEPOLLARO, Bianca and THOMMEN, T. 2019. "What's Wrong with Truth-Conditional Accounts of Slurs ." *Linguistics and Philosophy* 42(4): 333–347, doi:10.1007/s10988-018-9249-8.
- CROOM, Adam M. 2011. "Slurs." *Language Sciences* 33: 343–358, doi:10.1016/j.langsci.2010.11.005.
- DÍAZ-LEÓN, Esa. 2020. "Pejorative Terms and the Semantic Strategy." *Acta Analytica* 35(1): 23–34, doi:10.1007/s12136-019-00392-2.
- DUMMETT, Michael A. E. 1973. *Frege: Philosophy of Language*. London: Gerald Duckworth & Co.
- GLEZAKOS, Stavroula. 2013. "Words Gone Sour?" in *Reference and Referring*, edited by William P. KABASENCHE, Michael O'ROURKE, and Matthew H. SLATER, pp. 385–404. Topics in Contemporary Philosophy n. 9. Cambridge, Massachusetts: The MIT Press, doi:10.7551/mitpress/9581.001.0001.
- HIGGINBOTHAM, James. 1998. "Conceptual Competence." in *Philosophical Issues 9: Concepts*, edited by Enrique VILLANUEVA, pp. 149–162. Atascadero, California: Ridgeview Publishing Co., doi:10.2307/1522965.
- HOM, Christopher. 2008. "The Semantics of Racial Epithets." *The Journal of Philosophy* 105(8): 416–440, doi:10.5840/jphil2008105834.
- . 2010. "Pejoratives." *Philosophy Compass* 5(2): 164–185, doi:10.1111/j.1747-9991.2009.00274.x.
- . 2012. "A Puzzle about Pejoratives." *Philosophical Studies* 159(3): 383–405, doi:10.1007/s11098-011-9749-7.
- HOM, Christopher and MAY, Robert C. 2013. "Moral and Semantic Innocence." *Analytic Philosophy* 54(3): 293–313, doi:10.1111/phib.12020.
- . 2018. "Pejoratives as Fiction ." in *Bad Words. Philosophical Perspectives on Slurs*, edited by David SOSA, pp. 108–131. Oxford: Oxford University Press, doi:10.1093/oso/9780198758655.001.0001.
- HORNSBY, Jennifer. 2001. "Meaning and Uselessness: How to Think about Derogatory Words." in *Midwest Studies in Philosophy 25: Figurative Language*, edited by Peter A. FRENCH and Howard K. WETTSTEIN, pp. 128–141. Boston, Massachusetts: Basil Blackwell Publishers, doi:10.1111/1475-4975.00042.
- HYDE, Harford Montgomery. 1962. *The Trials of Oscar Wilde*. Mineola, New York: Dover Publications.
- JAMES, Aaron. 2012. *Assholes. A Theory*. Garden City, New York: Doubleday & Co.

- JESHION, Robin. 2013. "Expressivism and the Offensiveness of Slurs." in *Philosophical Perspectives 27: Philosophy of Language*, edited by John HAWTHORNE, pp. 231–259. Hoboken, New Jersey: John Wiley; Sons, Inc., doi:10.1111/phpe.12027.
- MCCONNELL-GINET, Sally. 2006. "Why Defining Is Seldom 'Just Semantics': Marriage, 'marriage,' and other minefields." in *Drawing the Boundaries of Meaning. Neo-Gricean Studies in Pragmatics and Semantics in Honor of Laurence R. Horn*, edited by Betty J. BIRNER and Gregory WARD, pp. 223–246. Studies in Language, Companion Series n. 80. Amsterdam: John Benjamins Publishing Co., doi:10.1075/slcs.80.13mcc.
- . 2011. *Gender, Sexuality, and Meaning. Linguistic Practice and Politics*. Oxford: Oxford University Press.
- MÜHLEBACH, Deborah Raika. 2021. "Semantic Contestations and the Meaning of Politically Significant Terms." *Inquiry* 64(8): 788–817, doi:10.1080/0020174X.2019.1592702.
- . 2022. "Tackling Verbal Derogation: Linguistic Meaning, Social Meaning and Constructive Contestation." in *The Political Turn in Analytic Philosophy. Reflections on Social Injustice and Oppression*, edited by David BORDONABA-PLOU, Victor Fernández CASTRO, and José Ramón TORICES, pp. 175–198. Eide n. 11. Berlin: Walter De Gruyter.
- . 2023a. "A Non-Ideal Approach to Slurs." *Synthese* 202(3), doi:10.1007/s11229-023-04315-y.
- . 2023b. "Meaning in Derogatory Social Practices." *Theoria* 89(4), doi:10.1111/theo.12476.
- TIRRELL, Lynne. 1999. "Derogatory Terms: Racism, Sexism, and the Inferential Role Theory of Meaning." in *Language and Liberation: Feminism, Philosophy, and Language*, edited by Christina HENDRICKS and Kelly OLIVER, pp. 41–80. New York: State University of New York Press.
- . 2012. "Genocidal Language Games." in *Speech and Harm: Controversies over Free Speech*, edited by Ishani MAITRA and Mary Kate MCGOWAN, pp. 174–221. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199236282.001.0001.
- WHITING, Daniel. 2007. "Inferentialism, Representationalism and Derogatory Words." *International Journal of Philosophical Studies* 15(2): 191–205, doi:10.1080/09672550701383483.
- . 2008. "Conservatives and Racists: Inferential Role Semantics and Pejoratives." *Philosophia* 36(3): 375–388, doi:10.1007/s11406-007-9109-1.
- . 2013. "It's Not What You Said, It's the Way You Said It: Slurs and Conventional Implicatures." *Analytic Philosophy* 54(3): 364–377, doi:10.1111/phib.12024.
- WILLIAMSON, Timothy. 2009. "Reference, Inference, and the Semantics of Pejoratives." in *The Philosophy of David Kaplan*, edited by Joseph ALMOG and Paolo LEONARDI, pp. 137–158. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780195367881.001.0001.

- YOUNG, Iris Marion. 1990. *Justice and the Politics of Difference*. Princeton, New Jersey: Princeton University Press. Second edition Young (2012).
- . 2012. *Justice and the Politics of Difference*. 2nd ed. Princeton, New Jersey: Princeton University Press. With a foreword by Danielle S. Allen, doi:10.1515/9781400839902.

# Sensitivity and Inductive Knowledge Revisited

GUIDO MELCHIOR

The orthodox view about sensitivity and induction has it that beliefs formed via induction are insensitive. Since inductive knowledge is highly plausible, this problem is usually regarded as a *reductio* argument against sensitivity accounts of knowledge. Some adherents of sensitivity defend sensitivity against this objection, for example by considering backtracking interpretations of counterfactuals. All these extant views about sensitivity and induction have to be revised since the problem of sensitivity and induction is a different one. Regardless of whether we allow backtracking interpretations of counterfactuals, some instances of induction yield insensitive beliefs whereas others yield sensitive ones. These results are too heterogeneous to provide a plausible sensitivity account of inductive knowledge. Induction remains a serious problem for sensitivity accounts of knowledge.

## 1 Sensitivity and induction: the discussion so far

Nozick suggests that if *S* knows that *p*, then *S*'s belief that *p* tracks truth. He thinks that subjunctive conditionals can best capture this truth-tracking relation. Moreover, he argues that we have to take the belief-forming method into account. Nozick (1981, 179) provides the following definition of knowing via a method:

*S* knows, via method (or way of believing) *M*, that *p* iff (1) *p* is true (2) *S* believes, via method or way of coming to believe *M*, that *p* (3) In the nearest possible worlds where *p* is false and where *S* uses *M* to arrive at a belief whether (or not) *p*, *S* does not believe, via *M*, that *p* (4) In the nearest possible worlds where *p* is true and where

S uses M to arrive at a belief whether (or not)  $p$ , S believes, via M, that  $p$ <sup>1</sup>

Condition (3) is the sensitivity condition, which I will focus on here, and condition (4) is the adherence condition.<sup>2</sup> Nozick formulates these modal conditions on knowledge as subjunctive conditionals, but he analyzes their truth conditions in terms of possible worlds. For the sake of convenience, and in accordance with the literature, I will use possible world terminology for formulating conditions (3) and (4). For the purposes of this paper, nothing hinges on this decision, as we acquire the same results for sensitivity and induction when talking in terms of subjunctive conditionals.<sup>3</sup>

Sensitivity accounts of knowledge face several major problems. First, it has been claimed that they preclude us from having inductive knowledge, as Vogel (1987, 1999) and Sosa (1999) contend. Second, they lead to implausible instances of closure failure as Kripke (2011) argues.<sup>4</sup> Third, Sosa (1999) and Vogel (2000) argued that sensitivity faces severe problems concerning higher-order knowledge about the truth of one's own beliefs. In this paper, I will

- 
- 1 Subjects can believe a proposition via various methods. Nozick argues that S knows that  $p$  *simpliciter* if the dominant method, the one that outweighs the other methods, fulfills conditions (3) and (4). These subtleties will not concern us here.
  - 2 Nozick does not provide a clear terminology. He suggests that condition (3) expresses the fact that S's belief is sensitive to the falsity of  $p$ , whereas (4) states that S's belief is sensitive to the truth of  $p$ . Accordingly, (3) and (4) jointly guarantee the complete sensitivity of S's belief. Hereinafter, I will stick to the terminology dominant in the literature that calls condition (3) the sensitivity condition and condition (4) the adherence condition.
  - 3 The situation is more subtle concerning condition (4). As Starr (2019) points out, "counterfactual conditional" and "subjunctive conditional" are usually used interchangeably in the philosophical literature. However, condition (4) is a so-called true-true subjunctive, since its antecedent and its consequent are both true. True-true subjunctives are not counterfactual conditionals in the literal sense. Nozick provides a specific semantics in terms of possible worlds that delivers a differentiated picture about the truth-values of true-true subjunctives, but true-true subjunctives are trivially true according to the standard Lewis/Stalnaker semantics. For discussions of the semantics of true-true subjunctives, see McGlynn (2012), Cogburn and Roland (2013), and Walters (2016). The sensitivity condition (3), in contrast, which is the focus of this paper, is a counterfactual conditional in the literal sense, given that the truth condition (1) for  $p$  is fulfilled. DeRose (2004) argues against safety and in favor of sensitivity, saying that we have clear intuitions about the truth conditions of real counterfactuals, i.e. of counterfactuals with false antecedents, but not of true-true subjunctives. This criticism can be extended to Nozick's adherence condition. In this paper, I focus on sensitivity. Consequently, we can ignore these subtleties concerning Nozick's adherence condition.
  - 4 For a defense of Nozick's knowledge account against Kripke's objection, see Adams and Clarke (2005).

focus on the first objection.<sup>5</sup> However, we will see in the last section that the problem of inductive knowledge has structural similarities to the problem of higher-order knowledge.

Despite the well-known challenges that sensitivity accounts of knowledge face, the sensitivity principle is intuitively appealing, leading to a “second wave” of sensitivity accounts, as Becker and Black (2012) label it. These accounts aim at defending a sensitivity-based theory of knowledge that avoids the problems that have been raised for Nozick’s (1981) original account.<sup>6</sup> Accordingly, the results about sensitivity and induction are not only relevant for Nozick’s original theory but also for these descendants.<sup>7</sup>

Vogel and Sosa argue for the claim that making sensitivity a necessary condition on knowledge rules out inductive knowledge by means of examples; they provide cases where a subject plausibly knows via induction although her belief is insensitive. Here are two cases:

CHUTE. On his way to the elevator, Ernie releases a trash bag down the chute from his high-rise condo. Walking along the street Ernie thinks about the trash and forms the belief that the trash is in the basement. Plausibly, Ernie knows that his bag is in the basement. But what if, having been released, it still (incredibly) were not to arrive there? That presumably would be because it had been snagged somehow in the chute on the way down (an incredibly rare occurrence), or some such happenstance. But none of these would affect Ernie’s belief, so he would still believe that the bag

---

5 A further type of problem is raised by Luper (1984) who points out that Nozick’s account of knowing via a method faces a technical problem when it comes to *one-sided methods* that can recommend believing that  $p$  but cannot recommend believing that  $\neg p$ . Intuitively, we want to allow knowledge via one-sided methods, but according to Luper-Foy, they necessarily violate the sensitivity condition and, therefore, cannot yield knowledge. Luper-Foy discusses a modification of Nozick’s sensitivity principle that avoids this problem but finally rejects this version too. For another discussion of this problem, see Williamson (2000). For an overview of the discussion about sensitivity and its problems, see Melchior (2020).

6 See DeRose (1995, 2017), Roush (2005), Becker (2007), and the contributions in Becker and Black (2012).

7 Sensitivity has not only been utilized for explaining knowledge but also for analyzing other epistemic concepts. Enoch, Spectre and Fisher (2012) argue that sensitivity is crucial for legal proof in order to explain why statistical evidence alone is not sufficient proof in the court. In this paper, I focus on the consequences for sensitivity accounts of *knowledge*. For a discussion of sensitivity, induction, and checking, see Melchior (2019). For a sensitivity-based theory of discrimination, see Melchior (2021).



has arrived in the basement. His belief seems not to be sensitive, therefore, but constitutes knowledge anyhow, and can correctly be said to do so. (See Sosa 1999, 145–146)

**HEARTBREAKER.** Sixty golfers are entered in the Wealth and Privilege Invitational Tournament. The course has a short but difficult hole, known as the “Heartbreaker.” Before the round begins, Jonathan thinks that, surely, not all sixty players will get a hole-in-one on the “Heartbreaker.” [See Vogel (1999), 165]<sup>8</sup>

These are cases of beliefs that are based on inductive reasoning, more specifically, inductive reasoning about particulars, as Vogel puts it.<sup>9</sup> He argues that knowledge about particulars via inductive reasoning is highly plausible. Intuitively, Ernie knows that the trash is in the basement, and Jonathan knows that not all sixty players will get a hole-in-one. However, in each case the target beliefs are insensitive. Consequently, sensitivity is not necessary for knowledge.

Sosa and Vogel argue against sensitivity accounts of knowledge by presenting examples of insensitive inductive beliefs that plausibly constitute knowledge. They need not argue for the stronger claim that *any* belief formed via induction is insensitive to make their point. The weaker claim that there are some plausible cases of inductive knowledge that involve insensitive beliefs is sufficient for their purpose. Nevertheless, the stronger view that any belief formed via induction is insensitive is the dominant one in the current debate.<sup>10</sup>

8 In conversation, it has been pointed out that *Heartbreaker* is a particularly convincing example. Nevertheless, one might regard its target proposition, that not all players get a hole-in-one, as a lottery proposition, which many think precludes it from being known. However, the problem of sensitivity and induction does not rely on assuming that the target proposition is a lottery proposition, as other cases presented in Vogel (1987, 1999) and in this paper show. For a discussion of lottery propositions, see Hawthorne (2004).

9 Sensitivity is, following Nozick (1981), usually defined as a feature of beliefs relative to a particular method. In cases of inductive knowledge, the relevant method is inductive reasoning. Hence, I will assume in the following that inductive reasoning is the relevant belief-forming method. Accordingly, for determining the sensitivity of inductive beliefs, we consider possible worlds where the method of inductive reasoning remains constant. In order to acquire the result that *Chute* and *Heartbreaker* are instances of induction, it has to be assumed that the belief-forming bases in these cases are instances of (tacit) inductive reasoning, an assumption that is usually only implicitly made in the literature.

10 One might object that induction is obviously not always insensitive because beliefs in necessities, which can also be formed via induction, are vacuously sensitive. This is true for orthodox seman-

Sosa's and Vogel's line of argumentation against sensitivity accounts of knowledge is not unopposed. One standard defense of sensitivity is proposed by Becker (2007). He accepts the view that induction yields insensitive beliefs, but he argues that this does not create a devastating objection to sensitivity accounts of knowledge. He admits that if we know propositions  $p_1 \dots p_n$ , then we do not have inductive knowledge that  $p_{n+1}$  is true. However, we still have knowledge about the *probability* of  $p_{n+1}$ .<sup>11</sup> Thus, our view about knowledge via induction rests on a confusion according to Becker. We cannot have knowledge via induction that  $p_{n+1}$ ; what we do know are propositions in the neighborhood of this proposition. Becker's account not only rejects inductive knowledge but also provides an explanation of our mistaken intuition that we can have this kind of knowledge. However, knowledge via induction is widely accepted. Accordingly, most philosophers are presumably not willing to bite the bullet of rejecting inductive knowledge for the gain of acquiring a sensitivity-based account of knowledge. We will take up Becker's account later and see that his solution faces additional problems.

Vogel, Sosa, and Becker agree that the subjects' beliefs in cases like *Chute* and *Heartbreaker* are insensitive, but draw conflicting conclusions as to whether this claim creates a serious problem for sensitivity accounts of knowledge. Until recently, the view that induction yields insensitive beliefs has remained unchallenged. Wallbridge (2018) takes up this objection to sensitivity accounts of knowledge and argues that properly understood, the purported counterexamples fail to succeed because the beliefs formed via induction are actually sensitive, not insensitive. Focusing on Sosa's chute case, Wallbridge argues that Ernie sensitively believes that the rubbish is in the basement. He claims that in some cases, in order to avoid "miracles", i.e. events that would not easily have happened, counterfactuals have to be interpreted as backtracking. According to a backtracking interpretation, counterfactual conditionals can be evaluated without keeping the past fixed until the time at which the counterfactual antecedent obtains. Wallbridge argues that, according to this

---

tics for counterfactuals and counterpossibles. However, it is still worth discussing whether the popular view about the insensitivity of induction holds also for the vast majority of contingent truths. Moreover, there is good reason to think that the orthodox semantics for counterfactuals should be rejected for having the counterintuitive consequence that all counterpossibles are vacuously true. For a discussion of an impossible worlds account of sensitivity, which delivers the result that not all beliefs in necessities are vacuously sensitive, see Melchior (2021).

11 For a similar take, see Roush (2005, 65f).

backtracking analysis, Ernie's belief is sensitive.<sup>12</sup> He suggests that other examples presented by Vogel (1987, 1999) and Pritchard (2012) can be analyzed analogously. Wallbridge is not particularly clear about his conclusion. In the abstract, he claims to show that inductive knowledge is sensitive. In the conclusion, Wallbridge (2018, 8) makes the weaker claim that "there are cases of sensitive inductive knowledge" and leaves the reader with a challenge, concluding that "if there are cases of insensitive inductive knowledge then they have yet to be pointed out."

In Section 2, I will show that the situation concerning induction and sensitivity is more subtle than opponents and defenders of sensitivity accounts of knowledge claim it to be. Some inductive processes yield sensitive beliefs, others yield insensitive beliefs, regardless of whether we opt for a backtracking or a non-backtracking interpretation of counterfactual conditionals. In Section 3, I will argue that this is problematic since the subjects in the cases presented are concerning inductive reasoning intuitively in similarly good epistemic situations. Hence, sensitivity accounts of knowledge are committed to making implausibly heterogeneous predictions about the knowledge status of subjects who believe via induction.

## 2 Sensitive and insensitive induction

In this section, I will discuss instances of enumerative and temporal induction and backtracking and non-backtracking interpretations of counterfactual conditionals. First, let me make some preliminary remarks about backtracking and non-backtracking counterfactuals. Lewis (1973) distinguishes between backtracking and non-backtracking counterfactuals. Non-backtracking counterfactuals keep the past fixed until the time at which the counterfactual antecedent obtains, whereas backtracking counterfactuals do not.<sup>13</sup> He argues that only non-backtracking counterfactuals can be used for analyzing causal dependencies. For example, in order to determine whether event *c*

---

12 In fact, Wallbridge's argumentation is more subtle. He distinguishes between a weak and a strong reading of sensitivity, analogously to weak and strong safety. To avoid miracles, the strong reading requires a backtracking interpretation according to which Ernie's belief turns out to be strongly sensitive. A weak reading of sensitivity does not require backtracking to avoid miracles, but Ernie's belief fulfills weak sensitivity even according to a non-backtracking analysis. Hence, Ernie's belief is sensitive under both readings of sensitivity. However, these subtleties are not crucial for the following argumentation.

13 See Menzies (2014). For a discussion of backtracking counterfactuals, see Khoo (2017).

caused event  $e$ , we consider those possible worlds that are identical with the actual world until the time where  $c$  does not obtain.<sup>14</sup>

In this paper, I will remain neutral about whether counterfactuals are correctly interpreted as backtracking or non-backtracking. Rather, I will investigate the consequences of these two interpretations for sensitivity accounts of knowledge. Let me emphasize the point of considering backtracking counterfactuals. We can say that whether  $S$ 's induction-based belief is sensitive depends on whether "the minimal 'change' from truth to falsity of  $p$  keeps the inductive evidence for  $p$  intact."<sup>15</sup> In terms of possible worlds, the sensitivity of  $S$ 's belief depends on whether the inductive evidence is available to  $S$  in the nearest possible worlds where  $p$  is false. If we interpret counterfactuals exclusively as non-backtracking, then we only consider possible worlds that do not differ from the actual world until the point at which the counterfactual antecedent obtains. If we allow for backtracking interpretations of counterfactuals, then we need not keep the past fixed until that point. Hence, backtracking or non-backtracking interpretations make a difference concerning which nearest possible worlds are considered and consequently whether a belief is judged to be sensitive or not.

In the following, I will present and analyze further cases of induction. We will see that some cases yield insensitive beliefs whereas others yield sensitive beliefs, regardless of whether counterfactuals can be backtracking or not. I will distinguish between enumerative induction where we draw an inference from objects  $o_1$ - $o_n$  to  $o_{n+1}$  and temporal induction where we draw an inference about an object  $o$  from time  $t_1$ - $t_n$  to  $t_{n+1}$ .<sup>16</sup> In each of these

---

14 We must distinguish two different claims about backtracking counterfactuals, a more specific claim that a particular counterfactual is backtracking and a general claim that there can be backtracking counterfactuals. Accordingly, we can distinguish two different dependence relations between backtracking counterfactuals and possible worlds. Given that we accept in general that there can be backtracking counterfactuals, whether a particular counterfactual is backtracking or not depends on what the nearest possible worlds are where the antecedent is false. In this case, the nearest possible worlds determine whether a counterfactual is backtracking. However, when backtracking counterfactuals in general are questioned, it is rather the other way around. Whether counterfactuals can be backtracking determines which of the nearest possible worlds where the antecedent is false we have to consider.

15 I am indebted to an anonymous reviewer for this formulation.

16 This distinction is not meant to be exhaustive as there might be instances of induction that cannot be clearly classified either as enumerative or as temporal. Moreover, like the contemporary discussion on induction and sensitivity, I will focus on inductive knowledge about particulars. Thus, generalizations of the form "All  $x$  are  $F$ " are not the conclusions of the inductive reasonings considered. For a discussion of inductive generalizations and sensitivity, see Roush (2005, 65f).

cases, the method of belief formation in question is induction.<sup>17</sup> Moreover, the cases have to be understood in a way such that the subjects are intuitively in equally good epistemic positions concerning the inductive conclusion in that (1) the evidence for believing the premises is equally strong; (2) the numbers of cases  $n$  observed is equally large (or the time interval observed is equally long); (3) the relevant similarity between the induced case  $c_{n+1}$  and observed cases  $c_1$  to  $c_n$  is equally strong (or the similarity between the basic conditions for  $o$  of the induced time point  $t$  and the observed interval  $i$ ); (4) there are no rebutting or undercutting defeaters available to the subjects; and (5) the predicates involved are equally projectible. Take, first, the following example of enumerative induction that yields an insensitive belief:

RAVEN (enumerative induction). Carl observes that  $\text{raven}_{n_1}$ – $\text{raven}_{n_n}$  is black and infers that  $\text{raven}_{n_{n+1}}$ , which he has not observed, is black. Ravens are typically black, though not necessarily, since there also exist rare mutations like albino ravens. In the nearest possible worlds where  $\text{raven}_{n_{n+1}}$  is not black, it is such a rare mutation. However,  $\text{raven}_{n_1}$ – $\text{raven}_{n_n}$  is black in these nearest possible worlds and Carl believes via observation of  $\text{raven}_{n_1}$ – $\text{raven}_{n_n}$  and induction that  $\text{raven}_{n_{n+1}}$  is black. Thus, his belief that  $\text{raven}_{n_{n+1}}$  is black is insensitive.

This analysis holds independently of whether counterfactuals are allowed to be backtracking or not. In both cases, the nearest possible worlds where  $\text{raven}_{n_{n+1}}$  is not black are such that it is an albino raven but where  $\text{raven}_{n_1}$ – $\text{raven}_{n_n}$  is black. In these possible worlds, Carl still believes via induction that  $\text{raven}_{n_{n+1}}$  is black. Hence, his belief is insensitive. Thus, in RAVEN, a case of enumerative induction, the subject believes insensitively, regardless of whether we allow backtracking interpretations of counterfactuals or not.<sup>18</sup>

Notably, a similar case of *temporal* induction yields a different outcome:

<sup>17</sup> However, I do not mean that the subjects in these cases explicitly draw inductive inferences. Rather they can be drawn implicitly and automatically. Nozick already developed an account of inferential knowledge. See Nozick (1981, 233f) and Baumann (2012). Since I do not understand induction here as an explicit process of drawing inferences, I will ignore this account. However, Nozick's account of inferential knowledge provides the same results as to whether the inductive processes investigated here are sensitive or insensitive.

<sup>18</sup> We will soon reflect on cases of enumerative induction that behave differently.

BLACKBIRD (temporal induction). Miles observes that blackbird<sub>n</sub> has been black until yesterday and believes via induction that blackbird<sub>n</sub> is black right now.

Non-backtracking: We only consider the nearest possible worlds where blackbird<sub>n</sub> is not black *right now*, and, hence, worlds where blackbird<sub>n</sub> has been black until yesterday. We ignore possible worlds where blackbird<sub>n</sub> changed its color earlier and worlds where it has never been black. In the nearest possible worlds considered Miles believes via observation and induction that blackbird<sub>n</sub> is black right now. Hence, his belief is insensitive.

Backtracking: If counterfactuals are backtracking the situation is different. In this case, the nearest possible worlds where blackbird<sub>n</sub> is not black right now are presumably such that it is an albino blackbird<sub>n</sub> that has been white all the time. They are not worlds where it changed the color since yesterday. Accordingly, in the nearest possible worlds where blackbird<sub>n</sub> is not black right now Miles does not believe via observation and induction that it is black right now. Thus, Miles' inductive belief that blackbird<sub>n</sub> is black right now is sensitive.<sup>19</sup>

So far we have seen that in the enumerative induction case of *Raven*, Carl's belief is insensitive no matter whether counterfactuals can be backtracking or not. However, in *Blackbird*, a case of temporal induction, Miles' belief is insensitive if counterfactuals are non-backtracking but sensitive if they are backtracking. At this point, one might suppose that enumerative induction is typically insensitive whereas the sensitivity of temporal induction depends on whether we opt for a non-backtracking interpretation or a backtracking one. However, this generalization is incorrect as the following cases will show. Take a second instance of enumerative induction that delivers *sensitive* beliefs in case of non-backtracking and backtracking counterfactuals:

EXAMINER (enumerative induction). Ina is a lazy examiner. When she has received all the exams she throws a dice and all the examinees get the same grade. For a particular test, she throws a 2 and, accordingly, marks all exams with B. Rachel is an examinee and

<sup>19</sup> It might be disputable whether instances of enumerative induction can plausibly have a backtracking reading. However, in the case of temporal induction, the concept of backtracking and non-backtracking interpretations is highly plausible.

does not know Ina's habits. Rachel asks numerous peers about their grades. Among them are peers of whom she knows that they were better prepared than herself and peers of whom she knows that they were worse prepared. All peers report that they got a B. Rachel forms the belief that she also got a B. The nearest possible worlds where Rachel does not get a B are such that Ina's dice throw delivered a different result than 2 and all students got a different grade than B, but the same one. In these possible worlds, Rachel does not believe via testimony and induction that she got a B. Thus, her belief that she got a B on the exam is sensitive.

The grades of all students are determined at the same time. Thus, no matter whether counterfactuals can be backtracking or not, the nearest possible worlds where Rachel does not get a B are such that all the other students do not get a B. In these possible worlds, Rachel does not believe via testimony and induction that she got a B. Thus, Rachel's belief is sensitive, regardless of whether counterfactuals can be backtracking or not. *Raven* and *Examiner* are both cases of enumerative induction. In *Raven*, the target belief is insensitive, no matter whether counterfactuals can be backtracking or not, and in *Examiner*, it is sensitive in both cases.

So far we have reflected on one case of temporal induction, *Blackbird*, where the belief is insensitive with a non-backtracking interpretation of counterfactuals and sensitive with a backtracking interpretation. We will now see that temporal induction can deliver different sensitivity results in different cases. Let's sketch a further case:

T – SHIRT (temporal induction). Sarah has seen Tim wearing a red T-shirt the whole day until 30 minutes ago and forms the inductive belief that Tim is wearing a red T-shirt right now.

Is Sarah's belief that Tim is wearing a red T-shirt right now sensitive? This depends on how we fill in the details. Let us consider two different scenarios:

Scenario 1: Tim and Sarah are on a hiking trail and they split thirty minutes ago. Sarah has seen Tim wearing a red T-shirt the whole day and forms the inductive belief that Tim is wearing a red T-shirt right now. Tim does not have another T-shirt with him. Thus, he could not easily get a fresh T-shirt. Suppose further that Tim accidentally grabbed a red T-shirt in the morning, but that he might easily have grabbed a T-shirt of a different color. If coun-

terfactuals can be backtracking, then the nearest possible worlds where it is false that Tim is wearing a red T-shirt right now are such that he grabbed a T-shirt of any other color in the morning. In these possible worlds, Sarah does not believe via observation that Tim was wearing a red T-shirt until thirty minutes ago and, therefore, does not believe via induction that he is wearing a red T-shirt right now. Thus, her belief is sensitive. If counterfactuals can only be non-backtracking, then we only consider possible worlds where Tim recently changed his T-shirt. In this case, Sarah's belief that Tim is wearing a T-shirt right now formed via observation and induction is insensitive. Hence, for Scenario 1, we acquire the same result as for *Blackbird*—a non-backtracking interpretation of counterfactuals implies insensitive beliefs, and a backtracking interpretation sensitive beliefs.

Scenario 2: Sarah and Tim were on a hiking trail until 30 minutes ago where Tim was wearing a red T-shirt. In fact, for security concerns, Tim only wears red T-shirts for hiking. Thus, it is not easily possible that he had a non-red T-shirt for hiking. After the hiking trail, Sarah and Tim split and Tim walks downtown for a drink. Sarah has seen Tim wearing a red T-shirt the whole day until 30 minutes ago and forms the inductive belief that Tim is wearing a red T-shirt right now. If counterfactuals can only be non-backtracking, then we only consider possible worlds where Tim recently changed his T-shirt. In these possible worlds, Sarah believes via observation and temporal induction that he is wearing a red T-shirt right now and, consequently, her belief is insensitive. However, even if we allow for backtracking counterfactuals, then the nearest possible worlds where Tim is not wearing a red T-shirt right now are such that he changed it recently downtown, given Tim's strict habit of only wearing red T-shirts for hiking.<sup>20</sup> Again, Sarah believes that Tim is wearing a red T-shirt right now and her inductive belief turns out to be insensitive.

In both scenarios, a non-backtracking interpretation of counterfactuals yields insensitive beliefs, but if we allow for backtracking interpretations, then Scenario 1 yields a sensitive belief whereas Scenario 2 yields an insensitive belief. In this respect, whether one's belief is sensitive in cases of temporal induction depends on how the cases are spelled out in detail.

---

<sup>20</sup> Due to the modal details of the case, possible worlds where Tim changed his T-shirt downtown are closer than possible worlds where he did not wear a red T-shirt until 30 minutes ago. Nevertheless, Scenario 2 has to be understood such that it is highly unlikely that Tim changes his T-shirt downtown. The inductive inference in Scenario 2 still has the same epistemic strength—according to the factors briefly mentioned earlier and more thoroughly analyzed later—as in the other cases considered, including Scenario 1.



We can now summarize the acquired results about sensitivity and induction: **Raven** is a case of enumerative induction. Carl’s belief that  $raven_{n+1}$  is black is *insensitive* regardless of whether counterfactual conditionals can be backtracking or not. **Examiner** is a further case of enumerative induction. However, Sarah’s belief that she got a B for the exam is *sensitive* regardless of whether counterfactuals can be backtracking or not. **Blackbird** is a case of temporal induction. Miles’s belief that  $blackbird_n$  is black is *insensitive* if counterfactuals can only be non-backtracking, but it is sensitive if they can be backtracking. As for **T-shirt**, a further case of temporal induction, sensitivity depends on how we fill in the details. In Scenario 1 and 2, Sarah’s belief that Tim is wearing a red T-shirt right now is insensitive if counterfactuals can only be non-backtracking. If they can be backtracking, then her belief is sensitive in Scenario 1 but insensitive in Scenario 2. These results are captured in table 1:

Table 1: Whether the belief is sensitive in non-backtracking and backtracking variants of the four cases.

Induction type	Case	Non-backtracking	Backtracking
Enumerative	Raven	–	–
	Examiner	+	+
Temporal	Blackbird	–	+
	T-shirt	–	+/-

Let me provide a more systematic analysis: Suppose S observes that object  $o$  has property F from  $t_1$  to  $t_n$  and believes via temporal induction that  $o$  has property F at  $t_{n+1}$ . If we generally accept that counterfactuals can be backtracking, then S’s inductive belief is sensitive only if, in the nearest possible worlds where  $o$  is not F at time  $t_{n+1}$ ,  $o$  is not F from  $t_1$  to  $t_n$ .<sup>21</sup> This is the case if worlds where  $o$  lost property F from  $t_n$  to  $t_{n+1}$  are more remote than worlds where  $o$  does not have property F from  $t_1$  to  $t_n$ , e.g. if it is more crucial for

21 I assume here that observation is sensitive concerning  $o$  being F, i.e. observation of  $o$  from  $t_1$  to  $t_n$  would not deliver that  $o$  is F from  $t_1$  to  $t_n$  if  $o$  were not F from  $t_1$  to  $t_n$ . If observation does not fulfill this sensitivity condition, then the sensitivity conditions for backtracking counterfactuals about  $o$  being F at  $t_{n+1}$  are different. This assumption is not problematic for my purposes of establishing that sensitivity and induction suffer from a heterogeneity problem. This result is also gained (or even strengthened) if we take further varying factors into account.

$o$  to constantly be F (or to constantly be not-F) from  $t_1$  to  $t_{n+1}$  than to be F from  $t_1$  to  $t_n$ , as in *Blackbird* and Scenario 1 of *T-shirt*, where Tim does not walk downtown. In Scenario 2, where Tim walks downtown after the hiking trail, the worlds where Tim changed his T-shirt downtown are closer than the worlds where he was not wearing a red T-shirt on the hiking trail. Here, it is more crucial for  $o$  to be F from  $t_1$  to  $t_n$  than it is to be constantly F or constantly not be F from  $t_1$  to  $t_{n+1}$ . Consequently, Sarah's belief is insensitive. In contrast, if counterfactuals can only be non-backtracking, then we can only consider possible worlds where  $o$  changed property F from  $t_n$  to  $t_{n+1}$ . In this case, beliefs formed via temporal induction are always insensitive. Thus, despite the heterogeneity of the overall results, at least we can say that temporal induction always yields insensitive beliefs if counterfactuals can only be non-backtracking.

We obtain slightly different results concerning *enumerative* induction. If the nearest possible worlds where  $o_{n+1}$  does not have property F are such that  $o_1$ - $o_n$  does not have property F, then S's belief that  $o_{n+1}$  is F formed via observation of  $o_1$ - $o_n$  and enumerative induction is presumably sensitive.<sup>22</sup> This condition is fulfilled if it is rather accidental that  $o_{n+1}$  is F but characteristic for  $o_1$ - $o_{n+1}$  that they have the same status of being F (or not being F), as in *Examiner*. However, if the nearest possible worlds where  $o_{n+1}$  does not have property F are such that  $o_1$ - $o_n$  still *has* property F, then S's belief formed via observation and induction is insensitive. This holds for *Raven*.

Notably, theories of counterfactuals that allow for non-backtracking counterfactuals and theories that do not deliver the same results for each individual case of enumerative induction, i.e. both types of theories imply that the target belief is sensitive or both theory types imply that it is insensitive. Perhaps we can construct cases of enumerative induction such that backtracking and non-backtracking theories deliver different results with respect to sensitivity, but I suspect that these instances of induction also involve a temporal element.<sup>23</sup>

<sup>22</sup> Again, I assume here that observation is sensitive with respect to  $o_1$ - $o_n$  being F.

<sup>23</sup> In the description of the cases certain details are made salient, e.g. habits of the target person. These details determine which facts are kept fixed and which facts differ in the nearest possible worlds considered. One might think that this leads to a contextualist sensitivity account in that different facts are salient to the knowledge attributor (in this case the reader) in different contexts and this salience determines which cases the attributor takes into account and whether her sensitivity attribution is true or false. However, this is not the way the alternative cases have to be understood. In *T-shirt*, we do not have alternative descriptions of one case leading to alternative judgements about whether Sarah's belief about Tim's T-shirt is sensitive. Rather, there

### 3 Heterogeneity: The problem for sensitivity and induction

Let me now diagnose what I regard as the real problem of sensitivity and induction. There is at least a tendency among proponents and critics of sensitivity that there is a homogeneous picture of sensitivity and induction. Critics of sensitivity accounts of knowledge, but also some adherents such as Becker, tend to think that induction yields insensitive beliefs whereas Wallbridge suggests that it yields sensitive beliefs. However, none of these opposing views is correct, since some instances of induction yield sensitive beliefs whereas some others yield insensitive ones.

I developed various cases of enumerative and temporal induction. In each of these cases, the subjects make an empirical observation and draw an inductive inference. Importantly, we intuitively judge that the subjects in these cases are in equally good epistemic positions. This view about the equality of the epistemic positions is also supported when applying plausible parameters for induction. Let me briefly explain. The epistemic force of induction comes in degrees. The epistemic strength of inductive reasoning from cases  $c_1$ - $c_n$  to case  $c_{n+1}$  (or from time interval  $i$  to point in time  $t$ ) and whether it can yield justification and knowledge intuitively depends on various factors. The strength of induction depends first on the number  $n$  of cases observed (or on the length of the observed time interval). All else being equal, the larger  $n$  is, the greater the epistemic strength of a particular induction; Secondly, the epistemic strength of induction varies with the relevant similarity between the cases observed and the case induced (or on the relevant similarity between the observed time interval and the point in time induced).<sup>24</sup> The more similar  $c_{n+1}$  is to  $c_1$ - $c_n$  in the relevant sense, the stronger the inductive reasoning is. Third, the epistemic strength of inductive reasoning depends on whether there exists a defeater  $d$  for the inductive conclusion, either rebutting or undercutting, such that  $S$  is propositionally justified in believing  $d$  and this justification undermines  $S$ 's justification in holding an inductive belief about  $c_{n+1}$ . Finally, the predicate involved has to be projectible.

---

are different cases whose constitution determines which possible worlds we have to consider and whether the target beliefs are sensitive. Notably, DeRose (1995) defends sensitivity-based contextualism about "knows" where in some contexts, sensitivity is required for knowledge but in others, it is not. However, he does not develop a contextualist account of sensitivity itself.

<sup>24</sup> It is a non-trivial task to determine the relevant similarity between the cases observed and the case induced, but plausibly the cases discussed can be set up in a way that the criteria for relevant similarity are to the same extent fulfilled. This is sufficient for the purposes of the paper.

Moreover, the amount of justification for the conclusion of an induction is also affected by the strength of the justification for believing the premises. All else being equal, the stronger the justification for believing the premises, the stronger the justification for believing the inductive conclusion.

This paper aims at showing that sensitivity accounts of knowledge have highly implausible consequences when it comes to inductive knowledge. In order to make this point, it suffices to refer to intuitively plausible criteria for inductive knowledge. We need not develop a detailed theory of induction and confirmation, involving Bayesianism or alternative conceptions.<sup>25</sup> For the purposes of this paper, it suffices to accept that in the cases discussed, the inductive reasoning is intuitively of the same epistemic strength according to the plausible parameters specified. That means that the subjects in the cases have equally good evidence about an equally high number of  $n$  cases (or a sufficiently long time interval  $i$ ), case  $c_{n+1}$  is equally similar to the observed cases  $c_1$ - $c_n$ , there is no defeater  $d$  for  $S$ , rebutting or undercutting, such that  $S$  is justified to believe that  $d$  and this justification undermines her inductive justification, and the predicates involved are equally projectible.<sup>26</sup>

Since the subjects are intuitively all in equally good epistemic positions, the minimal standards that a theory of knowledge has to fulfill is that it delivers the same outcome with respect to knowledge in all cases discussed. Here there are two options, first, that the subjects know in all cases of induction presented and, second, that they are precluded from knowing in all cases.<sup>27</sup> I assume that there is a wide agreement among epistemologists that we can have knowledge via induction. Accordingly, the first, positive option is far more popular than the second, negative one. However, sensitivity accounts of knowledge cannot deliver any of these two uniform pictures.

Let me explain in more detail. I regard it as an open question whether counterfactual conditionals can only be correctly interpreted as backtracking or

---

25 For an overview of theories of confirmation, induction, and Bayesianism, see Crupi (2020).

26 *Raven*, *Blackbird*, and *T-shirt* involve ordinary color predicates whereas *Examiner* involves the more superficial property of getting a particular grade. However, *Examiner* could be reformulated as a case where a subject throws a dice to determine which color a certain set of objects should be or a group of persons should wear. Moreover, the kinds of objects in the discussed cases are of different types, *Raven* and *Blackbird* involve natural kinds whereas *T-shirt* and *Examiner* do not. However, I do not see any reason why induction should not be applicable to different types of objects.

27 Knowledge can be based on inductive reasoning and inductive justification of different strengths. There might exist a threshold that these inductive strengths must exceed to be able to constitute knowledge, but determining such a threshold is not crucial for the purpose of this paper.

also as non-backtracking. However, in any case we acquire an unsatisfactorily heterogeneous picture. Suppose first that counterfactuals can only be non-backtracking. Presumably, temporal induction always provides insensitive beliefs, given a non-backtracking analysis of counterfactuals as in *Blackbird* and *T-shirt*. However, some instances of enumerative induction can yield sensitive beliefs, e.g. *Examiner*, but some others not, e.g. *Raven*. Thus, according to a sensitivity account of knowledge, S does not know in *Raven*, *Blackbird*, and *T-shirt* but knows in *Examiner*, given that counterfactuals can only be non-backtracking.

This result is counterintuitive since the epistemic position of the subject is intuitively equally good in all four cases. Suppose now that counterfactuals can be backtracking. In this case, *Examiner*, *Blackbird*, and Scenario 1 of *T-shirt* yield sensitive beliefs, but *Raven* and Scenario 2 of *T-shirt* yield insensitive beliefs. Again, sensitivity accounts of knowledge are committed to accept that the subjects know in the first three cases but not in the latter two.

Thus, in both cases of backtracking and non-backtracking theories of counterfactuals, some processes of induction yield sensitive beliefs but some others insensitive beliefs. Hence, sensitivity accounts of knowledge deliver in both cases an implausibly heterogeneous picture of inductive knowledge.<sup>28</sup> In both cases, we know via some instances of induction but do not know via some other instances. This heterogeneous picture is no less problematic than the orthodox view, dominant so far, that sensitivity accounts of knowledge preclude us from any kind of inductive knowledge.

These results affect extant pessimistic and optimistic accounts of sensitivity in various ways. The orthodox view about sensitivity and induction is based on cases of insensitive inductive beliefs that plausibly constitute knowledge, as presented by Vogel (1987, 1999) and Sosa (1999). The popular generalization of these cases has it that any instance of induction yields insensitive beliefs and that we cannot have any inductive knowledge according to sensitivity accounts of knowledge. This generalization is incorrect. However, Vogel and Sosa mainly aim at arguing against sensitivity accounts of knowledge by

---

<sup>28</sup> We can directly derive the heterogeneity of knowledge from the heterogeneity of sensitivity only if sensitivity is not only necessary but also sufficient for knowledge. However, various sensitivity accounts of knowledge, for example, those of Nozick (1981) and Becker (2007), assume that sensitivity is only necessary. These accounts defend further conditions such as adherence (Nozick) or reliability (Becker), but these conditions are fulfilled by induction. Hence, inductive knowledge is determined by the sensitivity of induction. These accounts are thus also committed to accepting the heterogeneity of inductive knowledge.

presenting counterexamples of insensitive knowledge via induction. This goal can still be reached by pointing out that sensitivity accounts of knowledge imply that we do not know in *some* (paradigmatic) instances of induction that plausibly yield knowledge.

Becker (2007) accepts that induction yields insensitive beliefs but argues that this does not pose a serious problem since we can still acquire knowledge about the probability of the target proposition.<sup>29</sup> This is already problematic since knowledge via induction seems highly plausible. Becker suggests that any instance of induction provides insensitive beliefs. What he should say is that in some cases of induction, we have knowledge of the target proposition, but in some very similar cases, we only have knowledge about the probability of the target proposition. This outcome is too heterogeneous to be plausible and, thus, not more convincing than Becker's original conclusion.

Wallbridge (2018) claims that inductive knowledge is sensitive, given that we accept backtracking counterfactuals in some contexts, or at least he leaves the reader with the challenge of presenting cases of insensitive induction. He suggests that we should not exclude backtracking counterfactuals in evaluating modal knowledge conditions like sensitivity or safety. In this respect, Wallbridge's analysis advances the existing debate about sensitivity and induction. However, he does not tell the whole story about sensitivity and induction since his challenge of finding instances of insensitive induction can be easily met. Moreover, sensitivity is not a matter of backtracking or non-backtracking interpretations of counterfactuals, as he suggests, since there are cases of sensitive induction and cases of insensitive induction for backtracking and non-backtracking interpretations.

Thus, sensitivity accounts of knowledge do not face the problem of precluding us from any inductive knowledge, as the orthodox view suggests, nor is it true that induction typically provides sensitive beliefs, as Wallbridge argues. Rather, some processes of induction yield sensitive beliefs whereas some very similar processes yield insensitive beliefs. Given this heterogeneous outcome, I do not see how a sensitivity account of knowledge can plausibly integrate a theory of inductive knowledge.

At this point, adherents of sensitivity might stick to their guns and claim that the acquired results about inductive knowledge are correct, since a sensitivity account of knowledge is correct, even though these results seem implausible at first sight. Nozick (1981) himself frequently endorses a similar line of

---

29 For a discussion, see Roush (2005, 66) who defends a similar view as Becker.

argumentation, as when he argues that knowledge does not transmit via conjunction elimination, a principle that is highly plausible.<sup>30</sup> However, such lines of argumentation are usually regarded as a vice of Nozick's account rather than virtue. Even adherents of sensitivity usually do not choose this strategy when defending sensitivity accounts of knowledge. For example, DeRose (1995) and Roush (2005) develop sensitivity accounts that avoid Nozick's implausible consequences of closure failure and Adams and Clarke (2005) defend Nozick's account against Kripke's (2011) objection by arguing that in Kripke's particular case knowledge closure is not violated. None of these defenses of sensitivity simply claim that the *reductio* arguments against sensitivity accounts fail because their highly counterintuitive consequences are the correct ones. This strategy is not more plausible in the case of induction.

The state of the discussion about sensitivity and induction has evolved as follows. Sosa and Vogel started the discussion by arguing that sensitivity precludes us from any kind of inductive knowledge, or at least from paradigmatic instances of inductive knowledge. Wallbridge objected that inductive beliefs are typically sensitive, providing a rejoinder to the cases presented by Sosa and Vogel. We have seen that neither of these positions is correct, pointing out instead that the relationship between sensitivity and induction is actually quite heterogeneous.

Interestingly, this development resembles the development of the discussion concerning sensitivity and higher-level knowledge, another purported challenge to sensitivity accounts of knowledge. Sosa (1999) and Vogel (2000) pointed out that one's beliefs that one does not falsely believe that *p* are insensitive. From this, Vogel concludes that sensitivity accounts of knowledge preclude us from any kind of higher-level knowledge while Sosa argues that this fact leads to implausible instances of closure failure since one can know that *p* without knowing that one does not falsely believe that *p*. Becker (2007) and Salerno (2010) respond to these concerns, pointing out that beliefs in weaker propositions with the formal structure  $\neg(B(p) \wedge \neg p)$  are insensitive but beliefs in the stronger propositions with the formal structure  $B(p) \wedge p$  or  $B(p) \wedge \neg \neg p$  can be sensitive. They conclude that we can have the relevant kind of higher-level knowledge according to sensitivity accounts. In Melchior (2015), I argue that the outcome that we know stronger higher-level propositions but fail to know weaker higher-level propositions is too heterogeneous to be

---

30 Hawthorne (2005) calls knowledge by conjunction elimination "incredibly plausible."

plausible, calling this the heterogeneity problem for sensitivity accounts.<sup>31</sup> Sensitivity does not preclude us from all inductive knowledge, nor does every instance of induction yield sensitive beliefs. In fact, some instances of induction yield sensitive beliefs, but very similar processes of induction lead to insensitive beliefs. We face a further instance of the heterogeneity problem for sensitivity accounts of knowledge when it comes to sensitivity and induction. This supports the view that heterogeneity, along different dimensions, is a characteristic feature of sensitivity and a more systematic problem for sensitivity accounts of knowledge.<sup>32</sup>

## 4 Conclusion

The orthodox view about sensitivity and induction has it that induction always delivers insensitive beliefs. Critics conclude that sensitivity accounts of knowledge are mistaken. Adherents of sensitivity accounts also assume that induction is homogeneous with respect to sensitivity. Becker accepts that any instance of induction is insensitive but argues that we still can have knowledge about the probability of the target proposition via induction. Wallbridge, in contrast, claims that induction yields sensitive beliefs. A careful analysis reveals more differentiated results. Some instances of induction yield sensitive beliefs but some instances in the neighborhood yield insensitive ones, regardless of whether we interpret counterfactuals as backtracking or non-backtracking. Sensitivity accounts of knowledge must, therefore, accept that we can know in some instances of induction but in very similar ones we cannot, although the epistemic situations of the believing subjects are intuitively equally good. These results are too heterogeneous to provide a plausible picture of inductive knowledge in terms of sensitivity.\*


---

31 For an objection to the heterogeneity problem, see Wallbridge (2017), and for a response, see Melchior (2017). For a related generality problem for higher-level knowledge, see Melchior (2014). For solutions to the heterogeneity problem, see Zalabardo (2016) and Bjerring and Gundersen (2020).

32 In Melchior (2019), I develop a sensitivity account of checking, arguing that sensitivity is necessary for checking while it is plausibly not necessary for knowing. I defend this view by showing that the proposed sensitivity account of *checking* is not equally affected by problems of sensitivity and induction as sensitivity accounts of knowing.

\* An earlier version of this paper was presented at the 2018 Joint Session of the Aristotelian Society and Mind Associate Conference in Oxford. I am thankful to the audience for their suggestions, to Martina Fürst for insightful discussions, and to Wes Siscoe and three anonymous referees for



Guido Melchior  
 0000-0001-6494-560X  
 University of Graz  
 guido.melchior@uni-graz.at

## References

- ADAMS, Frederick and CLARKE, Murray. 2005. "Resurrecting the Tracking Theories." *Australasian Journal of Philosophy* 83(2): 207–221, doi:10.1080/00048400500111030.
- BAUMANN, Peter. 2012. "Nozick's Defense of Closure." in *The Sensitivity Principle in Epistemology*, edited by Kelly BECKER and Tim BLACK, pp. 11–27. Cambridge: Cambridge University Press, doi:10.1017/CBO9780511783630.
- BECKER, Kelly. 2007. *Epistemology Modalized*. Routledge Studies in Contemporary Philosophy n. 5. London: Routledge.
- BECKER, Kelly and BLACK, Tim, eds. 2012. *The Sensitivity Principle in Epistemology*. Cambridge: Cambridge University Press, doi:10.1017/CBO9780511783630.
- BJERRING, Jens Christian and GUNDERSEN, Lars Bo. 2020. "Higher-Order Knowledge and Sensitivity." *Canadian Journal of Philosophy* 50(3): 339–349, doi:10.1017/can.2019.36.
- COGBURN, Jon and ROLAND, Jeffrey W. 2013. "Safety and The True-True Problem." *Pacific Philosophical Quarterly* 94(2): 246–267, doi:10.1111/j.1468-0114.2012.01454.x.
- CRUPI, Vincenzo. 2020. "Confirmation." in *The Stanford Encyclopedia of Philosophy*. Stanford, California: The Metaphysics Research Lab, Center for the Study of Language; Information. Revision, January 28, 2020, of the version of May 30, 2013, <https://plato.stanford.edu/entries/confirmation/>.
- DEROSE, Keith. 1995. "Solving the Skeptical Problem." *The Philosophical Review* 104(1): 1–52, doi:10.2307/2186011.
- . 2004. "Sosa, Safety, Sensitivity, and Skeptical Hypotheses." in *Ernest Sosa and His Critics*, edited by John GRECO, pp. 22–41. Philosophers and Their Critics. Oxford: Basil Blackwell Publishers, doi:10.1002/9780470756140.
- . 2017. *The Appearance of Ignorance. Knowledge, Skepticism, and Context, Vol. II*. Oxford: Oxford University Press.
- ENOCH, David, SPECTRE, Levi and FISHER, Talia. 2012. "Statistical Evidence, Sensitivity, and the Legal Value of Knowledge." *Philosophy & Public Affairs* 40(3): 197–224, doi:10.1111/papa.12000.
- HAWTHORNE, John. 2004. *Knowledge and Lotteries*. Oxford: Oxford University Press.

---

this journal for their helpful comments on this paper. The research was funded by the Austrian Science Fund (FWF): P 33710.

- . 2005. “The Case for Closure.” in *Contemporary Debates in Epistemology*, edited by Ernest Sosa and Matthias Steup, 1st ed., pp. 50–81. Contemporary Debates in Philosophy n. 3. Malden, Massachusetts: Basil Blackwell Publishers. Second edition: Steup, Turri and Sosa (2014, 40–55).
- KHOO, Justin. 2017. “Backtracking Counterfactuals Revisited.” *Mind* 126(503): 841–910, doi:10.1093/mind/fzw005.
- KRIPKE, Saul A. 2011. “Nozick on Knowledge.” in *Philosophical Troubles*, pp. 162–224. Collected Papers n. 1. Oxford: Oxford University Press.
- LEWIS, David. 1973. *Counterfactuals*. Cambridge, Massachusetts: Harvard University Press. Cited after republication as Lewis (2001).
- . 2001. *Counterfactuals*. Oxford: Basil Blackwell Publishers. Republication of Lewis (1973).
- LUPER, Steven. 1984. “The Epistemic Predicament: Knowledge, Nozickian Tracking, and Scepticism.” *Australasian Journal of Philosophy* 62(1): 26–49. Published under the name “Luper-Foy”, doi:10.1080/00048408412341241.
- MCGLYNN, Aidan. 2012. “The Problem of True-True Counterfactuals.” *Analysis* 72(2): 276–285, doi:10.1093/analys/anso46.
- MELCHIOR, Guido. 2014. “A Generality Problem for Bootstrapping and Sensitivity.” *Croatian Journal of Philosophy* 14(40): 31–47.
- . 2015. “The Heterogeneity Problem for Sensitivity Accounts.” *Episteme* 12(4): 479–496, doi:10.1017/epi.2015.31.
- . 2017. “Sensitivity has Multiple Heterogeneity Problems: a Reply to Wallbridge (2017).” *Philosophia* 45(4): 1741–1747, doi:10.1007/s11406-017-9873-5.
- . 2019. *Knowing and Checking: An Epistemological Investigation*. New York: Routledge.
- . 2020. “Sensitivity Principle in Epistemology.” *Oxford Bibliographies Online*, doi:10.1093/obo/9780195396577-0404.
- . 2021. “A Modal Theory of Discrimination.” *Synthese* 198(11): 10661–10684, doi:10.1007/s11229-020-02747-4.
- MENZIES, Peter. 2014. “Counterfactual Theories of Causation.” in *The Stanford Encyclopedia of Philosophy*. Stanford, California: The Metaphysics Research Lab, Center for the Study of Language; Information. Revision, February 10, 2014, of the version of January 10, 2001, <https://plato.stanford.edu/archives/spr2014/entries/causation-counterfactual/>.
- NOZICK, Robert. 1981. *Philosophical Explanations*. Cambridge, Massachusetts: Harvard University Press.
- PRITCHARD, Duncan. 2012. “In Defence of Modest Anti-Luck Epistemology.” in *The Sensitivity Principle in Epistemology*, edited by Kelly BECKER and Tim BLACK, pp. 173–192. Cambridge: Cambridge University Press, doi:10.1017/CBO9780511783630.

- ROUSH, Sherrilyn. 2005. *Tracking Truth. Knowledge, Evidence, and Science*. Oxford: Oxford University Press, doi:10.1093/0199274738.001.0001.
- SALERNO, Joseph [Joe]. 2010. "Truth Tracking and the Problem of Reflective Knowledge." in *Knowledge and Skepticism*, edited by Joseph Keim CAMPBELL, Michael O'ROURKE, and Harry S. SILVERSTEIN, pp. 73–84. Topics in Contemporary Philosophy n. 5. Cambridge, Massachusetts: The MIT Press, doi:10.7551/mitpress/9780262014083.003.0004.
- SOSA, Ernest. 1999. "How to Defeat Opposition to Moore." in *Philosophical Perspectives 13: Epistemology*, edited by James E. TOMBERLIN, pp. 141–153. Oxford: Basil Blackwell Publishers, doi:10.1111/0029-4624.33.S13.7.
- SOSA, Ernest and STEUP, Matthias, eds. 2005. *Contemporary Debates in Epistemology*. 1st ed. Contemporary Debates in Philosophy n. 3. Malden, Massachusetts: Basil Blackwell Publishers. Second edition: Steup, Turri and Sosa (2014).
- STARR, William B. 2019. "Counterfactuals." in *The Stanford Encyclopedia of Philosophy*. Stanford, California: The Metaphysics Research Lab, Center for the Study of Language; Information. Version of January 18, 2019, <https://plato.stanford.edu/entries/counterfactuals/>.
- STEUP, Matthias, TURRI, John and SOSA, Ernest, eds. 2014. *Contemporary Debates in Epistemology*. 2nd ed. Contemporary Debates in Philosophy n. 3. Oxford: Wiley-Blackwell. First edition: Sosa and Steup (2005).
- VOGEL, Jonathan. 1987. "Tracking, Closure, and Inductive Knowledge." in *The Possibility of Knowledge. Nozick and His Critics*, edited by Steven LUPER, pp. 197–217. Totowa, New Jersey: Rowman & Littlefield Publishers. Published under the name "Luper-Foy" .
- . 1999. "The New Relevant Alternatives Theory." in *Philosophical Perspectives 13: Epistemology*, edited by James E. TOMBERLIN, pp. 155–180. Oxford: Basil Blackwell Publishers, doi:10.1111/0029-4624.33.S13.8.
- . 2000. "Reliabilism Leveled ." *The Journal of Philosophy* 97(11): 602–623, doi:10.2307/2678454.
- WALLBRIDGE, Kevin. 2017. "Sensitivity hasn't Got a Heterogeneity Problem: a Reply to Melchior (2015)." *Philosophia* 45(2): 835–841, doi:10.1007/s11406-016-9782-z.
- . 2018. "Sensitivity, Induction, and Miracles." *Australasian Journal of Philosophy* 96(1): 118–126, doi:10.1080/00048402.2017.1328697.
- WALTERS, Lee. 2016. "Possible World Semantics and True-True Counterfactuals." *Pacific Philosophical Quarterly* 97(3): 332–346, doi:10.1111/papq.12067.
- WILLIAMSON, Timothy. 2000. *Knowledge and Its Limits*. Oxford: Oxford University Press, doi:10.1093/019925656X.001.0001.
- ZALABARDO, José L. 2016. "Reflective Knowledge and the Nature of Truth." *Disputatio* 8(43): 147–171, doi:10.2478/disp-2016-0009.

# Review of Imhof (2014)

ULRICH SCHWABE

Fichte's philosophy is still among the darkest of the German-language tradition. One approach to understanding it is through Kant's *Critique of Pure Reason* and the discussion that followed it. In his study, *Der Grund der Subjektivität*, Silvan Imhof follows this path—and with resounding success. Imhof shows how Fichte's *Wissenschaftslehre* can be understood as the preliminary endpoint of a discourse that was essentially concerned with overcoming skeptical arguments. Already Kant's *Critique of Pure Reason* can be understood in that way. In particular, in the *Transcendental Deduction of the Pure Concepts of the Understanding*, Kant tries to fend off Hume's attacks on the legitimacy of central concepts such as causality and substantiality. Kant's attempt, however, remained inadequate according to the diagnosis of some contemporaries such as S. Maimon and G.E. Schulze. This motivated first Reinhold, and then Fichte to search for a foundation of philosophy that is in fact indubitable. Fichte's *Wissenschaftslehre* can thus be understood, according to Imhof's central thesis, as an attempt to overcome skeptical objections from the post-Kantian discussion.

Imhof substantiates this thesis by tracing the discourse leading from Kant's *Transcendental Deduction of the Pure Concepts of the Understanding* (=“TD”) to Fichte's *Grundlage der gesamten Wissenschaftslehre* (=“GL”). This discourse is curiously shaped by a misunderstanding that Maimon commits in his interpretation of Kant's TD. Namely, as Imhof demonstrates, Maimon believes that Kant's TD takes its beginning from the fact of experience, without further substantiating it. It is this dogmatic assumption of the fact of experience that Maimon criticizes, claiming that such a fact is not a suitable starting point for a TD, because it is dubitable. However, according to Imhof, this criticism misses Kant's point, since his TD does not start from the fact of experience, but rather from subjects having ideas (= “*Haben von Vorstellungen*”).

Although Maimon's skeptical critique is based on a misunderstanding, Reinhold is so impressed by it that he sets out to find a new basis for a transcendental deduction. He finds one in his theorem of consciousness, according to which in consciousness the ideas are related to subject and object and are

distinguished from both. But even this theorem is not suitable as a basis for a transcendental deduction, as first Schulze and then Fichte note. For it expresses a mere *fact*. But facts can always be doubted. Thus, Reinhold's search for an indubitable basis also fails.

Fichte's GL now begins at this point of discussion. Its fundamental insight, according to Imhof, is that because every fact can be doubted, a skepticism-resistant philosophy must be built on something other than a fact. Fichte finds this other in the self-positing ("*Selbstsetzung*") of the I, which he conceives as "*Tathandlung*" ("fact-action" or "(f)act"). Fichte claims in Imhof's reconstruction that this *Tathandlung* has a special character, which makes it indubitable.

With the conception of the *Tathandlung*, Fichte overcomes the weaknesses of Reinhold's attempt at a foundation of philosophy. But how does Fichte's foundation of philosophy relate to Kant's TD? Since Reinhold's attempt, and with it its improvement by Fichte, is based on Maimon's wrong understanding of Kant, the question arises whether Fichte improves Kant's TD at all if properly understood. At first glance, this is not the case. For, Imhof argues, Kant already succeeded in building his TD on a foundation resistant to skepticism, namely on the mere having of ideas, which is expressed in Kant's phrase of the *I think that must be able to accompany all my ideas*. The indubitability of the mere having of ideas is conceded at least by skeptics like Hume and Schulze. Thus, it seems at first as if Kant's TD already stands on a secure foundation and therefore needs no improvement by Fichte.

But this view is wrong according to Imhof. For although Kant's TD starts from an indubitable foundation, it still fails. The reason for this is that Kant elaborates this foundation incorrectly: The having of ideas refers to a subject. Therefore, an adequate conception of the having of ideas requires an accurate theory of subjectivity. However, Kant does not provide such a theory. Its development is hindered by Kant's dogma of the strict separation of sensibility and understanding. This dogma prevents Kant from conceptualizing subjectivity by means of the figure of intellectual intuition. Because Kant refuses to resort to that figure, he arrives at an inconsistent characterization of subjectivity: on the one hand, the subject is conceived as a transcendental entity, which cannot appear sensually and therefore cannot be an object of insight. On the other hand, Kant speaks of an empirical subject that appears in the inner sense. According to Imhof, what one has to do with the other remains completely unclear in Kant.

Fichte overcomes this conceptual weakness by characterizing the subject through the figure of intellectual intuition. Since the subject is the basis both of understanding and of sensibility, it seems obvious that it must be the union of these two faculties, and this union is nothing other than intellectual intuition. Fichte does not yet express this thought in the GL, but all the more emphatically in his “*Versuch einer neuen Darstellung der Wissenschaftslehre*” of 1797.

With this revised understanding of the subject, Fichte now succeeds, according to Imhof, in actually finding a basis for philosophy that is resistant to skepticism. The further progress of the *Wissenschaftslehre* then consists in nothing else than deriving a multitude of central concepts of philosophy from this basis.

By tracing the philosophical development from Hume through Kant, Maimon, Reinhold, and Schulze to Fichte, Imhof provides the reader with a historical approach to Fichte’s *Wissenschaftslehre*. However, in order to make it plausible that Fichte’s philosophy can indeed be seen as overcoming its historical predecessors, Imhof also offers systematic reconstructions of central pieces of Kantian and Fichtean philosophy. It is primarily these systematically oriented exegetical offerings that make Imhof’s study so particularly valuable. One such offer is Imhof’s proposal to understand Fichte’s self-positing of the *I* centrally as an *intentional act*, and thus to view the *Wissenschaftslehre* as a *theory of intentionality*. Another innovative idea in Imhof’s reconstructions is the way in which Imhof derives the indubitability of the *Tathandlung* from an interpretation of positing. Imhof understands positing—following Strawson—as the presupposition of the subject in a sentence. Such a presupposition is found in every ordinary statement of the form “*x* is an *F*”. In such a statement it is presupposed that there exists an *x* to which *F* can be attributed. The statement itself, however, cannot ensure the fulfillment of this presupposition. Imhof now understands the self-positing of the *I* as that specific presupposition that guarantees its own fulfillment. Thus, because in the self-positing of the *I* a proposition is established whose content is real by virtue of merely being thought, the self-positing is indubitable.

Imhof’s project is highly ambitious in both historical and systematical terms. Historically, by discussing Kant’s TD and Fichte’s GL Imhof treats two of the most difficult texts that the philosophical tradition has to offer. Systematically, Imhof not only reconstructs the central ideas of these texts, but also examines their validity. What is admirable is Imhof’s argumentative concentration: he does not lose himself in out-of-the-way exegetical battles, but consciously

highlights core elements of Kant's and Fichte's argumentations. As a result, Imhof's reflections achieve exemplary clarity and stringency.

It is in the nature of such an ambitious undertaking that it is not immune to further inquiries. One question concerns Imhof's thesis that Fichte in his *Wissenschaftslehre* is essentially concerned with overcoming skepticism. This thesis would suggest itself if Fichte succeeded in such an overcoming. Imhof argues that he does, since the *Tathandlung* is a skeptic-resistant basis of philosophy. However, Imhof does not make it fully clear why this should be the case: first, it remains unclear in Imhof's reconstruction whether the foundation of the *Wissenschaftslehre* is supposed to be the *performance* of the *Tathandlung*, or the *description* of such performance. In the first case, the assumption of immunity to skepticism might be convincing, because only statements or propositions, but not actions, can be doubted. Usually, however, philosophical systems are regarded as networks of propositions. But only the description, not the performance, of an action can be a proposition. This suggests that the *Wissenschaftslehre* starts with the *description* of the *Tathandlung*. But then it is hard to see why such a description should not be exposed to skeptical objections in the same way as the description of a fact. For example, it can be doubted that anything at all corresponds to Fichte's descriptions of the *Tathandlung*. Such a doubt even suggests itself in certain respects, for these descriptions contradict familiar patterns of thought to such an extent that it is difficult to imagine how anything could correspond to them. Thus, despite Imhof's interpretations, it remains unclear whether Fichte actually succeeds in finding a skeptic-resistant foundation of philosophy.

If this is uncertain, the question arises as to whether Fichte actually aimed for immunity to skepticism with the vigor that Imhof attributes to him. Imhof's interpretation may well draw on relevant quotations from Fichte. However, there are also passages in Fichte's works that point in a different direction. In the introduction to the *Wissenschaftslehre nova methodo*, for example, Fichte points out that, at least in real life, no human being ever doubts the reality of the external world. Nevertheless, a foundation of the external world is necessary. But not because skepticism poses a serious threat to our belief in the external world. It is necessary solely because with skepticism our thinking is in danger of colliding with the obvious fact of the reality of the external world. Such a collision would be a scandal to reason. Thus, a foundation of our belief in the external world is necessary to save confidence in our reason, but not to save our belief in the external world. The goal of such

a foundation, then, is to reconcile reason with itself, not to escape skeptical threats.

Accordingly, it is not entirely evident that Fichte's conception of subjectivity springs essentially from the attempt to meet skepticism. Imhof believes that Fichte's conception of subjectivity escapes skeptical objections to which, among others, Reinhold's system fell victim. These objections point out that even if something must necessarily be thought, it need not be real. According to Imhof, the essential point of Fichte's conception of subjectivity is that it refutes this objection. It does so by conceiving of subjectivity as something that is necessarily real if it is only thought.

However, Fichte is not forced to respond to such skeptical objections with a special conception of subjectivity. Rather, he can assert more broadly against these objections that they are based on false presuppositions: they presuppose that something could not exist even though it must be regarded as existent by necessity of thought. But this presupposition could be wrong because reality could be in its essence nothing else than a certain form of necessity of thought. This is exactly how Fichte conceives reality when he reconstructs it as the boundedness ("*Gebundenheit*") of our thinking. The external world arises through the reification of this boundedness. Therefore, the objection that in reality nothing could correspond to something we necessarily have to assume fails to recognize that reality is only a reification of what we are bound to think. Since this is so, everything that is contained in this boundedness must be real. A skepticism resulting from the assumption that necessary thoughts and reality could go different ways is to be met not by developing a special form of subjectivity, but by clarifying the misunderstanding that underlies it.

But such considerations show no more than that even Imhof cannot clear up all ambiguities with respect to Fichte's philosophy. What he does succeed in doing, however, is to make an extremely plausible and well-comprehensible offer for understanding Fichte's philosophical approach. And this alone is an achievement of inestimable value in the case of an author like Fichte.

Ulrich Schwabe  
ulrich.schwabe@uni-tuebingen.de

IMHOF, Silvan. 2014. *Der Grund der Subjektivität. Motive und Potenzial von Fichtes Ansatz*. Philosophica n. 15. Basel: Schwabe Verlag.





# Review of Lutz (2016)

TONI RØNNOW-RASMUSSEN

Ideally, this book will not go under the radar of metaethicists who wish to deepen their views on metaphysics. Nor, for that matter, should metaphysicians who want to develop their metaethics ignore it. Both are in for a treat. That said, its 300 pages or so could have been parcelled in a more reader-friendly way. The chapters are long (one runs to over 70 pages) and dense with information and argument, and there is also no index. This makes for a challenging reading experience. However, these are but pimples and blemishes. Lutz has written an otherwise impressive and captivating work. It will amply reward colleagues who are ready to roll up their sleeves and scrutinize new and familiar views on supervenience, grounding, and the in-virtue-of relation.

*Good In Virtue Of* has six chapters. The first mainly summarizes the author's objectives, the most central of which pertains to the following question: "What kind of relation is this 'in-virtue-of' or 'making' relation that holds between the instantiation of other properties and the instantiation of a certain normative property?" (Lutz 2016, 2).

The final chapter recapitulates Lutz's main conclusions. Chapters 2–5, then, are the real body of the work.

In Chapter 2 Lutz develops the intuition, familiar to moral philosophers, that evaluative properties do not, as she puts it, obtain "brutally". Much should be of interest here to metaethicists. For instance, Lutz's take on the formal features of the in-virtue-of relation is illuminating. I particularly enjoyed her discussion of Väyrynen's argument against the transitivity of the in-virtue-of relation (albeit I am not sure I fully agree with her). However, the chapter also reveals important scope restrictions and assumptions informing Lutz's discussion. Let me mention just two. First, in her attempt to understand the in-virtue-of relation she confines herself to evaluative (rather than normative) properties. Second, she assumes moral realism, because "if there are no evaluative properties at all, or if there are at least no instantiated evaluative properties at all, then the desired relation never obtains" (Lutz 2016, 16). Thus, we need, she thinks, to accept some version of moral realism, since otherwise

there would be no in-virtue-of relation in the first place. Is this right? It would seem so: “no relata no relation” is an important truth. However, things may be less straightforward. For instance, the issue depends on how one regards a certain kind of conditional fact. If there are these peculiar facts, and they obtain in virtue of some other facts, then, I think Lutz is mistaken.<sup>1</sup> Somewhat ironically, the required adjustment would have made her book more interesting to a wider readership. For then it would not only be realists, with their commitment to the idea that evaluative properties are instantiated, who would gain from learning more about her approach to the desired in-virtue-of relation. An example of the kind of fact I have in mind is that *some object is good on condition that goodness is instantiated*.<sup>2</sup> If there are such peculiar facts, which do not depend on, or require, if they are to obtain, that, goodness is ever actually instantiated, we might wonder: In virtue of what do these obtain? To answer that question, we need a grip on what this relation is all about, and I think Lutz’s book is an excellent starting point for this.

Chapter 3 is a penetrating inquiry into why supervenience cannot be the desired in-virtue-of relation. A small caveat is in place here. At the outset, I had some problems following Lutz’s setup. I suspect they were age-related. Long ago, I was trained to regard supervenience as something other than merely strict covariance between evaluative and natural properties.<sup>3</sup> My guess is that, for many philosophers of my generation, what we had in mind all the time was a relation of dependency—one we often expressed by employing the in-virtue-of idiom. Admittedly, much has been said about supervenience, some of which points in various different directions. Lutz is perfectly aware of this. She maintains, then, that there is an important line of thought which conceives of supervenience, precisely, as covariance. As she argues at length, the in-virtue-of relation cannot merely boil down to covariance. She points to several reasons for this, the most important being that the in-virtue-of relation is one of determination and metaphysical priority while the relation of covariance is not.

I can’t help wondering whether some of the issues relating to whether supervenience/grounding is the desired in-virtue-of relation may turn out to

- 
- 1 I should add here that according to Lutz, nothing much hinges on whether we talk about properties or facts (Lutz 2016, 147–148). I am inclined to agree. However, in light of my example in the main text, she might change her view on this.
  - 2 For more on this kind of fact, see Rønnow-Rasmussen (2016).
  - 3 Lutz is aware that not everyone in the past considered supervenience mere covariance. E.g. see Lutz (2016, 82).

be verbal in character. For instance, I believe that many of those in the past who were thinking of supervenience as covariance had a special kind of covariance in mind—namely, a one- and not a two-way direction of covariance. I suspect they would argue that their accounts express a kind of metaphysical priority. As Lutz herself points out, many philosophers recognize the phenomenon of multiple supervenience bases, and for those who accept this supervenience *qua* covariance will in effect be a one-direction variance: it will not be the case then that, necessarily, if something instantiates, say, goodness, it instantiates the natural property *N*. Is this a way of expressing determination of some sort? I know people who believe it is. Whether or not we agree with them, I am inclined to concur with Lutz that there is more to the in-virtue-of relation than covariance. At all events, this is an impressive chapter in which Lutz shows how well-versed she is in the relevant literature. In fact, she spreads metaphysical and metaethical insights better than a farmer spreads seed, and her illuminating comments and arguments make for a most worthwhile read. I thoroughly enjoyed the discussion.

On top of what I have already referred to, in an extensive excursus Lutz serves the reader a buffet of core realist metaethical views (and, not least, some of the challenges they face). You get the sense that you are in the hands of an excellent chef who carefully points out that, however appetizing these dishes may seem to you, they all contain ingredients that make them more or less difficult to digest. Lutz does not take a stand and state which realist view she endorses. Some might believe this to be a fault on her part. I had no problem with it, mostly because I found her discussion thorough and highly informative. Readers familiar with metaethics will also appreciate her decision not to beat around the bush, and to directly address the core issues. The facility and clarity with which she unwraps numerous complex issues made me envious. For instance, I suspect her discussion of the alleged identity of value properties and natural properties will be illuminating for many readers, whether they are metaethicists or metaphysicians. Should we, for instance, maintain that the instantiation of an evaluative property is token-identical with the instantiation of a natural property? Lutz is wary of such a position because she takes it to imply a trope theory of properties. She regards the trope theory as highly controversial (Lutz 2016, 100–101). Personally, I wish she hadn't set it aside quite as swiftly.

So-called “response theories” also come under Lutz's powerful lens. Again, without much ado, she quickly goes to the central problems and identifies the challenging questions. I'm not sure I always agree with her on the weight

she places on some of her worries. I also think there are responses to some of the challenges she raises that appear in work with which she is apparently unfamiliar. But this shouldn't detract from the fact that she provides a concise account of the main issues surrounding response theories. She categorizes different so-called Fitting-attitude (FA) analyses as response theories. I think this is misleading. She also believes that an FA analysis must identify the subject for whom it is fitting to favour something, and that this should be a matter of concern for the FA analyst. It is a worry shared by many. However, I am less troubled by it. Arguably, if you cast the FA account in terms of *pro tanto* reasons, there is no need to specify the subject when something is, say, admirable, if you think (real or possible) subjects who can respond to *pro tanto* reasons are capable of admiring something. I also think that attempts to understand goodness, period, as problematic unless it is understood as a kind of relational goodness-for-someone (or vice versa) rest on mistaken views on goodness, period, and goodness-for (see [Rønnow-Rasmussen 2021](#)).

Chapter 4 presents a long and detailed discussion packed with metaphysical minutiae of the notion of grounding. It asks whether we should apply the grounding framework to evaluative facts and eventually settle on grounding as the *in-virtue-of* relation.

So should we? Lutz does not aim to provide a definitive answer, but she argues that grounding is a plausible candidate as the *in-virtue-of* relation. Why? The swift answer is that it meets the following criteria: “it is a non-causal, metaphysical determination relation which imposes hierarchical structure on reality” ([Lutz 2016, 178](#)); it introduces metaphysical priority. However, Lutz is explicit that she is not giving a “full-blown defence of grounding in this chapter” ([Lutz 2016, 135](#)). As far as I can see, this defence is not provided in the remaining chapters.

But what is grounding? Lutz identifies two core notions in the literature. The one she is attracted to is (somewhat puzzlingly) not the one she adopts. She sticks with the less controversial variety, which identifies grounding with metaphysical explanation (rather than with what is identical to the “relation that backs metaphysical explanation” ([Lutz 2016, 144](#))). Despite its being widely employed nowadays, metaphysical explanation is, to say the least, far from being a transparent notion. For one thing, it is debatable whether reality contains explanation relations on its own—as opposed to there being merely people who offer explanations of things the success of which is conditional, in part, on the way the person who is given an explanation understands it. The last kind of explanation is an epistemic success notion. Attempts to

understand grounding in terms of metaphysical explanation are therefore challenging. Lutz's attempt to meet the challenge is certainly laudable.

Things heat up when Lutz begins to outline the formal properties of grounding. Besides agreeing that grounding is an asymmetric, transitive, and hyper-intensional relation, Lutz stands with those who conceive of grounding as an irreflexive relation: if  $x$  grounds  $y$ , then  $x$  and  $y$  are non-identical. If  $x$  and  $y$  are not identical, it seems that  $x$  cannot be reduced to  $y$  (or vice versa, for that matter). Or so Lutz thinks. However, as Gideon Rosen (2010) has argued, that is not an uncontroversial inference. For, briefly, if we conceive of reduction as a relation between facts, reducing one fact to another is not a matter of identifying the one with the other.

The idea that you cannot reduce something to that to which it is identical will, I suspect, appear plausible to some, perhaps many, of Lutz's readers. However, she maintains that it requires a fine-grading of facts (something she resists). For instance, consider the fact that  $ABCD$  is a square. On the fine-grained approach, it would not turn out to be identical with the fact that  $ABCD$  is an equilateral rectangle. Against this, Lutz (in line with Paul Audi's (2012) critique) suggests that the approach is committed to a "wordy" instead of "worldly" view of facts. She makes two important points in this connection. First, she rightly stresses that "grounding and reduction can only go together if one adopts a different conception of reduction to what we might call the identity conception" (Lutz 2016, 152). Second, she makes it clear (Lutz 2016, 153) that she is not ready to do the latter. She accepts an identity conception, and "hence grounding and reduction excludes each other" (Lutz 2016, 153). This is, of course, an important statement in the book. Unfortunately, she is not terribly forthcoming with the reasons for her choice, and so readers might feel shorth-changed at this point. In all fairness, the issue is a tricky one. However, because I believe we can explain value most successfully with a combination of worldly and wordy facts,<sup>4</sup> I am inclined to side against Lutz on this matter. On the other hand, given that what we are talking about here is "metaphysical explanation"—a notion the conceptual contours of which are still very much in need of clarification—I do wonder whether retaining an open mind on this matter would have been preferable.

Lutz is open-minded on other issues. For instance, there is an important discussion in the grounding literature of the idea that metaphysical necessity

---

<sup>4</sup> See Rabinowicz and Rønnow-Rasmussen (2021) (esp. the discussion of two kinds of fact on pp. 2479–2480).

is what distinguishes grounding from causal relations. If a fact,  $p$ , grounds  $q$ , is it the case that necessarily (metaphysically speaking) if  $p$ , then  $q$ ? While so-called necessitarians affirm this, contingentists deny it. This is a vexed issue, and if we are to make headway with it, as Lutz elegantly shows (Lutz 2016, 155–166), we will need a clearer picture of what we have in mind by “metaphysical explanation”. I find it easy to agree with Lutz, and therefore I think her openness on this issue is understandable.

Another question about grounding that Lutz addresses with care is whether it is non-monotonic. Non-monotonicity guarantees that grounds do not contain arbitrary facts. Whatever is in the ground is part of what makes the fact ground some other fact. This is an important feature, but it is also not obvious how we should understand it. In metaethics, for instance, it is common, following Dancy, to distinguish between a value’s resultance (base), which contains only those properties of the value bearer that make it valuable, and the supervenience base, which is understood as a larger base containing all the facts on which the value, in a broad sense, depends (Dancy 2004). The larger base may contain so-called enablers, which are facts (or features) that enable other facts (or features) to be value-making. Dancy typically takes enablers to be facts about the context in which the valuable object is located. However, as Rabinowicz and I have recently argued (Rabinowicz and Rønnow-Rasmussen 2021), enabling facts need not be facts of this sort. Consider, admirability. The explanation of this kind of value seems to require us to refer to those features that make the object admirable (valuable) and those that enable the properties of the value bearer to be value-making. In this particular case, these are in part the features that make an attitude one of admiration, and facts about these essential features are arguably “wordy” (conceptual) facts.

This discussion raises some fundamental meta-questions. To what extent should we allow our metaphysical views and intuitions to govern our value-taxonomic views? Should we perhaps adjust our metaphysics in light of our value notions? There are convincing arguments in the literature that not every final value is an intrinsic value (and even that not every final intrinsic value accrues to its bearers in virtue of features that necessarily belong to the value bearers). This suggests that the grounds of final extrinsic values may, in one sense of “arbitrary”, contain arbitrary facts. It is perhaps to ask too much of an already rich book, but it would have been enlightening to read about Lutz’s views on these meta-issues.

One of the many strengths of Chapter 4 is that it brings out the ways in which arguments in the metaphysical literature correspond to arguments that

are discussed by metaethicists, and vice versa. Thus, in her overview of the metaethical holism-atomism debate (Lutz 2016, 166–178), Lutz finds interesting similarities with metaphysical debates over necessitarianism. Perhaps she is even right in thinking that this resemblance is an indication that her quest for the in-virtue-of relation within the grounding framework is on the right track.

The grounding framework comes, she thinks, with an important extra advantage. Since there is so much grounding about (as it were) that does not concern the evaluative, or the normative, at all, realists are handed a response to Mackie's queerness objection—namely, that there is nothing queer about grounding. Lutz advances the following argument (Lutz 2016, 179):

- (P1) The in-virtue-of relation is a grounding relation
- (P2) The grounding relation is ubiquitous in our world; we know it from many other philosophical contexts, and hence it is not metaphysically queer.
- (C) *Pace* Mackie, the in-virtue-of relation is not a metaphysical queer relation.

Lutz identifies two related problems with her argument (Lutz 2016, 180). First, if there are different kinds of grounding, the worry is that grounding *qua* the in-virtue-of relation might still come out as queer in comparison with other kinds of grounding. For instance, we might follow Kit Fine and distinguish different kinds of grounding in terms of metaphysical, normative, and natural necessity (Fine 2012). So if value has a normative grounding (is grounded in normative necessity) it might be regarded as queer in comparison with metaphysical grounding. Lutz is not really worried, though. She is sceptical about enriching the notion of necessity beyond conceptual and metaphysical necessity. Whether or not we agree with her, there may yet be other kinds of grounding. For instance, if I have understood her properly, she takes grounding to be factive. That is, she assumes that it is impossible for grounding to be exemplified when the relata of the grounding relations are not facts. This is a reasonable view, but since grounding eventually comes down to metaphysical explanation, why couldn't there be such explanations when we consider abstract entities that do not obtain? (Admittedly, this would require some work on how best to understand relations).

The second difficulty Lutz raises centres on scepticism about the grounding relation in the first place. In effect, she identifies three kinds of scepticism:



one can be a sceptic about the primitiveness, and/or intelligibility, and/or usefulness, of grounding. However, after discussing these varieties, she assumes they can all be resisted. She reminds us that her aim is to show that using the framework of grounding leads to interesting insights in metaethics, and that it is not her intention to “establish and defend this framework” (Lutz 2016, 183). Interestingly, Lutz’s assumption that scepticism about grounding fails seems to backfire. Ultimately, isn’t the need for this assumption just an expression of scepticism? Perhaps I am wrong. However, some of the things that Lutz herself recognizes appear to open the door to a sceptical conclusion. While it is certainly conceivable that someone with strong “grounding intuitions” would reach (C), one can also easily imagine error theorists being unconvinced by the argument. From what I know about Mackie’s queerness argument, and in particular why he thought the nature of supervenience provides a strong incentive to be a sceptic about it, I would expect him to reason in the following way. We do not quite know what metaphysical explanation is (something that Lutz acknowledges), and therefore we do not quite know what grounding is. Hence we do not quite know that (P<sub>1</sub>) is correct, and for this reason, we do not quite know that (P<sub>2</sub>) is correct. It would therefore be a mistake not to be sceptical about grounding, so we should not endorse (C).

Another challenge to the argument comes from the idea that grounding is not the only kind of metaphysical relation that can be identified with the elusive in-virtue relation. Lutz discusses the following three alternatives in detail: composition, constitution, and realization. She rejects the first two of these proposals. Composition is not a relation of priority, as the in-virtue-of relation is, and constitution is either an identity relation or, more plausibly, in effect boils down to composition (implying that that which is doing the constituting is part of that which is constituted). If it is neither of these things, then it is a *sui generis* relation which, very probably, we cannot invoke as the in-virtue-of relation, if we are to make progress with Mackie’s scepticism (Lutz 2016, 194). Having compared what has been said about grounding with the ways in which realisation is generally characterized, Lutz draws the conclusion that realisation is a subspecies of grounding (Lutz 2016, 200).


Some metaethicists have explored a relation that seemed to be absent from Lutz’s discussion. It goes back at least to a paper by Rabinowicz and Österberg (1996) in which it is suggested that what value subjectivists have in mind by value is something that is “constituted” by the non-cognitive attitudes of subjects. However, it is clear that what is meant here by constitution is not

what metaphysicians generally have in mind. Still, it is certainly a view that can be interpreted as having metaphysical implications.

In Chapter 5, “The Explanatory Challenge Revisited”, Lutz turns her attention to two questions. Can grounding explain evaluative supervenience? Can we explain why certain natural facts ground evaluative facts?

As we move further into the chapter, the metaphysical focus steps up another notch, and it becomes quite clear that several tough challenges face anyone wanting to apply the grounding framework—both about value and about natural facts. Lutz probes deeply into metaphysics in her attempt to develop her own answer to these questions, and the result is close to a metaphysical tour de force. In carving out her position, she oscillates between more or less reasonable views about what the fundamental metaphysical entities are. Frequently fascinating, at a few points the discussion also borders on the puzzling. Some readers may struggle to follow it in places. This happened to me a few times, but I always suspected that this showed I needed to think harder about the issues I had begun to find puzzling.

Toni Rønnow-Rasmussen

 0000-0001-8599-4814

Lund University

toni.ronnow-rasmussen@fil.lu.se

- AUDI, Paul. 2012. “A Clarification and Defense of the Notion of Grounding.” in *Metaphysical Grounding. Understanding the Structure of Reality*, edited by Fabrice CORREIA and Benjamin Sebastian SCHNIEDER, pp. 101–121. Cambridge: Cambridge University Press, doi:[10.1017/CBO9781139149136](https://doi.org/10.1017/CBO9781139149136).
- DANCY, Jonathan. 2004. *Ethics Without Principles*. Oxford: Oxford University Press, doi:[10.1093/0199270023.001.0001](https://doi.org/10.1093/0199270023.001.0001).
- FINE, Kit. 2012. “Guide to Ground.” in *Metaphysical Grounding. Understanding the Structure of Reality*, edited by Fabrice CORREIA and Benjamin Sebastian SCHNIEDER, pp. 37–80. Cambridge: Cambridge University Press, doi:[10.1017/CBO9781139149136.002](https://doi.org/10.1017/CBO9781139149136.002).
- LUTZ, Anika. 2016. *Good in Virtue Of. A Metaethical Application of Grounding*. Analytica: Investigations in Logic, Ontology, and the Philosophy of Language. München: Philosophia Verlag, doi:[10.2307/j.ctv2nrzgw0](https://doi.org/10.2307/j.ctv2nrzgw0).
- RABINOWICZ, Włodzimierz [Wlodek] and ÖSTERBERG, Jan. 1996. “Value Based on Preferences.” *Economics and Philosophy* 12(1): 1–27, doi:[10.1017/s0266267100003692](https://doi.org/10.1017/s0266267100003692).
- RABINOWICZ, Włodzimierz [Wlodek] and RØNNOW-RASMUSSEN, Toni. 2021. “Explaining Value: On Orsi and Garcia’s Explanatory Objection to the

Fitting-Attitude Analysis.” *Philosophical Studies* 178(8): 2473–2482,  
doi:10.1007/s11098-020-01558-0.

RØNNOW-RASMUSSEN, Toni. 2016. “On-Conditionalism: On the Verge of a New  
Metaethical Theory.” *Les ateliers de l'éthique / The Ethics Forum* 11(2–3): 88–107,  
doi:10.7202/1041768ar.

—. 2021. *The Value Gap*. Oxford: Oxford University Press,  
doi:10.1093/oso/9780192848215.001.0001.

ROSEN, Gideon. 2010. “Metaphysical Dependence: Grounding and Reduction.” in  
*Modality: Metaphysics, Logic, and Epistemology*, edited by Bob HALE and Aviv  
HOFFMANN, pp. 109–136. Oxford: Oxford University Press,  
doi:10.1093/acprof:oso/9780199565818.001.0001.

Published by *Philosophie.ch*

Verein philosophie.ch  
Fabrikgässli 1  
2502 Biel/Bienne  
Switzerland  
[dialectica@philosophie.ch](mailto:dialectica@philosophie.ch)

<https://dialectica.philosophie.ch/>

ISSN 0012-2017

ISBN 1234-5678

This work is licensed under a [Creative Commons Attribution 4.0 International License](#).

*Dialectica* is supported by the [Swiss Academy of Humanities and Social Sciences](#).

#### Abstracting and Indexing Services

The journal is indexed by the Arts and Humanities Citation Index, Current Contents, Current Mathematical Publications, Dietrich's Index Philosophicus, IBZ — Internationale Bibliographie der Geistes- und Sozialwissenschaftlichen Zeitschriftenliteratur, Internationale Bibliographie der Rezensionen Geistes- und Sozialwissenschaftlicher Literatur, Linguistics and Language Behavior Abstracts, Mathematical Reviews, MathSciNet, Periodicals Contents Index, Philosopher's Index, Repertoire Bibliographique de la Philosophie, Russian Academy of Sciences Bibliographies.

# Contents

JARED WARREN, <i>Gruesome Counterfactuals</i> . . . . .	314
ARINA PISMENNY, <i>When Is Jealousy Appropriate?</i> . . . . .	331
BRYCE GESSELL, <i>The Legend of Hermann the Cognitive Neuroscientist</i> . . . . .	359
MARIA SEKATSKAYA & GERHARD SCHURZ, <i>Alternative Possibilities and the Meaning of ‘Can’</i> . . . . .	379
DEBORAH RAIKA MÜHLEBACH, <i>Neopragmatist Inferentialism and the Meaning of Derogatory Terms—A Defence</i> . . . . .	409
GUIDO MELCHIOR, <i>Sensitivity and Inductive Knowledge Revisited</i> . . . . .	441
ULRICH SCHWABE, <i>Review of Imhof (2014)</i> . . . . .	463
TONI RØNNOW-RASMUSSEN, <i>Review of Lutz (2016)</i> . . . . .	469