JOURNAL *of* ETHICS
& SOCIAL PHILOSOPHY

VOLUME XXIV · NUMBER 1
February 2023

ARTICLES

DISCUSSION

JOURNAL *of* ETHICS & SOCIAL PHILOSOPHY
http://www.jesp.org

The *Journal of Ethics and Social Philosophy* (ISSN 1559-3061) is a peer-reviewed online journal in moral, social, political, and legal philosophy. The journal is founded on the principle of publisher-funded open access. There are no publication fees for authors, and public access to articles is free of charge and is available to all readers under the CREATIVE COMMONS ATTRIBUTION-NONCOMMERCIAL-NODERIVATIVES 4.0 license. Funding for the journal has been made possible through the generous commitment of the Gould School of Law and the Dornsife College of Letters, Arts, and Sciences at the University of Southern California.

The *Journal of Ethics and Social Philosophy* aspires to be the leading venue for the best new work in the fields that it covers, and it is governed by a correspondingly high editorial standard. The journal welcomes submissions of articles in any of these and related fields of research. The journal is interested in work in the history of ethics that bears directly on topics of contemporary interest, but does not consider articles of purely historical interest. It is the view of the associate editors that the journal's high standard does not preclude publishing work that is critical in nature, provided that it is constructive, well-argued, current, and of sufficiently general interest.

Journal of Ethics and Social Philosophy *https://doi.org/10.26556/jesp.v24i1.1696*
vol. 24, no. 1 · february 2023

YOU JUST DIDN'T CARE ENOUGH

QUALITY OF WILL, CAUSATION, AND BLAMEWORTHINESS FOR ACTIONS, OMISSIONS, AND OUTCOMES

Mattias Gunnemyr and Caroline Torpe Touborg

We may ask at least two different questions about blameworthiness: What makes someone blameworthy *for* something, and *how* blameworthy is she?[1] In this paper, we shall focus on the first of these two questions. In the case below, for illustration, our immediate reaction is that Suzy is blameworthy for breaking the window:

> *Solo Suzy*: Suzy is walking down the street. When she reaches the big house on the corner, she stops and considers. She has an intense dislike for the elderly couple who lives in the house, and she has just had an idea: she is going to upset them by breaking their window on the first floor. She carefully selects a stone and hurls it toward the window. She feels a jolt of satisfaction when she hears the sound of breaking glass. Then she walks on as if nothing has happened.

In virtue of what is Suzy blameworthy for breaking the window in this case? More generally: What distinguishes cases where an agent is blameworthy for something—an action, omission, or outcome—from cases where she is not?[2]

We aim to develop a compatibilist answer to this question. In doing so, we shall draw on two important approaches in the literature: the quality-of-will approach and the actual-sequence approach.

[1] These two questions correspond closely to Zimmerman's distinction between the *scope* of blameworthiness, i.e., what you are blameworthy *for*, and the *degree* of blameworthiness, i.e., *how* blameworthy you are ("Taking Luck Seriously"). Zimmerman uses this distinction to reject moral luck by arguing that while luck matters for what you are blameworthy *for*, it does not matter for your *degree* of blameworthiness. In this paper, we stay neutral on the question of moral luck and focus simply on understanding blameworthiness *for*.

[2] A parallel question may be asked about praiseworthiness for actions, omissions, and outcomes. In the following, we set this aside.

The quality-of-will approach is based on Strawson's suggestion that blame is tied to the reactive attitudes, particularly resentment, and that those attitudes in turn respond to an agent's quality of will: "The reactive attitudes I have so far discussed are essentially reactions to the quality of others' wills towards us, as manifested in their behaviour: to their good or ill will or indifference or lack of concern."[3] This idea has been developed by Arpaly, Björnsson, McKenna, Wallace, and others.[4] Proponents of the quality-of-will approach tend to focus on the question of *how* blameworthy an agent is (the exception is Björnsson). However, as we will argue, quality of will also fits naturally when we are thinking about blameworthiness *for*. That is, quality of will is a natural starting point when analyzing the conditions under which others are warranted in reacting negatively to us (with resentment, indignation, or the like) in virtue of what we have done or brought about.

The actual-sequence approach takes its inspiration mainly from Frankfurt-style cases. On this approach, what matters in determining whether an agent is blameworthy for an action, omission, or outcome is the actual causal sequence leading up to that action, omission, or outcome.[5]

Björnsson's account elegantly combines these two approaches.[6] The basic idea of his account is:

> *Basic Idea*: You are blameworthy for *X*—where *X* may be an action, omission, or outcome—just in case there is a time *t* such that your poor quality of will at *t* stands in the right causal-explanatory relation to *X*.

Exactly how to understand "poor quality of will" is a matter of debate. Strawson characterizes poor quality of will in terms of manifesting ill will, indifference, or lack of concern; Björnsson characterizes it as caring less than is required; and McKenna characterizes it as showing insufficient regard.[7] In our formal

[3] Strawson, "Freedom and Resentment," 15.

[4] Arpaly, *Unprincipled Virtue*; Björnsson, "Explaining Away Epistemic Skepticism about Culpability" and "Explaining (Away) the Epistemic Condition on Moral Responsibility"; McKenna, *Conversation and Responsibility*; and Wallace, *Responsibility and the Moral Sentiments*. See also, e.g., Shoemaker, *Responsibility from the Margins*; Smith, "Responsibility for Attitudes"; Talbert, "Moral Competence, Moral Blame, and Protest"; and Watson, "Responsibility and the Limits of Evil."

[5] For developments of this idea, see Fischer and Ravizza, *Responsibility and Control*; and Sartorio, *Causation and Free Will*.

[6] See Björnsson, "Being Implicated," "Explaining Away Epistemic Skepticism about Culpability," and "Explaining (Away) the Epistemic Condition on Moral Responsibility."

[7] Strawson, "Freedom and Resentment"; Björnsson, "Being Implicated," "Explaining Away Epistemic Skepticism about Culpability," and "Explaining (Away) the Epistemic Condition on Moral Responsibility"; and McKenna, *Conversation and Responsibility*.

definitions, we shall refer simply to poor quality of will, leaving it open precisely how this should be understood. In our discussions, though, we often adopt Björnsson's proposal and understand quality of will in terms of care. The reader is free to substitute their own preferred understanding.

In this paper, we present a new way to develop the Basic Idea. First, we argue that it needs to be refined in a number of ways (section 1). Next, we present an account of the relevant causal-explanatory relation (section 2) and finalize the account of when you are blameworthy for actions, omissions, and outcomes, testing it on a number of cases (section 3). Finally, we show that this account also gives the right verdict in Frankfurt-style cases (section 4) and in collective-harm cases (section 5).

1. Developing the Basic Idea

The Basic Idea already gives the intuitively right verdict in paradigm cases of blameworthiness *for*, such as Solo Suzy, where the intuitive verdict is that Suzy is blameworthy for throwing the rock toward the window (an action) and for breaking the window (an outcome). Here, Suzy has a poor quality of will—she dislikes the elderly couple who lives in the house and wants to break their window in order to upset them. Furthermore, Suzy's poor quality of will just before she throws her rock stands in the right causal-explanatory relation both to her throwing the rock and to the breaking of the window: Suzy's poor quality of will causes/explains her throwing the rock and the breaking of the window. Thus, Suzy is blameworthy both for throwing the rock and for the breaking of the window.

In Solo Suzy, Suzy intentionally breaks the window. In other cases, however, you may be blameworthy for something even though you did not do it or bring it about intentionally.[8] When you have a poor quality of will, you may forget things you should remember, you may fail to notice things, or neglect to consider them. Suppose, for example, that you do not care as you should and therefore forget your best friend's birthday. In that case, we think that you are blameworthy for forgetting the birthday—even though, of course, you did not

[8] Voluntarists about moral responsibility such as Fischer and Ravizza (*Responsibility and Control*) and Rosen ("The Alethic Conception of Moral Responsibility") would not agree. According to them, voluntary control is a precondition on being blameworthy, and you do not have voluntary control over, e.g., forgetting something. Still, it seems that you *are* blameworthy for something in such cases. Here, voluntarists typically argue that you are blameworthy for some earlier action or decision that you did have voluntary control over (given that you also satisfy some epistemic condition), such as failing to add a note in your calendar about your friend's birthday. This is the *tracing* strategy. There are, however, problems with the tracing strategy (see, e.g., Smith, "Attitudes, Tracing, and Control").

do this intentionally.[9] The Basic Idea easily captures this: you are blameworthy for forgetting the birthday, because your poor quality of will—your not caring enough—stands in the right causal-explanatory relation to your forgetting. For another example, suppose that you plan a weekend at the golf course with your colleagues, without even considering visiting your injured daughter at the hospital.[10] Here too, the Basic Idea captures why you are blameworthy: you failed to even consider visiting your daughter because you did not care enough about her.

However, the Basic Idea needs a number of refinements. In the remainder of this section, we introduce these refinements gradually, motivated by a series of cases.[11]

The first refinement is motivated by the following observation: when we blame someone for something, this seems to imply that what they are blamed for is *bad*. On its own, however, the Basic Idea delivers the result that you may be blameworthy for a *good* outcome, if it is caused/explained by your poor quality of will. A simple way to fix this is to add a further necessary condition to the Basic Idea: you are blameworthy for *X*—an action, omission, or outcome—only if *X* is bad. However, this is not quite right. First, it is at best difficult, and at worst impossible, to define what it is for something—an action, omission, or outcome—to be bad *tout court*. It seems much easier to make comparative judgments that an action or outcome is worse than some alternative. Second, there are cases where it seems that you can be blameworthy for making a negative difference, even though the outcome that happens does not seem bad as such. Suppose, for example, that Sally and Bob are cooking chili together. Bob is careful about following the recipe. Sally, on the other hand, is more attentive to her phone than to her cooking and fails to put in some of the ingredients. The chili still ends up being tasty, though not quite as tasty as it would have been with all the ingredients. In this case, we think it makes sense to say that Sally is blameworthy for the chili turning out as it did, even though this outcome is

[9] Smith, "Responsibility for Attitudes."

[10] McKenna, *Conversation and Responsibility*.

[11] In "Explaining Away Epistemic Skepticism about Culpability" and "Explaining (Away) the Epistemic Condition on Moral Responsibility," Björnsson suggests that a development of the Basic Idea should incorporate both an evaluative dimension (the object of blame must be bad) and the requirement that the agent's bad quality of will should explain the object of blame in a normal way. This allows his account to handle the kinds of cases we discuss in the following: in cases where the outcome is only comparatively bad, Björnsson would suggest that comparative badness is a way to be bad, and in Tragedy, Björnsson would suggest that the agent's bad quality of will does not explain the runaway consequences in a normal way. The refinements we suggest in this section draw on Björnsson's insights and implement them in a new way.

not bad: we are warranted in reacting negatively to her since the chili turned out as it did, rather than turning out even better.

Both considerations point toward the same solution: that blameworthiness *for* involves a comparative element. Fully spelled out, you are not simply blameworthy for *X*, where *X* is some action, omission, or outcome. Rather, you are blameworthy for the occurrence of *X* rather than *X**, where *X* is worse than *X**.[12] This yields the following refined version of the Basic Idea:

> *Blameworthiness For #1*: You are blameworthy for *X* rather than *X** just in case
>
> 1. *X* is worse than *X**, and
> 2. there is a time *t* such that your poor quality of will at *t* stands in the right causal-explanatory relation to *X* rather than *X**.[13]

This refined version easily handles the cases we have considered so far. However, problems still remain. Consider the following tragic variation of Solo Suzy:

> *Tragedy*: Everything is as in Solo Suzy up to the point where the window breaks. But the consequences of the window breaking are dire. The husband is so upset at seeing the broken window that he suffers a heart attack and dies. Unable to cope with her husband's sudden death, the wife has a nervous breakdown and never fully recovers. Her daughter has to abandon a promising artistic career in Australia and come home to take care of her mother for the next several years. If Suzy had not broken the window, none of this would have happened. Instead, the couple would have continued to live happily together for many years, and their daughter would have been free to pursue her promising artistic career in Australia.

We have no doubt that Suzy is blameworthy for throwing her rock and for breaking the window. But is she also blameworthy for the runaway consequences: the husband's heart attack? the wife's nervous breakdown? the end

[12] In ordinary language, we do not typically say that one is blameworthy for one thing *rather than another*. Instead, the relevant contrast is supplied by the context.

[13] On the intended reading, *X* is an event that actually occurred, while *X** is a merely possible event that is incompatible with *X*. This is, for example, the case with the chili: the chili turning out as it did is an actual event, while the chili turning out even better is a merely possible event. Sally is blameworthy for the fact that the chili turned out as it did, rather than turning out even better. There is an alternative reading where both *X* and *X** are events that actually occurred. Suppose, for example, that Ben is also involved in the cooking and botches the dessert. If someone were to blame Sally for the failed dessert, we might correct them by saying, "Sally is blameworthy for the chili rather than (being blameworthy for) the dessert." The rather-than construction is ambiguous between these two readings. Throughout the following, we intend the first.

of the daughter's promising artistic career? We do not think so. According to Blameworthiness For #1, however, she is: Suzy's poor quality of will causes/explains both the breaking of the window and the unfortunate events that follow.

The case shows that there has to be a tighter fit between *what* an agent is blameworthy for and the *way* in which her quality of will is poor. What is the required fit? Here is a suggestion: an agent is blameworthy for *X* rather than *X** only if her poor quality of will *specifically in relation to X versus X** stands in the right causal-explanatory relation to *X* rather than X*. We may state the modified condition as follows:

> *Blameworthiness For #2*: You are blameworthy for *X* rather than *X** just in case
>
> 1. *X* is worse than *X**, and
> 2. there is a time *t* such that your poor quality of will at *t* in relation to *X* versus *X** stands in the right causal-explanatory relation to *X* rather than *X**.

This captures what we need. To start with an easy case, suppose that although Suzy's quality of will is poor in that she wants to see the elderly couple upset, she still cares as she should about more serious outcomes, such as whether the elderly people might die or suffer a nervous breakdown, just as she still cares as she should about their daughter's artistic career. If she learned what happened next, she would be horrified and exclaim something like this: "It's true that I wanted to upset them, but I never wanted something like this to happen!" If we were to blame her, e.g., for the husband's death in this case, there clearly would not be the right fit between the *way* in which her quality of will was poor and *what* we blame her for: although her quality of will was poor in relation to the elderly couple's getting upset, it was *not* poor in relation to the possibility that the husband might die. Thus, condition 2 fails to be satisfied.

Blameworthiness For #2 successfully captures why Suzy is blameworthy for upsetting the elderly couple, but not for the husband's death, the wife's nervous breakdown, or the end of the daughter's promising artistic career. On further inspection, however, an unexpected difficulty arises: it is not actually clear that we still get the result that Suzy is blameworthy for throwing her rock or even for breaking the window. Consider Suzy's throwing her rock. According to Blameworthiness For #2, Suzy is blameworthy for throwing her rock rather than not only if there is a time when she has a poor quality of will in relation to throwing her rock rather than not. But as we have told the story so far, we have not said anything to the effect that Suzy has a poor quality of will in relation to throwing her rock rather than not—we have merely said that she has a poor quality of will in relation to the elderly couple's getting upset. In that case, Blameworthiness For #2 does not entail that Suzy is blameworthy for throwing the rock.

Fortunately, there is an easy way to solve this difficulty. Even though Suzy's throwing her rock is not intrinsically worse than her not doing so, Suzy's throwing her rock *is* worse than not throwing in virtue of how her throwing (rather than not) is related to other things—such as the elderly couple's getting upset. Suzy's quality of will is poor in relation to her throwing in precisely this sense: she does not care as she should about some of the outcomes (such as the elderly couple's getting upset) that make her throwing worse than not throwing. We may capture this as follows:

> *Blameworthiness For #3*: You are blameworthy for X rather than X^* just in case there is a Y and Y^* such that
>
> 1. X is worse than X^*, at least partly in virtue of Y being worse than Y^*, and
> 2. there is a time t such that your poor quality of will at t in relation to Y versus Y^* stands in the right causal-explanatory relation to X rather than X^*.

This secures the verdict that Suzy is blameworthy for throwing the rock: (1) Suzy's throwing the rock (X) is worse than her not throwing it (X^*), at least partly in virtue of the old couple's getting upset (Y) being worse than their not getting upset (Y^*); and (2) the time t just before she throws is such that Suzy's poor quality of will at t in relation to the elderly couple's getting upset (Y) versus not getting upset (Y^*) stands in the right causal-explanatory relation to her throwing the rock (X) rather than not (X^*). We similarly get the verdict that Suzy is blameworthy for breaking the window.

In cases where you do have a poor quality of will directly in relation to X versus X^*, we may set $Y = X$ and $Y^* = X^*$, effectively making Blameworthiness For #3 equivalent to Blameworthiness For #2, which is easier to use. In such cases, we will say that X *just is* worse than X^* (leaving it open in virtue of what X is worse than X^*). In this way, Blameworthiness For #3 straightforwardly gives the verdict that Suzy is blameworthy for the elderly couple's getting upset rather than not.

2. Characterizing the Right Causal-Explanatory Relation

Until now, we have relied on an intuitive understanding of "the right causal-explanatory relation." In this section, we suggest that the relevant relation just is *causation*. The success of this kind of suggestion depends critically on the account of causation that is used. We consider this in detail and suggest that the account of causation proposed by Touborg works well with our account of blameworthiness *for*.[14]

According to Touborg's account, there are two necessary and jointly sufficient conditions for causation. First, a cause has to *produce* its effect, in the sense that it has to be connected to its effect via a genuine process. Second, the effect has to *depend* on the cause, in the sense that the security of the effect has to depend on the cause.[15]

In the following, we first present the condition of production and then the condition of dependence. Fully spelled out, both conditions are complex; here we only include as much detail as we need to explain our account of blameworthiness *for*. For the sake of simplicity, we only consider causation in worlds with deterministic laws. Correspondingly, we assume determinism in the examples we consider below. However, we believe the account could be extended to also apply to causation in worlds with indeterministic laws.

2.1. Production as Process-Connection

Let us begin with the production condition. The guiding idea behind this condition is that a cause must be connected to its effect via a genuine process. This idea is familiar from the proposal that causation should be understood in terms of physical processes.[16] In its simplest form, this proposal may be stated as follows:

> *Physical Process*: *C* is a cause of *E* just in case *C* is connected to *E* via a physical process.

A physical process is here understood in terms of transfers of physical quantities—mass, energy, etc. To illustrate the idea, consider a paradigm case of causation, such as Suzy's throwing her rock and breaking the window. Here, there is indeed a physical process connecting Suzy's throw, through the trajectory of her rock and its impact on the window pane, to the shattering of the window.

[14] See Touborg, *The Dual Nature of Causation*.

[15] Touborg's account of causation is inspired by Hall's proposal that there are two concepts of causation: the concept of production and the concept of dependence. See Hall, "Two Concepts of Causation." Hall originally gave demanding conditions for production and dependence, and suggested that production and dependence were each individually sufficient for causation. However, he later abandoned this proposal in the face of counterexamples (Hall, "Structural Equations and Causation"). By contrast, Touborg suggests conditions for production and dependence that are much weaker, with production and dependence being individually necessary and jointly sufficient for a single concept of causation.

[16] See, e.g., Dowe, *Physical Causation*.

However, trouble is not far to seek: the proposal that a cause must be connected to its effect via a physical process cannot accommodate omissions and absences as causes and effects. This means, for example, that it cannot deliver the intuitively correct verdict on a case like the following:

> *Indifferent John*: John is walking along a beach and sees a child struggling in the water. John believes that he could save the child with very little effort, and in fact he could, but he is disinclined to expend any energy to help anyone. He decides not to save the child and continues to walk along the beach.[17]

Intuitively, John's failure to jump into the water and save the child is a cause of the child's death. However, John's failure to intervene does not transfer any physical quantities or exert any push or pull on the drowning child; it is a mere absence. Thus, Physical Process delivers the verdict that John's failure to jump into the water and save the child is *not* a cause of the child's death. This verdict is counterintuitive, and especially so in the context of blame.

The trouble extends further: even when the candidate cause and effect are both ordinary positive events, Physical Process delivers the verdict that there is no causal connection when an omission or absence features as an intermediary. Thus, proponents of Physical Process have to deny that pulling the trigger causes gunshot wounds, or that decapitation causes death, since there is an intermediary absence or omission in both cases: squeezing the trigger removes an obstacle that would have prevented the flight of the bullet; decapitation stops the blood flow, which would have prevented brain starvation.[18] (Such cases are called "double prevention cases," since in these cases *C* causes *E* by preventing *D*, which would have prevented *E*.)

These cases show that it cannot be a necessary condition for causation that a cause must be connected to its effect via a physical process, when this is understood in terms of transfers of physical quantities. To capture the intuitive idea that some kind of connecting process is necessary for causation, we instead need a more abstract notion of a process, which can include omissions and absences. Touborg suggests that we may get such a more abstract notion of a process by starting from *minimal sufficiency*.[19]

Minimal sufficiency is a relation between a set of simultaneous events *S* and a later event *E*, where events are understood broadly, so as to include omissions

[17] This case is based on Fischer and Ravizza's case "Sloth." See Fischer and Ravizza, *Responsibility and Control*, 125.

[18] See, e.g., Schaffer, "Causation by Disconnection."

[19] Touborg, *The Dual Nature of Causation*, ch. 6.

and absences. A set of simultaneous events S is minimally sufficient for a later event E just in case the occurrence of all the events in S guarantees (given the laws of nature) that E will occur; and if any event is removed from S, the remaining events no longer guarantee (given the laws of nature) that E will occur. Importantly, *minimal sufficiency* is a relation between actual events: only actual events—including actual omissions and absences—may feature in the set S; and the later event E also has to be an actual event (where this includes actual omissions and absences).

Let us say that there is an *apparent process* from C to E when C is connected to E via a *chain* of such relations of minimal sufficiency. This is so when C belongs to a set of simultaneous events S_0, which is minimally sufficient for some later event D_1; D_1 belongs to a set of simultaneous events S_1, which is minimally sufficient for some later event $D_2; \ldots;$ and D_n belongs to a set of simultaneous events S_n, which is minimally sufficient for the later event E. When we look more closely—by considering more and more intermediate times between C and E—we may sometimes find that the apparent process from C to E was not genuine: when we consider these intermediate times, we can no longer find a chain of relations of minimal sufficiency connecting C to E. In order for C to be *process-connected* to E, the connection must remain *when we consider more and more intermediate times* between C and E.[20]

This notion of process-connection is sufficiently abstract to accommodate omissions and absences. Returning to the case of Indifferent John, for example, we find that John's poor quality of will (in relation to the child's drowning versus surviving) is process-connected to the child's drowning. John's poor quality of will at the time t just before he decides not to intervene belongs to a set of simultaneous events that is minimally sufficient for the child's drowning. And this connection remains no matter how many intermediate times we consider. Take, for example, the intermediate time t' after John has decided not to intervene and before the child has drowned. Here, we find that John's poor quality of will at t belongs to a set of simultaneous events that is minimally sufficient for his failure to intervene at t' (remember, his failure to intervene is an actual event), and his failure to intervene at t' in turn belongs to a set of simultaneous events that is minimally sufficient for the child's drowning. Thus, John's poor quality of will at t is process-connected to the child's drowning.

The notion of process-connection also allows us to distinguish genuine causes from preempted backups in cases such as the following:

[20] The full definition of process-connection includes a further refinement: to be able to handle cases of late preemption, it makes use of a more demanding, *time-sensitive* relation of minimal sufficiency. For simplicity, we leave out this refinement. See Touborg, *The Dual Nature of Causation*, 143–48.

> *Backup Billy*: Everything is as in Solo Suzy, except that Billy also wants the window to break. On seeing that Suzy throws her rock, Billy is satisfied and walks away. However, if Suzy had not thrown her rock, Billy would have thrown a rock himself a moment later, and the window would still have broken.

In this case, Suzy's poor quality of will guarantees that the window will break, and so does Billy's. However, only Suzy's poor quality of will (in relation to the elderly couple's getting upset versus not) is process-connected to the shattering of the window. To see this, the key is to look at intermediate times. Let *t* be the time just before Suzy throws her rock. Then Suzy's poor quality of will at *t* belongs to a set of simultaneous events that is minimally sufficient for the shattering of the window; and similarly, Billy's poor quality of will at *t* belongs to a set of simultaneous events that is minimally sufficient for the shattering of the window. However, when we bring in more and more intermediate times, we find that we can keep filling in the details in the chain connecting Suzy's poor quality of will to the shattering of the window—going from Suzy's poor quality of will, to her decision to throw, to her throwing the rock, to the rock's trajectory and impact on the window pane, and finally to the shattering of the window. By contrast, the connection between Billy's poor quality of will at *t* and the window shattering breaks down when we consider intermediate times. Consider, for example, a time *t′* after Suzy has thrown her rock and Billy has turned away, but before the window shatters. To connect Billy's poor quality of will to the breaking of the window, we would need an event *D* at this time *t′*—such as Billy's rock flying toward the window—such that Billy's poor quality of will belongs to a set of events that is minimally sufficient for *D*, and *D* in turn belongs to a set of events that is minimally sufficient for the window shattering. But there is no such event *D* in the actual world. For this reason, Billy's poor quality of will at *t* is not process-connected to the breaking of the window. This fits the judgment that Suzy's poor quality of will at *t* is a cause of the shattering of the window, while Billy's poor quality of will is not. In this way, the notion of process-connection does crucial work in distinguishing genuine causes from preempted backups.

However, process-connection is not sufficient for causation. The condition of process-connection needs to be supplemented with a second necessary condition for causation, requiring that a cause must make a difference to its effect. The need for this is brought out by the following three considerations.

First, process-connection on its own cannot yield the intuitively correct verdict on counterexamples to the transitivity of causation, such as the following switching case:

> *Trolley Trouble*: Suzy is standing by a switch in the tracks as a trolley approaches in the distance. If she flips the switch, the trolley will travel down the left-hand track; if she does not flip the switch, it will travel down the right-hand track. Further ahead, the tracks converge again, and after that, five people are tied to the single track. Suzy wants the five to get run over, and she erroneously believes that they are tied to the left-hand track. She flips the switch so that the trolley travels down the left-hand track and subsequently runs over the five people. However, if she had not flipped the switch, the trolley would still have run over the five, reaching them via the right-hand track.[21]

Intuitively, Suzy's poor quality of will at *t* (the time just before she flips the switch) is not a cause of the five's death. However, Suzy's poor quality of will at *t* is process-connected to their death. Thus, process-connection is not sufficient for causation.

You might not immediately notice that Suzy's poor quality of will is process-connected to the death of the five. Indeed, Suzy's poor quality of will at *t* does not belong to a set of simultaneous events that is itself minimally sufficient for their death: a set that leaves out Suzy's poor quality of will and contains just the approach of the trolley, the layout of the tracks, etc., is sufficient for the trolley's running over the five. However, there is a *chain* connecting Suzy's poor quality of will at *t* to the death of the five: Suzy's poor quality of will at *t* belongs to a set of simultaneous events that is minimally sufficient for the trolley's journey along the left-hand track, and the trolley's journey along the left-hand track belongs to a set of simultaneous events that is minimally sufficient for the death of the five. This connection remains when we consider more intermediate times.

The problem arises since process-connection is a transitive relation—if *C* is process-connected to *D*, and *D* is process-connected to *E*, then *C* is process-connected to *E*. By contrast, causation is not transitive: it may happen that *C* is a cause of *D*, and *D* is a cause of *E*, but *C* is not a cause of *E*—as in Trolley Trouble.[22] This is the first reason why the condition of process-connection needs to be supplemented.

[21] This case is inspired by Foot, "The Problem of Abortion and the Doctrine of Double Effect"; Thomson, "Killing, Letting Die, and the Trolley Problem"; and Van Inwagen, "Ability and Responsibility." Switching cases like this are also common in the causation literature. See, e.g., Hall, "Structural Equations and Causation"; Paul and Hall, *Causation*, 232–37; and Sartorio, "Causes as Difference-Makers." Like we do here, Sartorio uses this kind of case to motivate the idea that a cause must make a difference to its effect, though her difference-making condition differs from ours.

[22] For an overview, see Paul and Hall, *Causation*, ch. 5.

Second, process-connection cannot on its own accommodate *contrastive* causal claims. Process-connection is simply a relation between two actual events: an actual event *C* is process-connected to an actual event *E*. However, contrastive causal claims include merely possible events as contrasts to the cause *C* or the effect *E*, and the truth value of a contrastive claim depends on what these contrasts are.[23] The need to handle such contrastive causal claims is especially pressing when we are concerned with blameworthiness for actions, omissions, and outcomes: as we have seen above, Blameworthiness For is based precisely on a contrastive claim—namely that your poor quality of will stands in the right causal-explanatory relation to *X rather than X**.

Third, process-connection on its own cannot distinguish between causes and background conditions. Suppose, for example, that Selma has no royal connections. Is the queen of Sweden's failure to water Selma's flowers a cause of their death? Intuitively, it is not.[24] However, the queen's failure to water Selma's flowers *is* process-connected to their death. So if we take process-connection to be sufficient for causation, we cannot accommodate the intuitive verdict in this case.

These difficulties have a common solution: recognizing that there is a second necessary condition for causation, which captures the intuitive idea that a cause must make a difference to its effect.

2.2. Dependence and Security

The core idea that causes are difference-makers is familiar. For example, David Lewis writes that "we think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have happened without it."[25] This idea is the starting point for counterfactual accounts of causation. In its contrastive form, it may be stated as follows:

> *Simple*: Suppose that *C* occurs at *t*, *E* occurs later, and *E** is incompatible with *E*. Then *C* is a cause of *E* rather than *E** just in case, if *C* had not occurred, then *E** would have occurred instead of *E*.

The heart of Simple is the counterfactual: "if *C* had not occurred, then ..." To evaluate this counterfactual, we first identify all the worlds where *C* does not occur. Among these, we consider the worlds that are *closest* to the actual world

[23] See, e.g., Schaffer, "Contrastive Causation" and "Causal Contextualism."

[24] See, e.g., Hart and Honoré, *Causation in the Law*, 38; and McGrath, "Causation by Omission."

[25] Lewis, "Causation," 557.

@. If the consequent is true in each of these worlds, then the counterfactual is true; otherwise it is false.

The relevant notion of closeness is standardly understood in terms of similarity between entire worlds. Following Paul and Hall, we prefer instead to understand it in terms of similarity between states of worlds at times.[26] Thus, we shall say that two possible worlds *w* and *w** are close-at-time-*t* to the extent that the state of *w* at *t* is similar to the state of *w** at *t*. Supposing that *C* occurs at time *t*, this means that the counterfactual "if *C* had not occurred, then …" is true just in case the consequent is true in each of the closest-to-@-at-*t* worlds where *C* does not occur.

Even with this clarification, a question remains: What replaces *C* in the closest-to-@-at-*t* worlds where *C* does not occur? An obvious answer is that *C* is replaced by an event that is as similar as possible to *C*, without satisfying *C*'s conditions of occurrence. However, this proposal yields intuitively false results. As Lewis writes: "if *C* had not occurred and almost-*C* had occurred instead, very likely the effects of almost-*C* would have been much the same as the actual effects of *C*. So our causal counterfactual will not mean what we thought it meant, and it may well not have the truth-value we thought it had."[27] That will not do. We need an alternative proposal about what replaces *C*.

Our preferred answer is that, when we evaluate counterfactuals, we do not in fact consider *all* possible worlds. Rather, we only consider a restricted class of possible worlds—namely, those possible worlds that we take to be relevant. This restricted class of possible worlds is itself a causal relatum; we shall call it a *possibility horizon*.[28] Our chosen possibility horizon will typically not contain any worlds where *C* is replaced by almost-*C*. Rather, it will typically contain only worlds where *C* either occurs or is replaced by a contextually salient alternative *C** that is quite different from *C*. When we only consider the worlds within such a possibility horizon, we find that in the *closest* worlds where *C* does not occur, it is replaced by *C**.

Based on this, we may now give the following more developed version of Simple:

> *Simple**: Suppose that *C* occurs at time *t*, *E* occurs later, and *E** is incompatible with *E*. Then *C* is a cause of *E* rather than *E** within possibility horizon *H* just in case there is at least one world in *H* where *C* does not occur, and in the closest-to-@-at-*t* world(s) in *H* where *C* does not occur, *E** occurs instead of *E*.

[26] Paul and Hall, *Causation*, 47–49.

[27] Lewis, "Causation as Influence," 90.

[28] See Touborg, *The Dual Nature of Causation*, ch. 5.

Simple* can capture our intuitions in a wide range of cases. In particular, it successfully handles the cases that presented difficulties for process-connection. In Trolley Trouble, Simple* entails that Suzy's poor quality of will is not a cause of the five's death, because Suzy's poor quality of will makes no difference to their fate—they would have died either way. Furthermore, Simple* is tailor-made to handle contrastive causal claims, such as "Sally's not caring about the chili caused the chili to be just good, rather than excellent." Finally, Simple* captures the verdict that the queen of Sweden's failure to water Selma's flowers did not cause them to die: in ordinary contexts, it is not a relevant possibility that the queen waters Selma's flowers. Rather, the queen's failure to water the flowers is treated as a background condition. Thus, the possibility horizon that is in play in an ordinary context only contains worlds where the queen does not water the flowers and thus the requirement that "there is at least one world in *H* where *C* does not occur" fails to be satisfied.

As is well known, however, Simple* does not give a necessary condition for causation: there are cases where *C* is clearly a cause of *E*, even though *E* would still have occurred if *C* had not. We have already seen such a case: in Backup Billy, it is clear that Suzy's throwing her rock is a cause of the window shattering. However, if Suzy had not thrown her rock, the window would still have shattered—because, in that case, Billy would have thrown *his* rock. Cases such as this show that in order to capture the idea that making a difference is necessary for causation, we need a more subtle notion of difference-making: one that can capture, e.g., how Suzy's throw makes a difference to the shattering of the window, even though the window would still have shattered if she had not thrown.

The key to developing such a more subtle notion of difference-making is to pay attention to the *modal* features of events. In particular, when an event actually occurs, we may ask how easily it could have failed to occur; and when an event does not occur, we may ask how easily it *could* have occurred.[29] Touborg uses the notion of *security* to capture this.[30]

On Touborg's account, whenever an event actually occurs, it has *positive security*. However, it may have a higher or lower *degree* of positive security. In some cases, an event *E* actually occurs, but when we consider what was the case at some earlier time *t*, we find that if things had been just slightly different at

[29] More carefully: when a particular *type* of event does not occur, we may ask how easily an event of this *type* could have occurred. Speaking of types of events solves the difficulty that we cannot refer determinately to an event that did not occur. For simplicity, we suppress this complication in the text.

[30] Here we use the simplified definition of security presented in Touborg, "Hasteners and Delayers." For more detail, see Touborg, *The Dual Nature of Causation*, ch. 8.

time *t*, *E* would not have occurred. In such cases, we shall say that *E* had a low degree of positive security at time *t*. Suppose, for example, that Suzy in Solo Suzy throws her rock toward the window and breaks it, but if there had been just a slight gust of wind at *t* (the time when Suzy threw her rock), a swaying branch would have deflected her rock, and the window would have remained intact. In this case, the breaking of the window has a very low degree of positive security at *t*. In other cases, an event *E* actually occurs, and when we consider what was the case at some earlier time *t*, we find that things would have had to be quite different at *t* in order for *E* not to occur. In such cases we shall say that *E* had a high degree of positive security at *t*.

Whenever an event fails to occur, this event has *negative security*. Once again, it may then have a higher or lower *degree* of negative security. Consider some event *E* that does not actually occur, and consider some time *t* prior to the time when *E* would have occurred, if it did occur. We may now ask: How different would things have to be at *t* in order for *E* to occur? If things would only have to be ever so slightly different at *t* in order for *E* to occur, we shall say that *E* has a low degree of negative security at *t*: although *E* does not happen, circumstances at *t* are such that it is *close* to happening. If, on the other hand, things would have to be quite different at *t* in order for *E* to happen, we shall say that *E* has a high degree of negative security at *t*: considering the circumstances at *t*, *E* is *far* from happening.

More formally, we may understand security-at-a-time in terms of the distance-at-a-time between worlds. We have already introduced the notion of distance-at-a-time above, when we discussed the evaluation of counterfactuals. As a reminder: two possible worlds *w* and *w** are close-at-time-*t* to the extent that the state of *w* at *t* is similar to the state of *w** at *t*. Based on this, we may define security-at-a-time as follows:

> If an event *E* occurs in *w*, then *E* has positive security in *w*, and its degree of positive security at an earlier time *t* is given by the distance-at-*t* between *w* and the closest-to-*w*-at-*t* world(s) where *E* does not occur.
>
> If an event *E* does not occur in *w*, then *E* has negative security in *w*, and its degree of negative security at an earlier time *t* is given by the distance-at-*t* between *w* and the closest-to-*w*-at-*t* world(s) where *E* occurs.

This notion of security allows us to capture a more subtle notion of difference-making: making a difference to the *security* of an event. A cause does not have to make a difference as to whether its effect occurs or not. But it *does* have to make a difference to the security of its effect: supposing that *C* occurs at time *t*, it has to be the case that if *C* had not occurred, *E* would have been *less secure* at *t* than it actually was. In the case of contrastive causal claims, such as "*C* is a cause of *E* rather than *E**," *C* has to make a difference to the security of both *E* and *E**: supposing again that *C* occurs at *t*, it has to be the case that if *C* had not occurred, *E* would have been *less secure* at *t* and *E** would have been *more secure* at *t* than what was actually the case.[31] We shall call this kind of difference-making *security-dependence*.

2.3. *Causation*

So far, we have introduced two necessary conditions for causation: the condition of process-connection and the condition of security-dependence. Neither of these two conditions can stand alone. The condition of process-connection needs help from security-dependence when dealing with switching cases, contrastive causal claims, and the distinction between causes and background conditions; the condition of security-dependence needs help from process-connection when dealing with preemption cases such as Backup Billy. But together, these two conditions are jointly sufficient for causation, yielding the following account:[32]

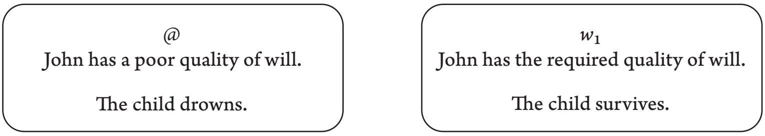
> *Causation*: Suppose that *C* occurs at *t*, *E* occurs later, and *E** is incompatible with *E*. Then *C* is a cause of *E* rather than *E** within possibility horizon *H* just in case
>
> a. *C* is process-connected to *E*, and
> b. there is at least one world in *H* where *C* does not occur, and in the closest-to-@-at-*t* world(s) in *H* where *C* does not occur, *E* is *less secure* at *t* and *E** is *more secure* at *t* than they are in @.

This account of causation handles the cases we have considered so far.

Consider first Indifferent John. As before, let *t* be the time just before John decides not to intervene. We have already seen that John's poor quality of will at *t* is process-connected to the child's drowning. We may now consider whether John's poor quality of will also satisfies the condition of security-dependence within the possibility horizon below:

[31] The suggestion that a cause must make its effect more secure is somewhat similar to the controversial suggestion that a cause must raise the probability of its effect. In particular, (apparent) counterexamples to the suggestion that causes are probability-raisers can be translated into (apparent) counterexamples to the suggestion that causes make their effects more secure. However, the notion of security within a possibility horizon offers resources to resist such counterexamples. We therefore do not think they threaten the proposal. For discussion, see Gunnemyr, *Reasons, Blame, and Collective Harms*, 284–91; and Touborg, *The Dual Nature of Causation*, 239–43.

[32] The account given here differs in one crucial respect from Touborg's account in *The Dual Nature of Causation*: Touborg does not include effect contrasts.



@
John has a poor quality of will.

The child drowns.

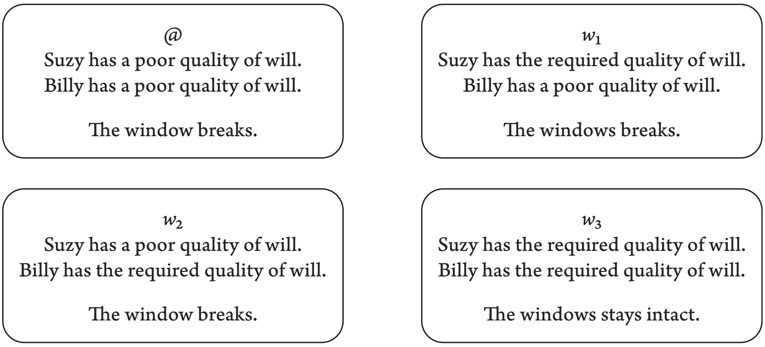
w_1
John has the required quality of will.

The child survives.

FIGURE 1 Possibility horizon H_J

Within H_J, the closest-to-@-at-t world where John does not have a poor quality of will at t is w_1, where he has the required quality of will at t. Here, John jumps into the water and saves the child. Thus, the child's drowning has negative security at t in w_1 (since it does not occur in w_1) and positive security in @ at t (since it occurs in @). From this, it immediately follows that the child's drowning is *less secure* at t in w_1 than it is in @. Similarly, the child's survival is *more secure* at t in w_1 than it is in @. Thus, Causation yields the result that John's poor quality of will at t is a cause (within H_J) of the child's drowning rather than surviving. Therefore, John is blameworthy for the child's drowning rather than surviving.[33]

Consider next Backup Billy. As before, let t be the time just before Suzy throws her rock. We may then consider what caused the window shattering within the following possibility horizon:



@
Suzy has a poor quality of will.
Billy has a poor quality of will.

The window breaks.

w_1
Suzy has the required quality of will.
Billy has a poor quality of will.

The windows breaks.

w_2
Suzy has a poor quality of will.
Billy has the required quality of will.

The window breaks.

w_3
Suzy has the required quality of will.
Billy has the required quality of will.

The windows stays intact.

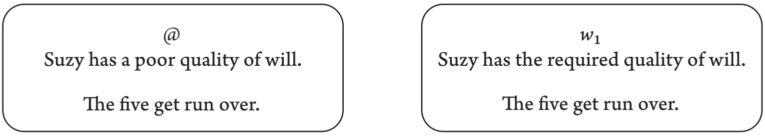
FIGURE 2 Possibility horizon H_B

The possibility horizon H_B includes a salient alternative to Suzy's poor quality of will (in relation to the elderly couple's getting upset versus not)—namely, her having the required quality of will; and it includes a salient alternative to Billy's poor quality of will—namely, his having the required quality of will. Independently

[33] Is John also blameworthy for killing the child? Killing is sometimes understood simply as causing someone's death. If killing is understood in this way, then we would have to say that John killed the child. However, we think there are further conditions on killing (roughly related to the distinction between doing and allowing, and maybe with a proximate-cause requirement; see, e.g., Woollard, *Doing and Allowing Harm*), and John does not satisfy those further conditions—he merely allows the death of the child.

of this choice of possibility horizon, we have already seen that Suzy's poor quality of will at *t* is process-connected to the shattering of the window, while Billy's poor quality of will is not. We may now verify that Suzy's poor quality of will at *t* makes a difference to the security of the window shattering. The closest-to-@-at-*t* world within H_B where Suzy does not have a poor quality of will at *t* is w_1, where she has the required quality of will at *t*. The window still shatters in w_1, since Billy has a poor quality of will and therefore throws his rock when Suzy does not. However, the window shattering is *less secure* at *t* in w_1 than it is in @: compared with @, w_1 is *closer-at-t* to w_3 where the window does not shatter. In w_1 only one thing needs to change at *t* in order for the window not to break (namely, Billy's poor quality of will), whereas in the actual world @, two things need to change at *t* in order for the window not to break (namely, both Suzy's poor quality of will and Billy's poor quality of will). And similarly, the window's remaining intact is *more secure* at *t* in w_1 than it is in @. Thus, Causation yields the desired result: within possibility horizon H_B, Suzy's poor quality of will at *t* is a cause of the window's shattering rather than remaining intact, while Billy's poor quality of will is not.

Finally, consider Trolley Trouble. We have already seen that Suzy's poor quality of will at *t* (the time just before she flips the switch) is process-connected to the five's death. Now consider the possibility horizon H_T below, where the relevant alternative to Suzy's having a poor quality of will at *t* (in relation to the five) is her having the minimally required quality of will:



@
Suzy has a poor quality of will.
The five get run over.

w_1
Suzy has the required quality of will.
The five get run over.

FIGURE 3 Possibility horizon H_T

Within this possibility horizon, Suzy's poor quality of will at *t* does not make any difference to the *security* of the five's death: there *is* no world where the five are not run over. Thus, their getting run over is infinitely secure, both in @ and in w_1. And so, their getting run over is *just as secure* in w_1 as it is in @. We therefore find, as we should, that Suzy's poor quality of will is not a cause (within possibility horizon H_T) of the five's getting run over rather than not.

3. COMPLETING THE ACCOUNT OF BLAMEWORTHINESS FOR

We suggest that the causal-explanatory relation that has to hold between an agent's poor quality of will and what she is blameworthy for is causation, understood as suggested above.

As we have seen, causation is relativized to a possibility horizon. Thus, it may sometimes be the case that C is a cause of E rather than E^* within possibility horizon H_1, while C is not a cause of E rather than E^* within a different possibility horizon H_2. This feature of the general account of causation has a number of advantages. However, it would be unsatisfactory to say, e.g., that you are blameworthy for X rather than X^* within possibility horizon H_1, but not within possibility horizon H_2. We may avoid this relativity by insisting that what matters for blameworthiness is causation within the *relevant* possibility horizon. This raises a crucial question: What is the relevant possibility horizon when evaluating what an agent is blameworthy for?

Suppose we are evaluating whether your poor quality of will at time t (in relation to Y versus Y^*) is a cause of X rather than X^*, and that the purpose of this evaluation is to determine whether you are blameworthy for X rather than X^*. To make this evaluation, we start from the actual state of the world at time t. We then identify relevant alternatives to the way things were at time t. If you had a poor quality of will at t (in relation to Y versus Y^*), we think it is relevant that you could instead have had the quality of will (in relation to Y versus Y^*) that you were minimally required to have.[34] By contrast, it is not relevant that you could have had an even worse quality of will or a saintly quality of will far above what was minimally required. Similarly, if someone else had a poor quality of will at t, we think it is relevant that *they* could have had the quality of will they were minimally required to have. But again, it is not a relevant possibility that they could have had an even worse quality of will or a saintly quality of will. Other changes to what actually happened at t may or may not be relevant as well. This gives a criterion for determining which possible worlds belong to the relevant possibility horizon: if a possible world w represents a relevant alternative to how things were at time t and evolves forward in accordance with the laws of nature, then it is included in the relevant possibility horizon. Otherwise not. We may summarize this in the following rule of thumb.[35]

> *Relevant Possibilities for Blame*: To determine, for the purpose of attributing blame, whether your poor quality of will at time t (in relation to Y versus Y^*) is a cause of a later event X rather than X^*, it is a relevant

[34] This proposal is closely related to Björnsson's proposal that what matters is how your quality of will falls short of what could be demanded. See Björnsson, "Explaining Away Epistemic Skepticism about Culpability" and "Explaining (Away) the Epistemic Condition on Moral Responsibility."

[35] What matters here is simply that these possibilities are relevant alternatives to the state of the actual world at time t. It does not matter whether the actual world could, from an earlier state, evolve into one of these alternative states (given determinism, it of course could not).

> possibility that you could instead have had the minimally required quality of will at *t* (in relation to *Y* versus *Y**). Similarly, it is a relevant possibility that anyone else involved in the situation who had a poor quality of will at time *t* could have had the minimally required quality of will at time *t*. Every combination of these possibilities is relevant. Other possibilities may or may not be relevant as well.[36]

When we discuss collective harm cases in section 5, we motivate why we have to include the possibility that each agent involved in the situation could have had the required quality of will.[37]

With this, we may now complete our account of blameworthiness for actions, omissions, and outcomes as follows:

> *Blameworthiness For*: You are blameworthy for *X* rather than *X** just in case there is a *Y* and *Y** such that
> 1. *X* is worse than *X**, at least partly in virtue of *Y* being worse than *Y**, and
> 2. there is a time *t* such that your poor quality of will at *t* in relation to *Y* versus *Y** is a cause of *X* rather than *X**, within the relevant possibility horizon *H*.[38]

We will now test this account on the cases we have considered so far.

Consider first Indifferent John. In this case, (1) the child's drowning *just is* worse than its surviving. Furthermore, (2) John has a poor quality of will at *t* (the time just before he decides not to intervene) in relation to the child's drowning versus surviving. By Relevant Possibilities for Blame, the relevant possibility

[36] This principle presupposes that everyone involved is an agent. If someone is not an agent at the relevant time—if, for instance, he or she is insane or under the influence of drugs—this person may be exempt from the requirement that they should have a particular quality of will. In that case, the relevant possibility horizon for determining blameworthiness may not include the possibility that they could have had a different quality of will. As a result, our account will yield the result that they are not blameworthy. In the following, we set such cases aside.

[37] You might wonder how to decide which agents are involved in a situation. Here, it is important to note that Relevant Possibilities for Blame is meant to ensure that enough agents are included. The verdict of our account will not change if even more agents are included. If you are unsure whether an agent is involved, the rule of thumb is to include them.

[38] We here set aside cases of deviant causation (for examples, see, e.g., Bernstein, "Moral Luck and Deviant Causation"). To handle such cases, something further is needed—at the very least, a requirement that your bad quality of will is a *nondeviant* cause of *X* rather than *X**. However, since cases of deviant causation present trouble for everyone, we think it is appropriate to set them aside for now. For discussion, see, e.g., Gunnemyr, *Reasons, Blame, and Collective Harms,* 246–51.

horizon is H_J. And as we have already seen, John's poor quality of will at t in relation to the child's drowning versus surviving is a cause of the child's drowning rather than surviving within H_J. Therefore, Blameworthiness For gives the intuitively correct verdict that John is blameworthy for the death of the child.

Consider next Backup Billy. This case is just like Solo Suzy, except that Billy is lurking in the background. There is a Y and Y^*—the elderly couple's getting upset versus not getting upset—such that (1) the breaking of the window (X) is worse than its staying intact (X^*), at least partly in virtue of the elderly couple's getting upset (Y) being worse than their not getting upset (Y^*). The relevant possibility horizon is H_B, and we have already seen that (2) within H_B, Suzy's poor quality of will at t (the time just before she throws her rock) in relation to the elderly couple's getting upset (Y) versus not (Y^*) is a cause of the window's breaking (X) rather than staying intact (X^*). Thus, Blameworthiness For yields the result that Suzy is blameworthy for the window's breaking rather than staying intact. By contrast, Billy is not blameworthy for this, since there is no process connecting Billy's poor quality of will to the breaking of the window.

Finally, consider Trolley Trouble. Here, (1) the five's getting run over *just is* worse than their not getting run over. Furthermore, (2) Suzy has a poor quality of will at t (the time just before she flips the switch to the left) in relation to the five's getting run over. However, as we have seen, Suzy's poor quality of will at t is *not* a cause of the five's getting run over rather than not within the relevant possibility horizon H_T, since it does not make any difference to the security of their getting run over. Blameworthiness For therefore delivers the intuitively correct result: Suzy is not blameworthy for the five's getting run over rather than not.[39]

Thus, Blameworthiness For gives the intuitively correct verdict in all three cases.

4. Blameworthiness and Frankfurt-Style Cases

Frankfurt-style cases are important test cases for accounts of blameworthiness.[40] In these cases, there is a backup ensuring that an agent will act in a

[39] This is of course consistent with Suzy's being blameworthy for other things—for example, for intending to kill the five. We may wonder whether it makes any difference for *how* blameworthy Suzy is that, because of circumstances outside of her control, she does not actually kill the five. This is related to the question of moral luck, which we have set aside in this paper.

[40] Such cases were first introduced by Frankfurt, "Alternate Possibilities and Moral Responsibility."

certain way. However, as things turn out, the backup does not have to intervene. Consider the following Frankfurt-style variation of Solo Suzy:

> *Backup Neuroscientist Billy:* Everything is as in Solo Suzy, except for the following: unbeknownst to Suzy, the mischievous neuroscientist Billy has implanted a chip in Suzy's brain and is now monitoring her process of deliberation. Billy wants Suzy to throw a rock and break the window. If he thinks, as a result of his monitoring of Suzy's process of deliberation, that Suzy is not going to throw a rock and break the window, he will have the chip induce this intention in her and make her act on it. As things happen, Billy does nothing.

Frankfurt-style cases such as Backup Neuroscientist Billy have a similar structure to early preemption cases such as Backup Billy, where Billy would have thrown a rock at the window had Suzy not done so. In both cases, Billy is lurking in the background with sinister intent. The difference is that, in Backup Neuroscientist Billy, Billy is going to intervene by modifying Suzy's intention, whereas in Backup Billy, he is going to intervene by breaking the window himself. If we consider the process connecting Suzy's poor quality of will to the breaking of the window, we see three salient features. First, there is her poor quality of will, then there is her throw, and finally there is the breaking of the window.

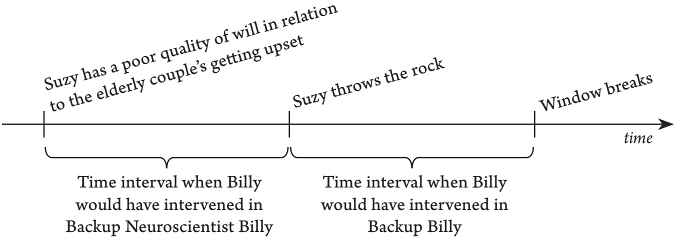
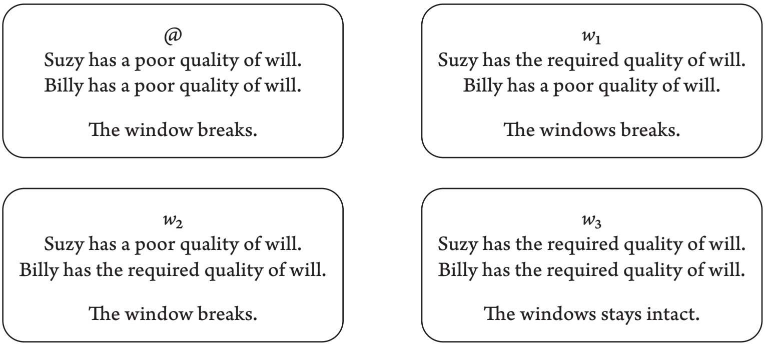

Suzy has a poor quality of will in relation to the elderly couple's getting upset
Suzy throws the rock
Window breaks
time
Time interval when Billy would have intervened in Backup Neuroscientist Billy
Time interval when Billy would have intervened in Backup Billy

FIGURE 4

Blameworthiness For gives the same result no matter where Billy is ready to enter as a backup in the process connecting Suzy's poor quality of will to the window breaking. This way, Blameworthiness For straightforwardly gives the intuitively right answer in Frankfurt-style cases. In Backup Neuroscientist Billy, for instance, Blameworthiness For gives the verdict that Suzy is blameworthy for breaking the window.

To begin with, (1) the window's breaking is worse than its staying intact, at least partly in virtue of the elderly couple's getting upset being worse than their not getting upset. Furthermore, (2) let *t* be the time right before Billy would otherwise have intervened via the chip. At this time, Suzy has a poor quality of will, but it is a relevant possibility that she could have had the required quality of will, and the same goes for Billy. Thus, the relevant possibility horizon is H_{NB}:



@
Suzy has a poor quality of will.
Billy has a poor quality of will.

The window breaks.

w_1
Suzy has the required quality of will.
Billy has a poor quality of will.

The windows breaks.

w_2
Suzy has a poor quality of will.
Billy has the required quality of will.

The window breaks.

w_3
Suzy has the required quality of will.
Billy has the required quality of will.

The windows stays intact.

FIGURE 5 Possibility horizon H_{NB}

Applying Causation, we find that Suzy's poor quality of will at *t* in relation to the elderly couple's getting upset caused the window to break within H_{NB}. We have already seen that (a) Suzy's poor quality of will at *t* (in relation to the elderly couple's getting upset) is process-connected to the breaking of the window in Backup Billy, and precisely the same reasoning shows that it is also process-connected to the breaking of the window here. Moreover, (b) the window breaking is *less secure* at *t* in the closest world where Suzy has the minimally required quality of will (namely, w_1): starting from w_1, only one thing would have to change at *t* in order for the window not to break (namely, Billy's poor quality of will); starting from @, two things would have to change at *t* in order for the window not to break (namely, both Suzy's poor quality of will and Billy's poor quality of will). For similar reasons, the window's staying intact is *more secure* at *t* in the closest possible world where Suzy has the required quality of will. Once again, this reasoning is entirely parallel to the reasoning we applied in the case of Backup Billy.

5. BLAMEWORTHINESS IN COLLECTIVE HARM CASES

Finally, our account sheds light on collective harm cases, i.e., cases where no single action is necessary or sufficient for bringing about some bad outcome. Consider the following case:

> *The Lake*: Ann, Beth, and Claire live close to a lake with a sensitive ecosystem. Each of them has a boat. They have all been using a cheap and hazardous paint. However, they have recently learned that this has brought

> the ecosystem to the verge of collapsing. Each of them now believes that if she were to switch to an environmentally friendly but more expensive paint, this might be enough to save the ecosystem from collapse. Still, none of the three cares sufficiently about the ecosystem in the lake. All three continue to use the hazardous paint, and the lake becomes a wet wasteland. As a matter of fact, the ecosystem would still have collapsed if just one of them had switched to the environmentally friendly paint. However, if two or more of them had switched to the environmentally friendly paint, the ecosystem would not have collapsed.[41]

Intuitions about this case might vary. It might seem that Ann, Beth, and Claire are blameworthy for the collapse of the ecosystem since the ecosystem collapsed as a result of what they did. It might also seem that none of them is blameworthy since it is true of each that the ecosystem would have collapsed whether or not she had switched to the environmentally friendly paint. If someone, for instance, blames Ann for the collapse of the ecosystem, she might make the following defense:

> *Defense:* "I accept that I may be blameworthy for using the hazardous paint. But I am not blameworthy for the collapse of the ecosystem. Given that Beth and Claire did not care, my lack of care did not matter."[42]

Moreover, even if we think that Ann, Beth, and Claire are blameworthy for the collapse of the ecosystem, there is a puzzle about whether they are to blame individually or collectively. On the one hand, it seems that they are not individually to blame since each can appeal to the defense just sketched. On the other hand, it seems that they could not be collectively to blame since they do not constitute a collective. They did not perform an intentional collective action, and they do not share a formal decision procedure.[43]

[41] Adapted from Björnsson, "Joint Responsibility Without Individual Control."

[42] Such a defense was considered in an early version of Björnsson, "Being Implicated," presented at the Manchester Centre for Political Theory (MANCEPT), 2019. Björnsson accepts the defense insofar as it concerns explanation: applied to The Lake, his view in "Being Implicated" is that we should reject the claim that "the ecosystem collapsed *because* Ann did not care." However, he argues that Ann is still a fitting target of reactive attitudes over the outcome because her substandard care is *involved* in a normal explanation of the ecosystem's collapse.

[43] According to standard accounts of intentional collective action, they did not perform a collective action. See Bratman, *Shared Agency*; Gilbert, *Joint Commitment*; and Searle, "Collective Intentions and Actions." For accounts of what it is to share a formal decision procedure, see French, *Collective and Corporate Responsibility*; and List and Pettit, *Group Agency.*

We suggest that the conflicting intuitions about The Lake can be understood as arising from different choices of possibility horizon—where the choice of possibility horizon may, in turn, reflect either a collective or an individual perspective. Consider the time *t* before any of the three has painted their boats. Following the standard procedure for generating the relevant possibility horizon based on what is happening at time *t* (as stated in Relevant Possibilities for Blame), we get a possibility horizon that contains every combination of Ann, Beth, and Claire having their actual poor quality of will at *t*, and having the minimally required quality of will at *t*. Call this possibility horizon H_{large}. In some worlds within this possibility horizon, the ecosystem will be saved. In other worlds, the ecosystem will collapse (as in the actual world).

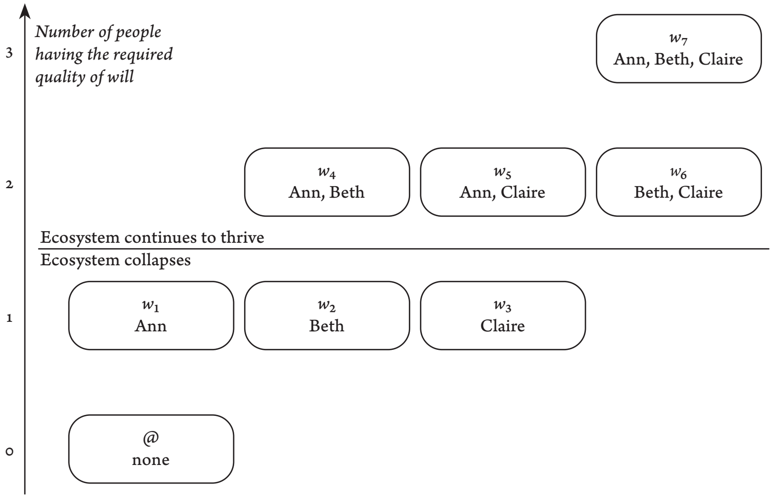


FIGURE 6 Possibility horizon H_{large}

Now consider Ann. When we use H_{large}, we find that Ann is blameworthy for the collapse of the ecosystem, since (1) the collapse of the ecosystem *just is* worse than its continuing to thrive, and (2) Ann's poor quality of will at *t* in relation to the ecosystem's collapsing versus continuing to thrive caused the ecosystem to collapse rather than thrive. In support of 2, note that (a) Ann's poor quality of will at time *t* is process-connected to the collapse of the ecosystem: her poor quality of will is, together with, for instance, Beth's poor quality of will, minimally sufficient for the collapse of the ecosystem (likewise, Ann's and Claire's poor qualities of will are minimally sufficient for the collapse), and this connection remains when we consider more intermediate times. Furthermore, (b) the collapse of the ecosystem is *less secure* at *t* in the closest world where Ann does not have a poor quality of will, i.e., in world w_1, where she has the required quality of will. Here, only one instead of both of the other boat owners would also need to have the required quality of will in order for the ecosystem not to collapse. For similar reasons, the ecosystem's continuing to thrive is *more secure* at *t* in the closest world where Ann has the required quality of will. Thus, Blameworthiness For yields the result that Ann is blameworthy for the ecosystem's collapsing rather than continuing to thrive. The same applies to Beth and Claire.

However, Ann might argue that H_{large} is not the relevant possibility horizon. As we saw above, she might defend herself by saying something like the following: "given that Beth and Claire did not care . . ." There is a straightforward reading of this statement in terms of which worlds should be included in the relevant possibility horizon: Ann is saying, essentially, that we should hold fixed that the others did not care about the lake and treat this as a background condition. If we comply with this request, we get a much smaller possibility horizon, H_{small}. This possibility horizon contains just two possible worlds: the actual world and a world where Ann has the minimally required quality of will. The ecosystem collapses in both worlds since Beth and Claire continue to use the hazardous paint.

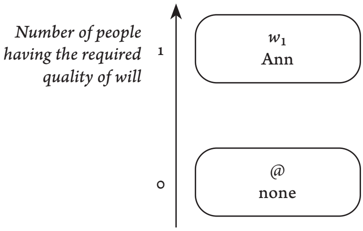


FIGURE 7 Possibility horizon H_{small}

Within H_{small}, the collapse of the ecosystem is *just as secure* at *t* in the actual world as it is in the closest world where Ann has the required quality of will—namely, infinitely secure. Thus, condition b of Causation is not satisfied. Given H_{small}, which holds fixed the motivations of the other boat owners, Blameworthiness For therefore gives the verdict that Ann is *not* blameworthy for the collapse of the ecosystem.

This suggests that our conflicting intuitions about The Lake can be explained as the result of employing two different possibility horizons: employing the large possibility horizon H_{large} yields the result that Ann (as well as Beth and Claire) is individually blameworthy for the collapse of the ecosystem. Employing the small possibility horizon H_{small}, which treats Ann's quality of will as variable while holding the quality of will of the other boat owners fixed, yields the result that Ann is *not* individually blameworthy for the collapse of the ecosystem. Thus, it matters to our assessment of Ann's *individual* responsibility whether we consider a possibility horizon that treats her as part of a group where each member's poor quality of will is a candidate cause, or whether we instead consider a possibility horizon that treats her poor quality of will as the *only* candidate cause. The conflicting intuitions are explained, not as arising from two kinds of entities that can be blameworthy—the group and the individual—but instead as arising from two different perspectives we can take when we are assessing the blameworthiness of an individual.

Of course, we cannot say that Ann both is and is not individually blameworthy for the collapse of the ecosystem. We need to single out either H_{large} or H_{small} as being the relevant possibility horizon for assessing Ann's blameworthiness. We have already suggested in Relevant Possibilities for Blame that we should treat it as a relevant possibility that everyone involved could have had the required quality of will. If this is accurate, the larger possibility horizon is the correct one. We will now support this with two considerations.

First, given that it is important to understand what brought about the collapse of the ecosystem, we have reasons to widen and adjust H_{small}, while H_{large} can do the job satisfactorily. If we address Ann alone, her defense that her lack of care did not matter seems persuasive. At least, it seems to be an open question whether we should consider the smaller possibility horizon she insists on, or the larger one. However, Beth could also argue in the same way that *she* is not blameworthy for the collapse of the ecosystem. And so could Claire. The situation is symmetrical, so if the defense is open to one, it is open to all. If we accept Ann's defense and follow this argument to its logical conclusion, we end up concluding that neither Ann, nor Beth, nor Claire is a cause of the collapse of the ecosystem, and therefore none of them is blameworthy. At this point, we might suspect that something has gone wrong. Surely, we might think, some of the boat owners must be blameworthy for the collapse of the ecosystem. After all, it seems clear that *their* lack of care for the ecosystem played a crucial role in bringing about its collapse.[44] So we face two (related) puzzles: one concerning causation and one concerning blameworthiness. When facing such puzzles, we think there are reasons to reconsider one's choice of possibility horizon:

> *Widening and Adjusting:* If it is important to understand what brought about *X*, and if the causes of *X* are unsatisfactorily explained, we have reasons to look for more possibilities to include in the possibility horizon under consideration and to scrutinize the possibilities we have already included.

[44] See Björnsson, "Joint Responsibility Without Individual Control," 192–93.

If we accept Ann's defense, and the similar defenses of the other boat owners, we find no cause of the collapse and no one who is blameworthy. If it is important to understand what brought about the collapse of the ecosystem, this gives us reason to turn to the larger possibility horizon H_{large}. Within H_{large}, we do find a satisfactory explanation—namely, that all three boat owners are causes of and blameworthy for the collapse of the ecosystem.

Importantly, the reason to widen and adjust our possibility horizon is not grounded in the fact that widening the possibility horizon allows us to hold Ann, Beth, and Claire responsible for the collapse, or say that they caused the collapse. If so, the reasoning would be circular. Rather, the reason to widen and adjust our possibility horizon is grounded in the fact that it is important to find out what brought about the collapse of the ecosystem. Maybe we seek to ensure that similar things do not happen in the future, or maybe we seek to find out who, if anyone, is to blame for the collapse of the ecosystem. If it would turn out upon closer scrutiny—even after widening the possibility horizon—that only Ann caused the collapse, or that none of them did (perhaps something else entirely caused the collapse), this is the conclusion we should accept.

Second, the smaller possibility horizons that Ann, Beth, and Claire must appeal to in their defenses do not fit together. As stated before, if we address Ann alone, her defense seems persuasive. However, this defense is only open to Ann as long as we treat Beth and Claire's poor quality of will as fixed. Ann's defense presupposes that *her* poor quality of will is a candidate cause, while Beth's (and Claire's) poor quality of will is a mere background condition. Beth's defense, on the other hand, presupposes that *her* poor quality of will is a candidate cause, while Ann's (and Claire's) poor quality of will is a mere background condition. In order to accept Ann, Beth, and Claire's defenses, we thus have to adopt one perspective when evaluating whether Ann is blameworthy, a different perspective when evaluating whether Beth is blameworthy, and a third when evaluating Claire's blameworthiness. But these perspectives do not fit together. We cannot at the same time treat, e.g., Ann's quality of will as a candidate cause and as a mere background condition.[45] Given the social function of blame, it should be possible to assess blameworthiness from a single perspective that everyone involved in the situation can accept, and that can be used to assess the blameworthiness of everyone involved. This disqualifies H_{small} since it does not provide such a perspective.

We should point out that there would be no problem of possibility horizons not fitting together if we were to consider a possibility horizon where

[45] Jamieson proposes a similar argument in "When Utilitarians Should Be Virtue Theorists," 176.

the poor quality of will of *each* of Ann, Beth, and Claire was treated as a mere background condition. Thus, there are in fact two ways to treat Ann, Beth, and Claire equally: treating them all as candidate causes (as in H_{large}), or treating them all as mere background conditions. If we, for instance, were to evaluate whether the company producing the hazardous paint is blameworthy for the collapse of the ecosystem, we might for the sake of simplicity treat the poor quality of will of Ann, Beth, and Claire as a mere background condition and instead focus on the quality of will of the company. However, when the question at issue is whether, e.g., Ann is to blame for the collapse of the ecosystem, there is a compelling reason not to treat Ann's poor quality of will as a mere background condition: doing so would *prejudge* the question of whether she is blameworthy by not treating her poor quality of will as a candidate cause at all. Furthermore, we would be failing to treat Ann as an agent if we were to reject the possibility that she could have had the required quality of will. Thus, when Ann's blameworthiness is at issue, we have to treat Ann's poor quality of will as a candidate cause; and then we also have to treat Beth's poor quality of will and Claire's poor quality of will as candidate causes.

We therefore conclude that H_{large} is the relevant possibility horizon. Thus, Blameworthiness For entails that Ann is directly individually blameworthy for the collapse of the ecosystem. By contrast, writers like Held and Wringe would argue that Ann is only *in*directly individually blameworthy for this outcome.[46] On their view, individual blameworthiness in cases like The Lake derives from the blameworthiness of the group. This presupposes that unstructured groups—i.e., groups that lack collective intentions or formal decision procedures—also can be blameworthy. Blameworthiness For does not presuppose this controversial idea. Still, our view captures Held's and Wringe's crucial insight that individual blameworthiness disappears from the scene if we lose sight of the group. For instance, if we fail to consider the possibility that each of Ann, Beth, and Claire could have cared enough about life in the lake, we may end up basing our assessment of Ann's blameworthiness on H_{small}, according to which Ann is not blameworthy for the collapse of the ecosystem.

6. Conclusion

We have suggested that Blameworthiness For together with Causation gives an accurate account of when you are blameworthy for an action, omission, or

[46] See Held, "Can a Random Collection of Individuals Be Morally Responsible?"; and Wringe, "Collective Obligations."

outcome. This account captures the intuitive idea that you are blameworthy for something if it happened because you did not care enough.

One virtue of the account is that it gives the right verdict in a wide range of cases. These include cases of forgetting, making a negative difference to a nevertheless good result (as when Sally did not pay attention to the chili recipe), or doing an action with runaway consequences (like Tragedy). They also include cases where we disregard irrelevant possibilities (like when we think the queen of Sweden did not cause Selma's flowers to die), omission cases (like Indifferent John), switching cases (like Trolley Trouble), (early) preemption cases (like Backup Billy), Frankfurt-style cases (like Backup Neuroscientist Billy), and collective harm cases (like The Lake). In addition, our account gives the right verdict in a wide range of other cases which we do not have space to discuss here, including cases of overdetermination, late preemption, and double prevention, as well as collective harm cases without a threshold.[47]

Another virtue of the account is that it can explain the conflicting intuitions we have about some cases. In collective harm cases like The Lake, for instance, it explains why it might be tempting to accept Ann's defense that we should not blame her: if we treat it as a background condition that Beth and Claire did not care, as Ann insists we should, Ann is correct that *her* lack of care did not cause the collapse. At the same time, our account also explains why it seems that Ann, Beth, and Claire *are* individually blameworthy for the collapse: when we treat it as a relevant possibility that each of them could have had the required quality of will, we find that each of them is a cause of the bad outcome.

As might be evident by now, our account entails that it matters for blameworthiness which possibilities are relevant. Correspondingly, it matters for our judgments about blameworthiness which possibilities we consider to be relevant. This explains, for example, why our intuitions are torn in The Lake: we are torn between only treating it as relevant that Ann could have had the required quality of will, or treating it as relevant that each of the three could have had the required quality of will. To say something about blameworthiness itself, rather than merely about our judgments, we need to answer the question: Which possibilities are relevant? We have argued that whenever an agent involved in the situation has a poor quality of will, it is a relevant possibility that they could instead have had the required quality of will, and the relevant possibility horizon includes, as a minimum, worlds representing every combination of

[47] For a discussion of such cases, see Gunnemyr, *Reasons, Blame, and Collective Harms*, chs. 12–13.

such possibilities. This means, for example, that the relevant possibility horizon in The Lake is the larger one, and so Ann, Beth, and Claire are all individually blameworthy for the collapse of the ecosystem.[48]

Lund University
gunnemyr@gmail.com

Umeå University
caroline.touborg@gmail.com

REFERENCES

Arpaly, Nomy. *Unprincipled Virtue: An Inquiry into Moral Agency*. Oxford: Oxford University Press, 2002.

Bernstein, Sara. "Moral Luck and Deviant Causation." *Midwest Studies in Philosophy* 43, no. 1 (September 2019): 151–61.

Björnsson, Gunnar. "Being Implicated: On the Fittingness of Guilt and Indignation over Outcomes." *Philosophical Studies* 178, no. 11 (November 2021): 3543–60.

———. "Explaining Away Epistemic Skepticism about Culpability." In *Oxford Studies in Agency and Responsibility*, edited by David Shoemaker, 141–64. Oxford: Oxford University Press, 2017.

———. "Explaining (Away) the Epistemic Condition on Moral Responsibility." In *Responsibility: The Epistemic Condition*, edited by Philip Robichaud and Jan Willem Wieland, 146–62. Oxford: Oxford University Press, 2017.

———. "Joint Responsibility Without Individual Control: Applying the Explanation Hypothesis." In *Moral Responsibility: Beyond Free Will and*

[48] We would like to thank Gunnar Björnsson, Erik Carlson, Anton Emilsson, Jens Johansson, and Per-Erik Milam for their close reading and valuable comments, and we are grateful to our colleagues at the Lund Gothenburg Responsibility Project. We are also grateful to audiences at the Group Agency and Collective Responsibility workshop in Flensburg, the Group Agency and Collective Responsibility workshop in Gothenburg, the Higher Seminar in Practical Philosophy at Lund University, the conference Social Ontology 2020, and the Higher Seminar in Practical Philosophy at Uppsala University. We would especially like to thank Per Algander, Franz Altner, Emil Andersson, Olle Blomberg, Dan Egonsson, Anna Folland, Frits Gåvertsson, Niels de Haan, Giulia Lasagni, Samuel Lee, Marianna Leventi, Anne-Sofie Munk Autzen, Carlos Núñez, Anna Nyman, Grace Paterson, Andrew Peet, Björn Petersson, Matthew Rachar, Wlodek Rabinowicz, Toni Rønnow-Rasmussen, Andrew Reisner, David Schweikard, Matthew Talbert, Alexander Velichkov, Marta Johansson Werkmäster, Jakob Werkmäster, and Michael Wilby. Finally, we would like to thank an anonymous reviewer from this journal.

Determinism, edited by Jeroen van den Hoven, Ibo van de Poel, and Nicole Vincent, 181–99. Dordrecht: Springer, 2011.

Bratman, Michael. *Shared Agency: A Planning Theory of Acting Together*. Oxford: Oxford University Press, 2014.

Dowe, Phil. *Physical Causation*. Cambridge: Cambridge University Press, 2000.

Fischer, John Martin, and Mark Ravizza. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press, 1998.

Foot, Philippa. "The Problem of Abortion and the Doctrine of Double Effect." *Oxford Review* 5 (1967): 5–15.

Frankfurt, Harry G. "Alternate Possibilities and Moral Responsibility." *Journal of Philosophy* 66, no. 23 (December 1969): 829–39.

French, Peter. *Collective and Corporate Responsibility*. New York: Columbia University Press, 1984.

Gilbert, Margaret. *Joint Commitment: How We Make the Social World*. Oxford: Oxford University Press, 2014.

Gunnemyr, Mattias. *Reasons, Blame, and Collective Harms.* PhD Thesis, Lund University, 2021. https://portal.research.lu.se/en/publications/reasons-blame-and-collective-harms.

Hall, Ned. "Structural Equations and Causation." *Philosophical Studies* 132, no. 1 (January 2007): 109–36.

———. "Two Concepts of Causation." In *Causation and Counterfactuals*, edited by John Collins, Ned Hall, and L. A. Paul, 225–76. Cambridge, MA: MIT Press, 2004.

Hart, H. L. A., and Tony Honoré. *Causation in the Law*. Oxford: Clarendon Press, 1985.

Held, Virginia. "Can a Random Collection of Individuals Be Morally Responsible?" *Journal of Philosophy* 67, no. 14 (July 1970): 471–81.

Jamieson, Dale. "When Utilitarians Should Be Virtue Theorists." *Utilitas* 19, no. 2 (June 2007): 160–83.

Lewis, David. "Causation." *Journal of Philosophy* 70, no. 17 (October 1973): 556–67.

———. "Causation as Influence." In *Causation and Counterfactuals*, edited by John Collins, Ned Hall, and L. A. Paul, 75–106. Cambridge, MA: MIT Press, 2004.

List, Christian, and Philip Pettit. *Group Agency: The Possibility, Design, and Status of Corporate Agents.* Oxford: Oxford University Press, 2011.

McGrath, Sarah. "Causation by Omission: A Dilemma." *Philosophical Studies* 123, nos. 1–2 (March 2005): 125–48.

McKenna, Michael. *Conversation and Responsibility*. Oxford: Oxford University Press, 2012.

Paul, L. A., and Ned Hall. *Causation: A User's Guide*. Oxford: Oxford University Press, 2013.

Rosen, Gideon. "The Alethic Conception of Moral Responsibility." In *The Nature of Moral Responsibility: New Essays*, edited by Randolph Clarke, Michael McKenna, and Angela M. Smith, 65–87. Oxford: Oxford University Press, 2015.

Sartorio, Carolina. *Causation and Free Will*. Oxford: Oxford University Press, 2016.

———. "Causes as Difference-Makers." *Philosophical Studies* 123, nos. 1–2 (March 2005): 71–96.

Schaffer, Jonathan. "Causal Contextualism." In *Contrastivism in Philosophy*, edited by Martijn Blaauw, 35–63. Abingdon: Routledge, 2013.

———. "Causation by Disconnection." *Philosophy of Science* 67, no. 2 (June 2000): 285–300.

———. "Contrastive Causation." *Philosophical Review* 114, no. 3 (July 2005): 327–58.

Searle, John. "Collective Intentions and Actions." In *Intentions in Communication*, edited by Philip R. Cohen, Jerry Morgan, and Martha E. Pollack, 401–15. Cambridge, MA: MIT Press, 1990.

Shoemaker, David. *Responsibility from the Margins*. Oxford: Oxford University Press, 2015.

Smith, Angela M. "Attitudes, Tracing, and Control." *Journal of Applied Philosophy* 32, no. 2 (May 2015): 115–32.

———. "Responsibility for Attitudes: Activity and Passivity in Mental Life." *Ethics* 115, no. 2 (January 2005): 236–71.

Strawson, P. F. "Freedom and Resentment." In *Freedom and Resentment and Other Essays*, 1–28. Abingdon, UK: Routledge, 2008.

Talbert, Matthew. "Moral Competence, Moral Blame, and Protest." *Journal of Ethics* 16, no. 1 (March 2012): 89–109.

Thomson, Judith Jarvis. "Killing, Letting Die, and the Trolley Problem." *Monist* 59, no. 2 (April 1976): 204–17.

Touborg, Caroline Torpe. *The Dual Nature of Causation: Two Necessary and Jointly Sufficient Conditions*. PhD Thesis, University of St Andrews, 2018. http://hdl.handle.net/10023/16561.

———. "Hasteners and Delayers: Why Rains Don't Cause Fires." *Philosophical Studies* 175, no. 7 (July 2018): 1557–76.

Van Inwagen, Peter. "Ability and Responsibility." *Philosophical Review* 87, no. 2 (April 1978): 201–24.

Wallace, R. Jay. *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press, 1994.

Watson, Gary. "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme." In *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*, edited by Ferdinand Schoeman, 256–86. Cambridge: Cambridge University Press, 1988.

Woollard, Fiona. *Doing and Allowing Harm*. Oxford: Oxford University Press, 2015.

Wringe, Bill. "Collective Obligations: Their Existence, Their Explanatory Power, and Their Supervenience on the Obligations of Individuals." *European Journal of Philosophy* 24, no. 2 (June 2016): 472–97.

Zimmerman, Michael J. "Taking Luck Seriously." *Journal of Philosophy* 99, no. 11 (November 2002): 553–76.

Journal of Ethics and Social Philosophy
VOL. 24, NO. 1 · FEBRUARY 2023

https://doi.org/10.26556/jesp.v24i1.2145

THE RED MIST

Maxime Lepoutre

In her 2018 autobiography, *Rage Becomes Her,* Soraya Chemaly memorably recalls one of the first times she experienced sexual harassment: "In cases like these, I usually freeze—like many of us do. My brain and heart race to determine *the nature of the risk* and calibrate my response." She continues:

> However, on the day when I was fourteen, and the man grabbed my arm, I didn't freeze; I punched him hard in his windpipe. This was my first memory of *blinding visceral rage* in these circumstances.[1]

Chemaly's rage at the harasser exemplifies a central feature of anger: namely, that anger often comes at an epistemic cost. Her anger, indeed, is experienced as "blinding." And it blinds her, more specifically, by distracting her from the "risks" involved in lashing out. In everyday life, we have a term for this epistemic cost of anger: we call it the "red mist."

The idea that anger gives rise to a red mist constitutes one of the most longstanding objections to this emotion. Seneca, for instance, condemns anger on the grounds that it seems to involve a "departure from sanity," which "does not disturb the mind so much as take it by force."[2] More recently, evidence from experimental psychology has reignited these epistemic concerns. Drawing on this evidence, Glen Pettigrove pessimistically concludes that anger "adversely affect[s] judgment."[3]

To critics such as these, anger's red mist is concerning for two related reasons. First, and most obviously, it suggests that anger might be detrimental to knowledge, especially knowledge of what risks one faces. Second, *because* of this epistemic deficiency, anger may lead to bad consequences. Chemaly's furious response to being harassed, for example, might have led to a catastrophic escalation of the situation.

[1] Chemaly, *Rage Becomes Her*, 122, emphases added.

[2] Seneca, *Moral and Political Essays*, iii.1.

[3] Pettigrove, "Meekness and 'Moral' Anger," 122; see also Nussbaum, *Anger and Forgiveness*, 38.

Some are unfazed by this concern. Defenders of anger commonly observe that, even if anger comes at an epistemic cost, it also yields significant epistemic benefits. In particular, anger highlights injustices that may otherwise have been overlooked.[4] Moreover, partly *because* it highlights injustices, anger often motivates actions of opposition to injustice that may result in good, rather than bad, consequences.

I am sympathetic to this response. But it is nevertheless limited in two important ways. First, absent further development, this response does not establish that anger is epistemically good, overall. *Perhaps* the epistemic benefits of anger outweigh its epistemic costs. But the opposite could conceivably be true. This, in fact, is typically what critics such as Seneca or Pettigrove believe. So, this response is unlikely to convince those who take issue with anger's epistemic costs.

However, there is a second and more fundamental problem. The problem is that this response leaves unchallenged the critics' central assumption: namely, that the epistemic costs of anger are necessarily a bad thing. Chemaly's testimony is intriguing, not just because it highlights anger's epistemic cost—its blinding red mist—but also, crucially, because it suggests that this cost has moral value. The red mist enables her to take a stand against an injustice, in a way that protects her sense of dignity and self-respect.

It is this suggestion that I wish to articulate and develop here. I will argue that the epistemic costs of anger are intimately bound up with one of its core moral benefits. Specifically, the red mist contributes to protecting the dignity and self-respect of those who experience it. Thus, anger is useful not just because of the knowledge that it facilitates, but also because of the knowledge that it prevents.

To be clear, this argument does not purport to show that the red mist is always a morally good thing, overall. In fact, I will highlight a number of circumstances where it is not. But it nevertheless establishes something important. It shows that it is far more difficult to move from the epistemic costs of anger to a moral rejection of this emotion than critics have usually supposed. This is for two reasons. First, my argument reveals an overlooked moral benefit of anger, which we must weigh against anger's epistemic costs. Second, this moral benefit is significant, because it pertains to the enjoyment of a fundamental good: namely, self-respect.

The rest of this paper is organized as follows. Section 1 defines anger, and clarifies the nature of its epistemic benefits and costs. In particular, it underscores what is perhaps anger's most central epistemic cost: that anger makes

[4] See, e.g., Lorde, "The Uses of Anger"; Frye, *The Politics of Reality*, ch. 5; Lepoutre, *Democratic Speech in Divided Times*, ch. 2.

risk less visible to the angry person. Next, section 2 argues that the perception of certain forms of risk can be deeply harmful to self-respect. Putting these two conclusions together, section 3 demonstrates that anger's red mist contributes meaningfully to protecting the self-respect of those who live under certain forms of risk. Lastly, section 4 qualifies this argument by considering more closely (a) the conditions under which the red mist protects self-respect, (b) the extent to which this protective role could be performed by other emotions than anger, and finally (c) whether this role comes at too high a price.

1. The Epistemic Costs of Anger

What is anger? Like all emotions, anger is a state characterized first of all by a physiological dimension. In other words, anger is associated with certain bodily feelings.[5] For example, anger can make our skin feel hot, our heart race, our breath quicken, our voice tremble, etc.

Anger also possesses a motivational dimension. Feeling angry typically involves being moved to do something. The angry person is disposed, in some broad sense, to act against the object of their anger. Since anger's natural object is injustice or wrongdoing, this means that anger characteristically involves a motivation to oppose what we perceive to be injustices or wrongdoings.[6]

But anger, crucially, is not merely a physiological and a motivational state. It is also a cognitive state, in that it modifies the way we represent the world around us. How, exactly, does anger alter our representations?

To answer this question, we need to step back and consider how emotions in general affect our representations. The answer, in short, is that emotions are sources of salience. As philosophers have widely argued, emotions highlight particular features of our environment, and focus our attention on them.[7] This salience role is absolutely essential to our cognition. We typically inhabit immensely complex environments that bombard us with information. By selecting some features, and making them salient to us, emotions allow us better to navigate the world, and to make sense of it.

Different emotions can be distinguished, in part, based on what they make salient to us. Grief makes loss more visible to us. Hope highlights reasons to be optimistic about our aims and projects. As for fear, it makes salient potential

[5] Deonna and Teroni, *The Emotions*, 2.

[6] Nussbaum, "Transitional Anger," 45–48; Bell, "Anger, Virtue, and Oppression," 168–69. Some philosophers refer to the emotion I am describing as "moral" anger, while allowing that some "non-moral" forms of anger (e.g., frustration) can have objects other than injustice or wrongdoing. See Pettigrove, "Meekness and 'Moral' Anger," 357–58.

[7] On the salience role of emotions, see Elgin, "Emotion and Understanding," 43–46.

sources of danger. As Elgin observes, for instance, if I am afraid as I walk home at night, and I hear footsteps behind me, my attention will naturally be drawn to the features of this situation that could make it dangerous.[8]

What, then, does *anger* make salient to us? If fear represents the world around us as dangerous, anger, by contrast, represents it as unjust.[9] So anger, as a source of salience, highlights possible sources of injustice or wrongdoing.

This, to anger's defenders, is undoubtedly its most important epistemic benefit. In non-ideal social and political settings, injustices are not always plainly visible. For example, political elites sometimes deploy spurious ideologies to make injustices appear morally legitimate (as, say, when the ideal of meritocracy is used to rationalize what are in fact deeply unjust inequalities). In settings such as these, the experience of anger is helpful because, to use Audre Lorde's memorable phrase, it casts a "spotlight" on wrongs that may otherwise have been overlooked.[10] Thus, so long as we live in societies that contain grave injustices, as well as epistemic obstacles that make it difficult to fully recognize these injustices, anger's salience function constitutes an important consideration in its favour.

This epistemic argument for anger does require some qualification. In particular, it requires distinguishing between warranted (or rational) anger, and unwarranted (or irrational) anger.[11] At least two conditions must be satisfied for anger to be warranted. First, anger is warranted only if the object or situation it is directed at actually involves an injustice or moral wrong. If I am angry that you broke your promise to me, but you did not in fact break your promise, then my anger is unwarranted and irrational.

The second factor has to do with intensity. Anger can vary in intensity. It can take mild and non-violent forms, such as minor irritation or moderate indignation. But it can also be far more intense, as exemplified by Chemaly's violent rage at the sexual harasser. Intense anger is not necessarily unwarranted or irrational. Rather, what is crucial, as Lee McBride III notes, is that anger must be proportionate to the severity of the injustice it is responding to.[12] Thus rage may be a rational response to grave injustices, such as slavery. But it is obviously not a fitting response to a minor moral violation, such as your forgetting to

[8] Elgin, "Emotion and Understanding," 43–44.

[9] See, e.g., Frye, *The Politics of Reality*, 85–86; Srinivasan, "The Aptness of Anger," 128–29 Callard, "The Reason to Be Angry Forever."

[10] Lorde, "The Uses of Anger," 278. See also Lepoutre, *Democratic Speech in Divided Times,* ch. 2.

[11] I will be using the terms "warranted anger" and "rational anger" interchangeably.

[12] McBride, "Anger and Approbation," 5–6. See also Cogley, "A Study of Virtuous and Vicious Anger," 202–3. These two conditions are necessary for anger to be warranted. But they are not intended to be sufficient. For further ways in which anger can go wrong, see, e.g., McBride, "Anger and Approbation," 5.

return my favourite book. Conversely, minor irritation is not a warranted or rational response to the grave injustice of slavery.

What does this mean for anger's epistemic value? Defenders of anger generally agree that, for anger to be epistemically valuable, it must warranted or rational.[13] When anger is unwarranted or irrational—e.g., when it is directed at a situation devoid of injustice, or when it is disproportionate relative to how severe the injustices are—it tends to lack epistemic value. Indeed, unwarranted anger highlights injustices where there are none, or fixes our attention on injustices to a greater degree than is justified by their severity.

But, even with this qualification, critics of anger insist that it faces a deeper problem: even when anger *is* warranted or rational, it still comes at an important epistemic cost. This is the flipside of anger's salience role. Salience is always a comparative matter. When we highlight something, we also necessarily place other things in the background of our vision. The upshot for (warranted) anger is that it makes injustices visible, but only at the expense of making other things less visible. To put this slightly differently: the fact that anger performs an epistemically valuable salience function partly explains why, as we saw in the Introduction, anger also involves an epistemically costly "red mist."[14]

Yet the problem is not simply that anger's salience function comes at an epistemic cost. It is, in addition, that *the more* anger draws our attention to injustices, *the more* it draws our attention away from other things. Intense rage fixes our attention on perceived injustices more strongly than mild indignation would. But this same increased fixation also means that, when we are full of rage, we ignore more things, and ignore them more completely, than when we are mildly indignant.

So, despite (warranted) anger's clear epistemic benefits, it is difficult to argue that it is all-things-considered epistemically valuable. The epistemic benefits of anger entail epistemic costs; and when its epistemic benefits increase, the epistemic costs are likely to do so too.

[13] Srinivasan, "The Aptness of Anger," 130–33.

[14] In ordinary usage, the term "red mist" is commonly used to refer to the epistemic costs associated with very intense forms of anger (e.g., blinding rage). It is worth highlighting that, here, I am using the term "red mist" more broadly, to refer to anger's epistemically costly tendency to conceal risk. As the evidence from experimental psychology discussed below shows, this epistemic cost applies not just to very intense anger, but to milder forms of anger as well. However, in practice my emphasis will be on cases of intense anger (e.g., Chemaly's rage). This is because these are the cases where the risk-related epistemic costs of anger are greatest, and where, therefore, the self-respect protecting function I am discussing is most clearly exemplified. Thus, even though I take the "red mist" to be a feature of anger generally, my account will focus predominantly on a particularly intense subset of cases involving the red mist. I am grateful to a reviewer for pressing me to clarify this.

The foregoing concerns about the epistemic costs of anger are not mere theoretical conjecture. Experimental psychologists have widely corroborated the proposition that anger makes some things less visible to the angry person. What is more, they have made this proposition more specific, by identifying precisely *what* anger's red mist tends to make less visible to us. The most consistent result is that experiencing anger makes *risk*—understood as the possibility that a negative consequence will occur—less salient to those who are angry.[15]

Current experimental psychology suggests that there are two principal ways in which anger tends to suppress our perceptions of risk. First, anger can make the probability of a bad outcome less salient to us. Put differently, experiencing anger can make us judge that a bad outcome is less likely to arise than we would otherwise think.[16] Alternatively, anger can make the *badness* of a possible bad outcome less salient. Here, the issue is not that anger makes us underestimate the likelihood of the outcome. Rather, it is that anger makes us underestimate how problematic it would be if it took place.[17] Either way, experimental results strongly suggest that anger's red mist makes us less attentive to the risks we are exposed to than we would normally be.

But we needn't rely exclusively on the theoretical accounts provided by philosophers, nor on experimental studies taking place in non-political contexts, for the insight that anger can suppress our attentiveness to risk. Indeed, this insight is also familiar from the testimony of political actors, when they express or report on their anger. Chemaly, recall, explicitly notes that when she experienced rage towards the man who harassed her, that intense anger short-circuited her usual assessment of risk. Nor is Chemaly's case unusual. As we will see in greater detail in section 3, the anti-slavery abolitionist Frederick Douglass once observed that, when his fury boiled over, he found himself acting "heedless of consequences."[18]

We could readily add to these examples. But what leaves no doubt is that, just as critics of anger suggest, anger—and intense anger in particular—comes with an unmistakable epistemic cost. Anger diminishes or suppresses our perceptions of risk. And it does so, not just in carefully controlled experimental settings, but also in real-world political environments where the stakes are high.

[15] See, e.g., Lerner and Keltner, "Fear, Anger, and Risk," 146–59; Hemenover and Zhang, "Anger, Personality, and Optimistic Stress Appraisals," 363–82; Gambetti and Giusberti, "Dispositional Anger and Risk Decision-Making," 7–20.

[16] Lerner and Keltner, "Fear, Anger, and Risk," 154.

[17] Hemenover and Zhang, "Anger, Personality, and Optimistic Stress Appraisals," 370; Gambetti and Giusberti, "Dispositional Anger and Risk Decision-Making," 14.

[18] Douglass, *My Bondage and My Freedom*, 104.

2. Feeling at Risk

In this section, and the following, I want to suggest that this epistemic cost of anger nevertheless plays a positive and important moral function. Specifically, it helps to protect the self-respect, or dignity, of those who are subjected to risk.[19]

To see why, we first need to examine why feeling at risk might be detrimental to self-respect. Self-respect, as I am understanding this notion, consists in one's sense of oneself as having a basic and equal moral standing, in virtue of which one is owed respect.[20] Why might feeling at risk impair self-respect, thus understood?

The first thing to note is that *not all* risk impairs self-respect. To begin, some risks are voluntarily pursued or chosen because people find them exciting (e.g., the risks associated with extreme sports). Experiencing chosen risks such as these seems intuitively unproblematic for one's sense of moral worth.[21] Moreover, not all categories of unchosen risk are problematic for self-respect. Suppose I love the feeling of riding a motorcycle at full speed, but would prefer if this activity were not so risky. In this case, the experience of risk attached to riding my motorcycle is, in an important sense, unchosen: I perform the activity *despite* the risk. Still, it seems unclear why exposure to this risk would diminish my sense that I am owed respect in virtue of my basic moral standing.

My argument here will therefore be restricted to a specific category of risk, that is particularly salient in non-ideal conditions: risk that sustains or consolidates injustice ("injustice-sustaining risk"). Real-world politics is non-ideal not only because it involves grave injustices, but also because attempts at dismantling those injustices are often risky. Chemaly's story vividly exemplifies this kind of risk: standing up to unjust sexual harassment exposes Chemaly to the risk of backlash or violent escalation. Likewise, as Davin Phoenix argues, if a person of colour acts out against racial injustice in the US, "[they] risk being labelled a threat, targeted, monitored, and brought down by agents of the system [they] challenge."[22]

This injustice-sustaining risk threatens to impair self-respect via two mechanisms. First, perceiving such risks can deter actions aimed at opposing injustices. If taking a stand against injustice exposes me to severe violence, I may simply decide *not* to act. Chemaly is explicit about this. Awareness of risk, she suggests, normally leads her to "calibrate her response." In practice, she later

[19] I am using the terms "self-respect" and "dignity" interchangeably here.

[20] I am drawing on Robin Dillon's influential characterization of self-respect (in Dillon, "Respect,"). For other characterizations that emphasize the importance of recognizing one's moral equality, see also Shelby, *Dark Ghettos*, 98; and Bell "Against Simple Removal," 784.

[21] Tomasi, *Free Market Fairness*, 80; Baderin and Barnes, "Risk and Self-Respect," 1430–32.

[22] Phoenix, *The Anger Gap*, 17.

indicates, this means that she, like many other women, often ends up "biting her tongue" when faced with unjust harassment.[23]

This matters, because acts of opposition to injustice are intimately bound up with the preservation of self-respect. To see this, consider first that injustice has an expressive dimension. In her influential analysis of speech-acts, Mary Kate McGowan observes that an action typically presupposes its own appropriateness.[24] What this means is that injustices, and the actions that sustain them, express their own appropriateness. When a group violates the rights of another, for instance, that violation expresses, by presupposing it, the proposition that this group may appropriately be violated in this way. More generally, then, injustices express disregard for the moral standing of their targets.

This has an important implication for acts of resistance to injustice. If injustice expresses disregard for the moral status of its targets, acts of resistance by contrast express the rejection of this proposition. In other words, the act of resisting injustice helps to reaffirm one's equal moral standing when one has been wronged. Hence, Tommie Shelby concludes that "we surrender or sacrifice our self-respect when we acquiesce to mistreatment or when we suffer indignities in silence."[25]

My point here is not that perceiving injustice-sustaining risks will *always* deter actions aimed at opposing the relevant injustices. As we will see in section 4, it *is* possible to recognize such risks and nevertheless take action. But injustice-sustaining risk nevertheless makes such actions less likely. The risk, for example, that one will be subjected to a violent backlash creates strong pressure not to take a stand against an injustice one has suffered. Some may be able to withstand this pressure. But many others are likely to respond (as Chemaly notes) by "biting their tongue." Insofar as this is the case, injustice-sustaining risk poses a threat to self-respect.

So far, my argument has been that perceiving injustice-sustaining risk threatens self-respect indirectly, via its deterrent effect on action. But this category of risk also poses a second, and arguably more direct, threat to self-respect. Independently of its impact on action, the existence of risks that sustain injustices can *aggravate* the demeaning message conveyed by those injustices.[26]

Take, once more, the case of sexual harassment. Acts of sexual harassment already express a degrading attitude—e.g., the attitude that it is permissible to

[23] Chemaly, *Rage Becomes Her,* 128.

[24] McGowan, *Just Words*, 140.

[25] Shelby, *Dark Ghettos*, 99.

[26] For the idea that the existence of risk might express a degrading message, see, e.g., Baderin and Barnes, "Risk and Self-Respect," 1424.

treat women as mere sexual objects. But the fact that women who resist such harassment risk incurring verbal or physical violence *adds* to this degrading message. It sends the message, not only that women may permissibly be treated as sexual objects, but that—in addition—they have no right to protest against, or otherwise oppose, this treatment. Awareness of this risk therefore exacerbates the original insult faced by victims of sexual harassment—and with it, the potential damage to their sense of self-worth. This is once more explicit in Chemaly's testimony. The experience of such risk, she suggests, "is how we come to accept the harsh fact of our violability."[27]

Let us take stock. I have argued that perceiving injustice-sustaining risk threatens self-respect in two ways. For one thing, it makes it more difficult—and so, less likely—that one will act against injustices to which one is subjected. From the perspective of self-respect, this matters because taking a stand against injustices is a crucial way of reasserting one's basic moral standing. Yet perceiving risk is a problem even for those who do take action. This is because injustice-sustaining risk exacerbates the demeaning message associated with the injustices it consolidates—not least, by denying one's moral entitlement to protest or resist those injustices. Thus, to be aware, not just of an injustice, but also of the risk that sustains it, is to be aware of even greater disregard for one's moral status.

3. THE RED MIST AND SELF-RESPECT

We can now appreciate a significant value associated with anger's epistemic cost, or red mist. Perceiving that we live under risk (in particular, injustice-sustaining risk) can pose a deep threat to our self-respect. But, importantly, we have also seen that anger's red mist makes risk less visible to us. Putting these two

[27] Chemaly, *Rage Becomes Her,* 123. Note that perceiving this exacerbated message of disrespect can aggravate the injury to one's self-respect in at least two complementary ways. In the first place, simply *understanding* the disrespectful message associated with injustice-sustaining risk is hurtful. Indeed, it is hurtful to realize, via that understanding, that others do not respect one's basic moral standing (and more specifically, one's right to stand up to injustice). But there is arguably a second and deeper possible harm to self-respect. In some cases, as the Chemaly quote suggests, perceiving the exacerbated message of disrespect associated with injustice-sustaining risk is hurtful because it leads to some degree of *acceptance* of that message. This second mechanism may not obtain in all cases: there may be cases where one perceives the risk-induced message of disrespect yet does not accept this message to any degree. Nevertheless, it is a possible further harmful consequence associated with registering the exacerbated message of disrespect. I am grateful to a reviewer for helping me clarify this distinction.

conclusions together, the upshot seems to be that anger's red mist helps the angry person maintain or recover their sense of self-respect.[28]

It does so, more specifically, by blocking the two expressive mechanisms through which the perception of risk assaults our sense of dignity. First, we have seen that risk can exacerbate the disrespectful message associated with injustices. By concealing risk, the red mist shields us from experiencing this added insult. Second, insofar as the red mist blinds us to the risks we face, it diminishes our sense of vulnerability—which, for many of us, makes it more likely that we will take action against injustices we face, and thereby reassert our dignity. So, the red mist both helps to reduce our exposure to a degrading message, *and* facilitates a dignifying counter-message. By suspending her consideration of risk, for example, Chemaly's "blinding visceral rage" simultaneously shields her from its degrading message (as a symbol of her "violability") and emboldens her to fight back, and thereby reassert her dignity.

Now, for any purported benefit of anger, we can ask: "How important is this benefit?" I wish to suggest that, in the case of the red mist's self-respect protecting function, the answer is "Very important."

The first reason is simply that self-respect is of great moral significance. Indeed, possessing a sense of oneself as a moral equal, who is owed respect, is crucial to living a good and meaningful life. In his influential discussion of self-respect, Shelby suggests as much. He takes self-respect to be an "intrinsic" good, without which one is bound to live "an impoverished life."[29] Thus, for Shelby, it is often worth sacrificing material gain, and other important ingredients of one's welfare, to maintain one's self-respect.[30]

This is by no means an unusual view. As Robin Dillon observes, there is "surprising agreement among moral and political philosophers" that self-respect is "essential to the ability to live a satisfying, meaningful, flourishing life—a life worth living."[31] Nor does this value depend on having a particular or idiosyn-

[28] The claim that anger can protect self-respect can also be found in Bell, "Anger, Virtue, and Oppression," 168; McBride, "Anger and Approbation," 9. My argument develops this position in two significant ways. First, it explains in greater detail *why* anger protects self-respect. Second, it shows that this benefit supervenes on what is typically considered a negative feature of anger: namely, its epistemic cost.

[29] Shelby, *Dark Ghettos*, 99–100.

[30] Shelby, *Dark Ghettos*, 99–100. To say that self-respect is crucial to living a good and meaningful life does not imply that the value of self-respect reduces to its contribution to welfare and well-being. As Susan Wolf argues, what makes life meaningful is not reducible to considerations of well-being. See Wolf, *Meaning in Life and Why It Matters*, 1–7.

[31] Dillon, "Respect," sec. 4. As Dillon also notes, self-respect may be valuable in ways that go beyond its place in a meaningful and good life. According to Kant, for instance, there is a *moral duty* to respect oneself. While I am sympathetic to this view, the existence

cratic conception of what constitutes a good and meaningful life. Rawls, for instance, famously argues that self-respect is a "primary good" ("perhaps the most important primary good" there is).[32] In other words, it is something that one has reason to want, and without which it may be difficult to live a good and meaningful life, whatever one's conception of the good and meaningful life may be.

The second reason why the function at hand constitutes an important benefit has to do with frequency: not only does this function protect something that is of great moral significance, but it *often* protects it.

As we saw in section 2, the experience of risk is especially likely to impair self-respect in cases where risk sustains injustices. Now, crucially, these cases are very common in non-ideal conditions: oppressed groups often find themselves in tragic situations where, on the one hand, they are subjected to grave injustices; and, on the other hand, it is extremely risky to take action against these injustices.[33] Standing up to sexual harassment, for example, often comes with a risk of escalation. Likewise, protesting or rioting against police violence itself often involves a risk of subjection to police violence. Thus, there is reason to think that, in the real world, anger's red mist can frequently help to preserve self-respect: it offers protection against a threat that is rife in non-ideal conditions.

Even so, for all that I have said about the value of anger's red mist in non-ideal circumstances, this defence might still seem overly abstract. After all, I have so far only provided one concrete example of this value—namely, Chemaly's rage towards her harasser. To further illustrate this account, and provide a more concrete sense of the red mist's importance in non-ideal settings, I therefore wish to conclude this section by demonstrating how it makes sense of a further piece of testimony—Frederick Douglass's famous recollection of when, still a slave, he finally fought back against the slave-breaker Covey:

> Whence came the daring spirit necessary to grapple with a man who, eight-and-forty hours before, could with his slightest word have made me tremble like a leaf in a storm, I do not know.… The *fighting madness* had come upon me, and I found my strong fingers firmly attached to the throat of my cowardly tormentor; as *heedless of consequences*, at the moment, as though we *stood as equals* before the law.… Well, my dear reader, this battle with Mr. Covey—undignified as it was, and as I fear my narration of it is—was the turning point of my "life as a slave."… *It*

of self-regarding duties is controversial. Accordingly, my argument will not depend on accepting it.

[32] Rawls, *A Theory of Justice*, 386.

[33] Flanigan, "From Self-Defense to Violent Protest," 9.

> *recalled to life my crushed self-respect* and inspired me with a renewed determination to be A FREEMAN. A man, without force, is without the essential dignity of humanity.[34]

Douglass's testimony leaves no doubt that this episode restored his "dignity" and "self-respect." How did it do so? The answer revolves crucially around anger and risk perception. Douglass's blinding rage (his "fighting madness") left him acting "heedless of consequences." And this inattentiveness to risk in turn helped resurrect Douglass's self-respect in two ways: first, by allowing him to regard himself as Covey's equal in standing (indeed, Douglass overtly associates the perceived absence of consequences with a sense of equal status); and second, by emboldening him to fight back against his oppressor (where previously, he would have "tremble[d] like a leaf in a storm"). So, in Douglass's testimony, we have another remarkably vivid account of how anger's red mist can play an indispensable role—by concealing the degrading status symbolism associated with risk; and by emboldening him to act out against injustice—in protecting self-respect.

I have argued that the epistemic cost of anger can play a morally valuable function, and that this function is of great significance in non-ideal circumstances such as our own. The core upshot is that, to defend anger, we do not necessarily need to show that the epistemic benefits of anger outweigh its epistemic costs. Because the epistemic costs of anger can perform a morally valuable function, it is at best an open question (to which I return in the next section) whether, overall, they constitute a bad thing for anger.

4. THE LIMITS OF THE RED MIST

None of this means that the value of anger's red mist is without limits. To clarify my argument, the rest of this paper will examine more closely under what conditions anger's red mist protects self-respect; whether it is necessary to protect self-respect; and whether, even if it is necessary, it nonetheless comes at too high a cost.

Let us start with the circumstances under which anger's red mist helps to protect self-respect. One might worry that my account of the red mist's value overgenerates—in other words, that it lends support to intuitively unacceptable instances of anger. Consider a white supremacist who experiences violent rage directed at people of colour, and who lives in a society that legislates strongly against hate crimes. This legislation clearly imposes a risk: anyone who performs a hate crime faces lengthy incarceration. And this risk may well deter

[34] Douglass, *My Bondage and My Freedom*, 103–6, emphasis added.

the white supremacist from acting as they otherwise would. At first sight, my account of the red mist's value might seem to have an implausible implication in this case: it might seem to imply that the white supremacist's rage is a good thing, because the epistemic costs associated with that rage conceal risks, and thereby protect the white supremacist's self-respect.

In fact, my argument for the red mist's value does not extend to the white supremacist, for two reasons. The first is that, in this case, perceiving risk needn't undermine self-respect. My argument, recall, centres on risk *that sustains injustice*. But the risk to which the white supremacist is subjected does not sustain injustice—rather, it serves to uphold justice. This makes a crucial expressive difference. As discussed in section 2, unjust actions and states of affairs implicate the moral inferiority and violability of their targets. By contrast, just actions and states of affairs express the opposite message. They express a message of fundamental moral equality.[35] Indeed, the risks imposed by hate crime legislation express the idea that no one—not the white supremacist, nor their intended victim—should be harmed due to their race, ethnicity, etc. Insofar as this risk expresses the equal moral standing of all, its visibility seems protective of—not detrimental to—self-respect. My analysis therefore does not imply that the white supremacist's rage protects self-respect: their red mist conceals, not a disrespectful message, but rather a message of universal and equal dignity.

Second, even if the white supremacist's red mist *did* contribute to maintaining their self-respect, it would still not follow that it is good overall, or indeed that it is morally equivalent to Chemaly's or Douglass's red mist. This is because this (alleged) benefit would arguably be overridden by countervailing moral costs. For one thing, the white supremacist's red mist is constantly conjoined with an attitude of profound disrespect towards others. That is, their anger is premised on the perception that racial minorities are inferior. Moreover, the white supremacist's red mist is likely to motivate them to act in support of unjust and oppressive norms. This is in stark contrast to Chemaly and Douglass, whose anger emboldens them to *challenge* oppressive norms. Since both factors—the disrespectful attitude; and the oppressive actions–have great moral disvalue, the white supremacist's red mist seems overall bad in this case *even if* we assume (for the sake of argument) that it would preserve their self-respect.

But even with this qualification, my argument for the value of anger's red mist might still seem overstated. Even where people *are* subjected to risks that consolidate grave injustices, anger may not seem necessary to preserve their self-respect. To insist that it is necessary would imply that political figures such

[35] For discussion in the context of anti-discrimination law, see Anderson and Pildes, "Expressive Theories of Law," 1503–75.

as Mohandas Gandhi, Nelson Mandela, and Martin Luther King—who spearheaded struggles against injustice but are often regarded as having repudiated anger—lacked self-respect.[36] Yet this seems clearly false.

Strictly speaking, this observation is correct. It is indeed *possible* for someone who eschews anger to maintain their self-respect despite facing injustice-sustaining risk. There are different reasons why this might be. Perhaps they possess an unusually strong social support network, whose presence allows them to feel worthy of respect despite this risk's derogatory message, and despite the fact that it deters them from acting out against injustices they face. Or, to give another example, perhaps they have an extraordinary sense of self-sacrifice, such that perceiving such risks does not deter them from acting against injustice, and thereby reaffirming their self-respect.

Even so, this observation is compatible with recognizing that subjection to injustice-sustaining risk typically makes it harder to maintain one's self-respect. As I argued in section 2, these risks usually aggravate the demeaning message conveyed by injustices, and create strong pressure not to act out. Both Chemaly and Douglass, recall, vividly describe the pressure that the awareness of such risk placed on their willingness to act, and on their attending sense of dignity. Though withstanding this pressure is possible, it is hard—and, as the above examples suggest, it may require felicitous social circumstances, or rare character traits. In this context, anger's red mist is still helpful: though it may not be strictly necessary for the protection of self-respect, its impact on risk perception nevertheless meaningfully facilitates it.

This initial response may not be sufficient to appease the sceptic. After all, you might think that emotions other than anger could be equally effective at facilitating the preservation of self-respect in the face of risk. Hope seems like a promising candidate here. Jakob Huber argues that hope is capable of motivating political action. And hope, too, does so by altering our perception of the social environment. Notably, hope tends to make good outcomes appear more salient than they otherwise would. This outlook can encourage acts of resistance to injustice that—partly due to the risks they involve—would otherwise have seemed futile and not worth undertaking.[37]

I agree that hope is a valuable political emotion, and that it is valuable, in significant part, because of its capacity to motivate acts of resistance to injustice—acts which, in turn, help reaffirm our self-respect. But this does not undermine my defence of anger's red mist, for several reasons.

[36] Nussbaum, *Anger and Forgiveness*, ch. 7. I am assuming, for the sake of argument, that these three figures actually repudiated anger. But this claim is controversial. For disagreement, see Cherry, "Love, Anger, and Racial Injustice," 157–68.

[37] Huber, "Defying Democratic Despair," 720.

The first reason is more conciliatory. Even if we assume that hope and anger are equally capable of protecting self-respect, and of doing so in the same circumstances, this does not undermine my central contention in this essay. As I explained at the outset, my primary aim has been to challenge the inference from the observation that anger comes at an epistemic cost, to the conclusion that anger is morally undesirable. In response, I have argued that, on closer inspection, this epistemic cost can perform a morally valuable function. This point is not inherently comparative: it is compatible with thinking that *other* emotions can perform this valuable moral function as well.

But we can go further than this first response. There are respects in which anger's self-respect protecting function seems distinctive, such that hope could not fully replace it. To begin, hope and anger can be warranted in different circumstances. There are circumstances where hope is warranted, but anger is not (e.g., hoping, in a context where no injustice has occurred, that my friend likes the gift I have given them). And, more importantly for our purposes, there are circumstances where anger is warranted, but hope is not. When the good outcome one desires (e.g., the eradication of injustice) is impossible to achieve, hope is arguably unwarranted.[38] Anger, however, can in principle be warranted in these "hopeless" cases. Whether we are warranted in feeling anger does not depend on the possibility of good outcomes. Instead, it depends on the existence of injustices.[39] Accordingly, warranted anger can contribute to shielding us from the derogatory message conveyed by injustice-sustaining risk, and can motivate us to take an expressively powerful stand against the relevant injustices, even in situations where we cannot warrantedly hope for success. Imagine, counterfactually, that Douglass had no chance of defeating Covey in their physical struggle. Even in this "hopeless" scenario, anger would still have been warranted, and could still have helped him to reassert his dignity.

Moreover, even in cases where both anger and hope are warranted, anger has distinctive features that make it particularly well-suited to protecting self-respect. In particular, Samuel Reis-Dennis has argued that

> anger is distinctive because it is *scary*: its connection to action and (sometimes violent) threat allows those who employ it to stand up for themselves, to establish or re-establish social standing and self-respect.[40]

The thought, in other words, is that anger is distinctive partly due the kinds of actions it makes us willing to engage in. Specifically, anger often (though not

[38] Blöser, Huber, and Moellendorf, "Hope in Political Philosophy," 5–6.

[39] Srinivasan, "The Aptness of Anger."

[40] Reis-Dennis, "Anger," 451–52.

always) involves a willingness to engage in confrontational, sometimes even violently confrontational, behaviour—what Reis-Dennis refers to as a "willingness to fight."[41] This is clearly visible in our running examples: Chemaly's and Douglass's intense anger motivates, not just any action, but physically confrontational action. When directed at injustices, this willingness to fight has expressive significance: it communicates, with distinctive force, one's sense that one is owed respect. One possible reason for this distinctive expressive force—which Reis-Dennis alludes to—relates to norms of civility.[42] Confrontation (and particularly violent confrontation) is, in most contexts, a deep departure from conventional norms of civility. Accordingly, a willingness to engage in (violent) confrontation signals how deeply one is committed to defending one's dignity.

In sum, anger's red mist can help protect self-respect in circumstances where hope may be unwarranted; and even in cases where both are warranted, anger's particular motivational profile allows us to reassert our self-respect with distinctive strength. This is not to say that hope should not also play an important role in preserving our self-respect in the face of injustice. But the foregoing considerations suggest that anger's contribution to self-respect cannot fully be replaced by hope.

Still, even if anger's red mist plays a distinctive role in protecting self-respect, one might worry that this role comes at a significant, and perhaps excessive, cost. Anger helps protect self-respect by making injustice-sustaining risk less visible to us. Yet, if we disregard risks that are really there, the causal consequences of our resulting actions might be bad, overall. For example, lashing out at one's oppressor, irrespective of the risks involved in doing so, could lead to a violent backlash and increased oppression. This concern about counterproductivity is especially strong in "hopeless" cases, because in these cases achieving a good outcome, which could counterbalance the risk of negative repercussions, is by definition impossible.

But even if the red mist leads to counterproductive results in many cases, I have argued that there is still a moral reason to commend it: namely, that it contributes to preserving self-respect. Moreover, I have argued that, because

[41] Reis-Dennis, "Anger," 457. The point is not that anger always motivates us to fight (physically or otherwise). It is that anger is prototypically more strongly associated with a willingness to fight than other emotions (in particular, hope)—and this stronger association arguably affects its expressive force. For discussion of anger's characteristic association with "fight," see, e.g., Skitka et al., "Confrontational and Preventative Policy Responses to Terrorism," 375–84; and Berkowitz, "A Different View of Anger," 322–33.

[42] For Reis-Dennis, it is because of their "association with threat and danger that expressions of anger and resentment have their expressive … power. The suspension of civility demands attention" ("Anger," 457–58).

self-respect is a fundamental component of living a good and meaningful life, this reason is weighty.

This is not to say that the value of self-respect always has overriding force. Sometimes, the downstream consequences of the actions motivated by the red mist may be so bad that they override the value of self-respect. In these cases, the red mist is not a morally good thing, overall. But what matters for my purposes is that this is not necessarily the case. There is no reason to think that the disvalue of an action's bad causal consequences will always outweigh the value of self-respect. Indeed, it seems intuitively plausible that the value of self-respect at least sometimes outweighs the counterproductive consequences that may result from blindly lashing out.

Consider again Chemaly's enraged lashing out at the street harasser. Blindly lashing out could have led to more harassment, not less. But it is not clear that Chemaly would regret her act even then. Her proud retelling of the event, even in light of her subsequent awareness of the risks involved, suggests otherwise. It suggests, in other words, that taking a stand—and thereby reaffirming her status as an agent who deserves respect—may well have counted more to her. A similar observation applies to Douglass's attack on Covey. His autobiographical recollection suggests that, even in hindsight, regaining his self-respect (which he likens to a "resurrection") was worth risking his life for.[43]

This suggests that the red mist's benefit to self-respect at least sometimes outweighs its potential counterproductivity. When exactly—and relatedly, how commonly—will it do so? This question cannot fully be answered in the abstract. Determining when the red mist will be overall valuable depends partly on empirical facts about specific real-world contexts. Nevertheless, at least three moral parameters should guide our assessment of the red mist's overall value in particular settings.

The most obvious parameter concerns the *scale* of the red mist's potential negative consequences. The greater the red mist's negative causal consequences are, the less it is likely to be overall valuable. Second, it also matters *who* sustains these consequences. Acting without consideration of risks is morally worse when doing so leads to negative consequences that affect, not just oneself, but innocent bystanders too. This seems a positive feature of Douglass's and Chemaly's actions: prima facie, they do not expose innocent bystanders to harm. The third factor, finally, concerns the *status quo*. When we assess the overall value of the red mist in particular settings, we should consider how they compare to the consequences of inaction. The bad consequences of acting out of blind rage seem less problematic if inaction would have been nearly as bad, than if inaction

[43] Douglass, *My Bondage and My Freedom*, 106.

would have resulted in significantly better consequences. This moral dimension helps appreciate why, of the cases discussed, Douglass's red mist may intuitively seem to be the most valuable. Douglass compares the condition of slavery to a form of death.[44] Precisely because his existing situation was already so terrible, the potential negative consequences of his anger-fuelled resistance had less weight to him than they otherwise would have.[45]

These three moral parameters (which are not intended to be exhaustive) offer preliminary insight into how we may go about assessing the red mist's overall value in particular settings. Now, one complication here is that, once we are angry, it may be difficult to apply these parameters. After all, applying them to a particular case requires knowing about the potential consequences of a course of action, about whom these consequences affect, and about how they compare to the status quo. But this is precisely the kind of knowledge that anger's red mist makes less accessible to the angry person.

Nevertheless, the foregoing account of the conditions under which anger's red mist is overall valuable can still guide action at an earlier stage, *prior* to our becoming angry. Consider two ways it can do so. First, it can guide how we train our emotional dispositions. Emotions are typically not under our *direct* volitional control: we generally cannot simply choose, when confronted with an injustice, whether to become angry or not. But it is nevertheless possible to exercise *indirect* control over our emotions. In other words, our emotional dispositions can be trained through repeated behavioural and cognitive exercises.[46] McBride applies this insight to anger: though he recognizes that doing so is not an easy task, he suggests that we can and should train our anger to make it "attentive to various contexts."[47] Accordingly, the moral considerations outlined above can guide *how* we train or discipline our disposition to feel anger. We can train, for instance, to resist anger—and thus avoid its red mist—in conditions where innocent bystanders are involved.

Second, the parameters outlined above can also guide political rhetoric. Political speakers routinely aim to arouse emotion in their audiences. The decision to verbally arouse anger should be sensitive to whether its red mist would be valuable, overall, in the relevant settings. For example, a public speaker should refrain from verbally exciting anger in her audience, if she suspects that blindness to risk would cause excessive harms, or injure innocent third parties.

[44] Douglass, *My Bondage and My Freedom*, 106.

[45] Douglass, *My Bondage and My Freedom*, 106.

[46] Kristjánsson, *Virtuous Emotions*, ch. 9.

[47] On the importance and possibility of disciplining anger, see McBride, "Anger and Approbation," 7.

Thus, here too, the account I have offered of the red mist's overall value is capable of guiding action.

5. Conclusion

Anger comes at an epistemic cost. It clouds our vision with a red mist. To many, this constitutes one of the central reasons why we should avoid this emotion.

I have argued that this concern is overly hasty. It overlooks, notably, the fact that anger's epistemic cost performs an important moral function. By concealing risk—more specifically, risk that helps sustain injustices—anger helps us retain our self-respect. It does so in two main ways: first, by shielding us from the degrading message associated with injustice-sustaining risk; and second, by helping us to take a stand against injustices we face.

The moral value of this function is nonetheless qualified in at least two respects. First, not all instances of the red mist perform this self-respect protecting function. To reiterate, my argument applies principally to cases where risk sustains injustices. Second, even when it does protect self-respect, the red mist is not always valuable overall. As we have seen, acting without awareness of risk can sometimes engender bad consequences. In some cases, these could outweigh the value of self-respect.

Both qualifications are important. But neither constitutes a decisive problem for my argument. Even if the red mist does not always protect self-respect, the conditions in which it does so remain common in non-ideal conditions. As for the risk of countervailing bad consequences, we can work to elicit anger, and to train our emotional dispositions, so that the red mist arises predominantly in contexts where it does tend to be overall valuable.

Overall, then, my defence does not yield a blanket approval of blind rage. Morally speaking, the red mist is not a tool for everyone and at all times. But the fact that the red mist can be misused should not detract from the following basic insight: that, in the hands of those who face paralysing and degrading risk, the red mist can be, and often has been, a vital protector of dignity.[48]

University of Reading
m.c.lepoutre@reading.ac.uk

[48] For helpful discussions and/or feedback on previous drafts, I am deeply indebted to Alice Baderin, Samuel Bagg, Jacob Barrett, Emma Borg, Joanna Demaree-Cotton, Jamie Draper, Cecile Fabre, Nat Hansen, James Hutton, Alexander Prescott-Couch, and two anonymous reviewers, as well as participants in the Oxford Centre for the Study of Social Justice and the University of Reading's Philosophy Visiting Speaker Series.

REFERENCES

Anderson, Elizabeth, and Richard Pildes. "Expressive Theories of Law." *University of Pennsylvania Law Review* 148, no. 5 (2000): 1503–75.

Baderin, Alice, and Lucy Barnes. "Risk and Self-Respect." *British Journal of Political Science* 50, no. 4 (October 2020): 1419–37.

Bell, Macalester. "Anger, Virtue, and Oppression." In *Feminist Ethics and Social and Political Philosophy*, edited by Lisa Tessman, 165–83. London: Springer, 2009.

———. "Against Simple Removal: A Defence of Defacement as a Response to Racist Monuments." *Journal of Applied Philosophy* 39, no. 5 (November 2022): 778–92.

Berkowitz, Leonard. "A Different View of Anger." *Aggressive Behavior* 38, no. 4 (July/August 2012): 322–33.

Blöser, Claudia, Jakob Huber, and Darrel Moellendorf. "Hope in Political Philosophy." *Philosophy Compass* 15, no. 5 (May 2020): 1–9.

Callard, Agnes. "The Reason to Be Angry Forever." In Cherry and Flanagan, *The Moral Psychology of Anger*, 123–137.

Cogley, Zac. "A Study of Virtuous and Vicious Anger." In *Virtues and Their Vices*, edited by Kevin Timpe and Craig Boyd, 199–224. Oxford: Oxford University Press, 2014.

Chemaly, Soraya. *Rage Becomes Her*. London: Simon and Schuster, 2018.

Cherry, Myisha. "Love, Anger, and Racial Injustice." In *The Routledge Handbook of Love in Philosophy*, edited by Adrienne Martin, 157–68. London: Routledge, 2019.

Cherry, Myisha, and Owen Flanagan, eds. *The Moral Psychology of Anger*. London: Rowman and Littlefiend, 2018.

Deonna, Julien, and Fabrice Teroni. *The Emotions: A Philosophical Introduction*. New York: Routledge, 2012.

Dillon, Robin. "Respect." *Stanford Encyclopedia of Philosophy* (Fall 2018). https://plato.stanford.edu/entries/respect/.

Douglass, Frederick. *My Bondage and My Freedom*. New York: Miller, Orton, and Mulligan, 1855.

Elgin, Catherine. "Emotion and Understanding." In *Epistemology and Emotions*, edited by Georg Brun, Ulvi Doguoglu, and Dominique Kuenzle, 33–50. Aldershot, UK: Ashgate, 2008.

Flanigan, Edmund. "From Self-Defense to Violent Protest." *Critical Review of International Social and Political Philosophy* (forthcoming). Published online ahead of print, January 5, 2021. https://doi.org/10.1080/13698230.2020.1870859.

Frye, Marilyn. *The Politics of Reality: Essays in Feminist Theory*. Trumansburg, NY: The Crossing Press, 1983.

Gambetti, Elisa, and Fiorella Giusberti. "Dispositional Anger and Risk Decision-Making." *Mind and Society* 8, no. 1 (June 2009): 7–20.

Hemenover, Scott, and Shen Zhang. "Anger, Personality, and Optimistic Stress Appraisals." *Cognition and Emotion* 18, no. 3 (2004): 363–82.

Huber, Jakob. "Defying Democratic Despair: A Kantian Account of Hope in Politics." *European Journal of Political Theory* 20, no. 4 (October 2021): 719–38.

Kristjánsson, Kristján. *Virtuous Emotions*. Oxford: Oxford University Press, 2018.

Leboeuf, Celine. "Anger as a Political Emotion: A Phenomenological Perspective." In Cherry and Flanagan, *The Moral Psychology of Anger*, 15–30.

Lepoutre, Maxime. *Democratic Speech in Divided Times.* Oxford: Oxford University Press, 2021.

Lerner, Jennifer, and Dacher Keltner. "Beyond Valence: Toward a Model of Emotion-Specific Influences on Judgement and Choice." *Cognition and Emotion* 14, no. 4 (2000): 473–93.

———. "Fear, Anger, and Risk." *Journal of Personal and Social Psychology* 81, no. 1 (July 2001): 146–59.

Lorde, Audre. "The Uses of Anger." *Women's Studies Quarterly* 25, nos. 1/2 (Spring–Summer 1997): 278–85.

McBride, Lee A., III. "Anger and Approbation." In Cherry and Flanigan, *The Moral Psychology of Anger*, 1–14.

McGowan, Mary Kate. *Just Words*. Oxford: Oxford University Press, 2019.

Nussbaum, Martha. *Anger and Forgiveness*. Oxford: Oxford University Press, 2016.

———. "Transitional Anger." *Journal of the American Philosophical Association* 1, no. 1 (Spring 2015): 41–56.

Pettigrove, Glen. "Meekness and 'Moral' Anger." *Ethics* 122, no. 2 (January 2012): 341–70.

Phoenix, Davin. *The Anger Gap*. Cambridge: Cambridge University Press, 2020.

Rawls, John. *A Theory of Justice*. Rev. ed. Cambridge, MA: Belknap Press, 1999.

Reis-Dennis, Samuel. "Anger: Scary Good." *Australasian Journal of Philosophy* 97, no. 3 (2018): 451–52.

Seneca. *Moral and Political Essays*. Edited by John Cooper and J. F. Procope. Cambridge: Cambridge University Press, 1995.

Shelby, Tommie. *Dark Ghettos: Injustice, Dissent, and Reform*. Cambridge, MA: Harvard University Press, 2016.

Skitka, Linda J., Christopher W. Bauman, Nicholas P. Aramovich, and G. Scott

Morgan, "Confrontational and Preventative Policy Responses to Terrorism: Anger Wants a Fight and Fear Wants 'Them' to Go Away." *Basic and Applied Social Psychology* 28, no. 4 (2006): 375–84.

Srinivasan, Amia. "The Aptness of Anger." *Journal of Political Philosophy* 26, no. 2 (June 2018): 123–44.

Tomasi, John. *Free Market Fairness*. Princeton, NJ: Princeton University Press, 2012.

Wolf, Susan. *Meaning in Life and Why It Matters.* Princeton, NJ: Princeton University Press, 2010.

Journal of Ethics and Social Philosophy
vol. 24, no. 1 · february 2023

https://doi.org/10.26556/jesp.v24i1.1819

THE INTERESTING AND THE PLEASANT

Lorraine L. Besser

Think of the most recent remarkable experience you have had. Perhaps it was reading an engrossing novel that opened your eyes to a new depth of poverty, stamina, and kindness. Perhaps it was attending a sporting event you thought would exemplify stereotypes on the basest level yet turned out to deliver an unexpected but welcome insight into empowerment and dedication. Perhaps it was a walk in the woods, just after the fall leaves dropped, transforming the previously lush forest into a network of brown sticks. Perhaps it was simply a conversation you had with a stranger in line at the coffee shop, which quickly moved from the expected small talk to a brief but memorable exchange about the healing powers of sound baths.

A shared aspect of these experiences is that they are interesting; they are ones that engage, captivate, and enthrall a subject. Exactly how experiences become or are interesting is variable. The quality of any experience depends upon the interaction between the subject and the activity with which she is engaging. Whenever a subject engages in an activity, the mental states she brings to it shape her experience of it. While some qualitative experiences are fairly predictable—scary experiences arise when the activity generates fear within the subject, and pleasurable experiences arise when the activity generates positive affect within the subject—in the case of the interesting, whether or not an experience is interesting is much more difficult to predict, because it depends heavily on the *particular* interaction between the subject and the activity. Sometimes experiences are interesting because they are novel; here what the subject brings is simply a lack of experience, which generates its interesting feeling. Sometimes experiences are interesting because they are unexpected; here the subject brings expectations that turn out to be false and it is the clash of expectations that generates its interesting feeling. And sometimes experiences are simply just interesting; here the subject may bring a sense of curiosity that allows her to become stimulated and enthralled.[1] In

[1] The interesting is thus importantly different and not to be confused with "interested," which I take to describe experiences that align with an agent's particular, preexisting interests. Sometimes the interesting derives from preexisting interests, but, as the examples

each of these instances, the quality of the experience—its being interesting—arises from the particular interaction between the subject and the activity, which I will describe in terms of a "synthesis." But most importantly, in each of these instances, the subject finding the experience interesting adds value to her experience.

We live lives full of interesting experiences, and many seek out interesting experiences as a way to enrich their lives. But the concept of the "interesting," and its status as a prudential value, has received very little attention from philosophers.[2] It is time to remedy this neglect and explore what it means for something to be interesting and what kind of value interesting experiences embody.

In what follows here, after talking a little more about "experiential value" in general, I will begin my defense of the value of the interesting by showing the parallels between the interesting and the pleasant. I will argue that the interesting is an intrinsic prudential value, in largely the same sense that feeling pleasure is valuable: both present a value that is experientially realized and has its roots in the interaction between the agent and the activity. I will go on to argue that, despite sharing the same kind of value, the interesting is distinct from pleasure. Insofar as it challenges the hedonist's assumption that pleasure and pain are the only evaluative dimensions of our phenomenological experiences, my argument here serves both as a defense of the value of the interesting and as an important critique of hedonism.

1. PRELIMINARIES: EXPERIENTIAL VALUE

In identifying the interesting to be an experiential value, my suggestion is that one of the ways in which experiences can be valuable is in virtue of being interesting. While I suspect there are many experiential values, the most familiar is the pleasant. Pleasant experiences are widely taken to be valuable for the subject; that a subject finds pleasure within an activity makes that experience a valuable one. There may be other valuable aspects of the activity and of the

above show, often it is the case that an agent finds experiences interesting even when they do not align with her interests.

[2] Grimm discusses the interesting from an epistemological perspective ("What Is Interesting?"). Kraut hints at the interesting in his discussion of wonder and puzzlement (*The Quality of Life*, 49–50). Matthen describes aesthetic engagement in terms of a distinctive form of psychological engagement that is reinforced by pleasure such that it becomes self-motivated ("The Pleasure of Art" and "New Prospects for Aesthetic Hedonism"). His account comes the closest to mine in its emphasis on cognitive engagement, yet his analysis describes the value of this engagement in terms of the pleasure that attends arousal, whereas the suggestion of this paper is that there is value to the engagement itself.

subject's engagement in it. The activity might be morally valuable, insofar as it helps someone else. The activity might also be such that the subject's engagement in it allows her to develop her capacities, therein having perfectionist value. That the subject finds the experience to be pleasant is a separate and additional source of value. Its value lies solely at the phenomenological level, whereas moral and perfectionist values have their basis within the nature of the activity or within the nature of the subject. While some analyses of these values maintain that there is also a phenomenological component to them, rarely is the phenomenological level sufficient to explain them. Aristotle, for example, locates moral value in the development or exercise of the virtues; he maintains that the experience of exercising virtue generates pleasure within a virtuous agent, but that the virtuous agent finds pleasure in her activity does not *add* to its moral value. Rather, it more properly reflects the moral value of the subject—for she would not find pleasure in the exercise of virtue were she not herself virtuous.[3]

The concept of "experiential value" thus describes a fairly limited kind of value. It describes a value that is good for the subject, in the moment she experiences it. Pleasure has long been regarded as an experiential value. It is something a subject feels as she engages in an activity and gives value to that experience. While hedonism maintains that the experiential value of pleasure exhausts the category of prudential value—such that for the hedonist, experiential value just is prudential value—it is also possible to see experiential value, in general, as *one* form of prudential value. That is, experiential value presents one thing among possibly many things that are valuable to a subject. Just as it can be prudentially valuable for a person to develop virtue, or to exercise their capacities, it can be prudentially valuable for someone to experience value.

Exactly how experiential value weighs up against other sources of prudential value is an important question but will not be my focus here. My aspirations are modest; they are limited to showing the experiential value of the interesting, while leaving open the question of how this experiential value contributes to one's overall good life.[4] Answering this question requires taking stands on the overall nature of prudential value that are not necessary to the current aim, which is to show that the interesting is an experiential value, something that is good for the subject, in the moments in which she experiences it.

[3] Aristotle, *Nicomachean Ethics,* bk. x.

[4] In this respect my perspective on the role of experiential value differs from Kraut, *The Quality of Life*. Kraut's analysis of experiential value parallels mine, yet Kraut seeks specifically to defend experientialism as a form of well-being.

2. WHAT IS THE INTERESTING?

While there are many different senses of the word "interesting" used within ordinary language, I focus here on the sense of interesting as it is used to describe experiences: What do we mean when we say that an experience is interesting, such as when we walk away from a conversation with someone and think, "that was so interesting," or when we find ourselves so utterly enthralled by a film that we just cannot look away? We find an activity interesting when something within it captures our attention in a way that stimulates curiosity and leads us to both notice and seek out the nuances of the activity. When something is interesting this aspect of it tends to linger beyond the immediate engagement. When we walk away from an interesting conversation, we often find ourselves returning to that conversation in our thoughts, and in revisiting that conversation we may even find that thinking about the conversation is itself an interesting experience. Finding something interesting shapes and transforms our experiences and often impacts our perspective. Finding something interesting triggers something within us.

To better illustrate the nature of the interesting, and why some experiences are interesting and others are not, compare the following two experiences.

> Ingrid reads a Jack Kerouac biography and finds herself completely enthralled. She focuses on all of the details, thinking through how Kerouac's real life events play out and compares to the ones in his novels, finding the descriptions of his complicated relationships with others to be a gold mine full of examples of how our interactions with others can shape our values, and of how even the smallest encounter can set one on a new course in life. At the end of reading the biography, she just wants more and makes a plan to read all ten current biographies.

> Anna reads a Jack Kerouac biography and has to force herself to finish it. She cannot understand how (or why) one would choose to live the life that he did. She cannot understand his fascination with Neal Cassady, nor, for that matter, why Cassady deserves such a notorious position within the counterculture movements of the '50s and '60s. The one piece of Kerouac's life that she found worthy of reading about was about the daughter he left behind. While she found this piece of information worth knowing, it also operates as a deal breaker for her, as once she learned this aspect of Kerouac's life, it prevented her from finding anything about the rest of his life interesting. She made it through the biography but sets it aside and never again voluntarily thinks about Kerouac or his novels. When something else reminds her of him, all

> she remembers is that he left his kid. She does not really care about any other aspect of his life.

Here we have an activity that one person finds interesting while another does not.[5] The activity, *qua* reading the biography, is the same. But the experience could not be more different. Ingrid's experience was interesting; Anna's experience was annoying. Whereas Ingrid might describe her experience as life changing, Anna would describe it as a waste of time.

It cannot be that what makes the experience interesting is that the biography was itself interesting. Rather, what differentiates these two experiences is that they are distinct experiences: i.e., different subjects engaging in an activity. The activity, *qua* reading the biography, may be the same, but the subject's experience depends upon her interaction with the activity, which is informed by her specific mental states.

In this case, we might speculate that Anna's commitment to family roles, and to overall good citizenship, might have prevented her from finding the activity interesting. In contrast, we might speculate that Ingrid's rather conservative and sheltered upbringing prompted her to find learning about Kerouac's lifestyle an interesting activity. Our past experiences, and our beliefs and values, certainly factor into how we engage in the activity. Our personality does too: a naturally curious person probably finds a lot more activities interesting than does a person with less curiosity, an observant person might find certain activities such as walking or driving more interesting than a less observant person, and so on.

That which a person brings to an activity plays a central role in whether or not her experience turns out to be interesting; it is likely that the individual's contribution to the activity plays a more important role than features of the activity. The Kerouac example supports this, as do our countless experiences of finding something interesting that others do not. The best explanation of this common phenomenon appeals to the fact that experiences arise from the interaction between a subject and an activity, an interaction that is specified by the subject's mental states and comes together in a synthesis that makes each person's experience unique.

Whenever we engage in an activity, there is some kind of synthesis, and this synthesis determines the phenomenological feel of the experience itself. Attitudinal hedonists point out that when we engage in activities that we desire, we find ourselves pleased: the synthesis created in this case generates a pleasing phenomenological feeling. In this case, if attitudinal hedonism is correct, the

[5] Yet notice that it would not necessarily be apt to say that Anna finds it *boring* simply because she does not find it interesting. I discuss the relationship between the interesting and the boring in the appendix.

synthesis is clearly specifiable in terms of the interaction between a person's attitude and its direction of fit with the activity. But the synthesis that generates interestingness is not one that can be so specified. Sometimes a subject's values and desires drive the synthesis—as is the case with Anna, whose values inhibited her from finding learning about Kerouac's life to be an interesting experience. But sometimes values and desires factor into the synthesis in more dynamic ways, as is the case with someone who finds the experience of reading books about Charles Manson to be interesting. A teenage girl with limited experience of those outside of her rural community may find interest in the sheer differences between her lifestyle and the free yet dangerous lifestyle of those her own age living as part of Manson's values; here the distance between the values she inherited through her upbringing and the values embodied by those so similar in age yet so different in attitude and lifestyle stimulates her curiosity and generates interest. In this instance, values and desires may be relevant, but only because they clash with the content of the activity. And finally, sometimes values and desires do not influence the synthesis at all: sometimes we find reading a book interesting just because the author's writing style "clicks" with us.

We cannot provide a uniform analysis of the interactions that generate interesting experiences, for ultimately whether or not something is interesting depends on unique features of the subject and how she engages in the activity. There are many mental states a subject brings to an activity, including expectations, desires, values, beliefs, general likes and dislikes, and curiosity and other features of one's personality. And there are many ways in which these features can combine with the activity: we expect not to learn anything from our third or fourth reading of the same book, but find that we do; we have strong family values that prevent us from finding anything but negativity when we read about one who abandons his family, or we have values so different from another that they stimulate curiosity and interest, and so on. It is unreasonable to think that we can specify uniformly the features a subject brings to an activity that result in an interesting experience, and, indeed, I think it a central feature of the dynamic quality of the interesting that there can be no such specification.

That the details of the synthesis resist uniform specification does raise the question of whether the various experiences we find "interesting" track something that is uniform across all of its instantiations. When I find reading a philosophy book to be an interesting experience, is this the same phenomenological quality that others find when reading books that are of a very different kind and scope? Is it the same phenomenal quality that others might find on their Sunday drives, or while birdwatching? The concern is straightforward:

Given the vast array of experiences that we find interesting, is it reasonable to assert that there is a shared phenomenological aspect of these experiences?[6]

I find reading philosophy books to be an interesting activity, at least most of the time. I also often find looking at trees to be interesting, and always find thinking about what goes on in my dog's head to be an interesting activity. Even if we keep the subject constant (and so do not take into account that other people very well may not find these activities interesting), is it really plausible to think that all of these experiences share one phenomenological quality of being interesting? I think they do, and that we can helpfully describe this phenomenological quality in terms of a state of cognitive arousal. When we find something interesting, it is because we find that the experience has activated a state of cognitive arousal—it sets in motion the activation of cognitive mechanisms that were previously at rest. Experiences can be vastly different yet have the same phenomenological feel in virtue of the ways in which they stimulate our cognitive capacities.[7]

We can, of course, identify common factors that tend to facilitate or inhibit the degree of interestingness found within an experience. While ultimately an empirical question, it is plausible that facilitators include curiosity, open mindedness, a secure sense of self that allows a person to be open to risk taking and challenges, and a strong sense of autonomy that allows a person to fully emerge in her activities.[8] Inhibitors likely include dogmatism, judgmentalism, fear and insecurity, and depression. These tendencies tend to prevent a subject from fully embracing an activity and so from allowing oneself to be captivated by it. This would be unfortunate, because interesting experiences, as I will argue in the next section, are a source of intrinsic prudential value.

[6] This is the "problem of heterogeneity" often discussed within literature on hedonism. Within hedonism, the question is whether pleasure tracks a homogenous "felt quality" that explains the wide variety of ways in which we experience pleasure. Is the pleasure I take in reading a book the same as the pleasure I take in getting a massage or someone else takes in eating tripe? Some take the heterogeneity problem to be decisive against hedonism, and reflection on the heterogeneity problem has led to the development of attitudinal hedonism. The stakes of this issue are less pressing for the current discussion than they are for hedonism, however, for what makes the heterogeneity problem especially problematic for hedonism is hedonism's claim that there is *one* intrinsic value (pleasure). This makes it particularly pressing to show how the variety of forms that pleasure seems to take amount to the same thing. Where a plurality of experiential values is on the table, the concern is less pressing.

[7] Notice that my appeal to cognitive arousal helps to explain the phenomenology of the interesting but is not intended to explain the value of the interesting.

[8] Besser and Oishi, "The Psychologically Rich Life"; Oishi et al., "The Psychologically Rich Life Questionnaire."

3. THE VALUE OF THE INTERESTING: JUST LIKE PLEASURE

Having isolated the "interesting" as a qualitative feature of our experiences, let us now consider the kind of value found within the interesting. It is clear, I hope, that interesting experiences are prudentially valuable. They benefit us and they enrich our lives. Interesting experiences are ones that tend to stand out, penetrate our memories, and linger. But just what kind of value do they have? In this section, I will argue that the kind of value found within interesting experiences is parallel to the value of pleasant experiences. Both the interesting and the pleasant are intrinsic prudential values.

First, the above analysis shows the value of the interesting to be fundamental and not derivative of something else. It does not depend on any particular attitude; it does not depend upon a particular skill or the exercise of particular capacities; it does not depend upon success; it does not depend upon one's own values or any kind of objective value the experience might possess; nor does it depend on any value attached to our cognitive capacities. It does arise from the interaction between the agent and the activity, but the synthesis that generates an interesting experience is unspecifiable. The interesting is, indeed, a unique form of value. That it is unique gives it a fundamental value, insofar as its value is inexplicable by appeal to other sources.

Second, the value of the interesting is intrinsic, insofar as its value is inherent to and inseparable from the experience. Where an experience is interesting, it presents an intuitive, undeniable value to the subject in the moments she experiences it. Where an experience is not interesting, it lacks this value, although it may be valuable in other respects. Keeping in mind that being interesting is a feature of the experience but not the activity itself, let us return to our opening pair of examples. Ingrid's experience reading Kerouac had an undeniable value for her. This is *evidenced* by both her decision to read the entire catalog of biographies and by the positive attitude that she takes toward her experience, but the value itself is explicable solely in terms of the phenomenological feel of her experience. In contrast, Anna's experience reading Kerouac was not interesting and so lacks this value. The activity might be valuable, but Anna's experience of reading it was not itself valuable.

We can now see that the kind of value associated with the interesting very much parallels the value hedonists attribute to the pleasant. Hedonists maintain that pleasure is an *intrinsic* value—its occurrence is itself always valuable. Moreover, hedonism most often interprets pleasure to be a *prudential* or *relational* intrinsic value: its occurrence is always valuable to the subject. This is the kind of value we find within interesting experiences. Interesting experiences are intrinsically valuable for the subject. If I am right about this, then this challenges hedonism's position that there is only *one* intrinsic value. Yet before developing this line of criticism against the hedonist, let us first consider the similarities between the interesting and the pleasant. Doing so affirms the status of the interesting as an intrinsic prudential value.

As we move into this analysis, the form of pleasure I will focus on is the phenomenological account of pleasure, according to which pleasure is defined by its distinctive feel, or "felt quality." While some hedonists locate intrinsic value in attitudinal pleasure, it is the phenomenological account of pleasure that parallels the interesting in several respects, and as such will be our focus here.[9]

When we think about pleasure and try to wrap our heads around why it is valuable, we likely find ourselves stuck with the simple fact that *pleasure just is valuable*. The value associated with pleasure is undeniable and therein intuitive. We can argue about degrees of value and whether or not the value of pleasure outweighs other concerns, but the claim that an experience is pleasant, yet *entirely lacking in value*, lacks plausibility. The pleasure counts for something. As Katz observes:

> Pleasure presents as good and attractive—itself, when it comes to our notice, and all else that appears aglow in its light. This suggests simple explanations both of why people pursue pleasure and why there are reasons to do so. That we may prefer and choose something for its pleasure suggests that there are facts about pleasure that make some such choices better than others. Philosophers, taking this suggestion further, have sometimes taken pleasure to be a *single simple (feature of) experience that makes experiences good and attractive to the extent it is present*.[10]

The presence of pleasure itself is valuable. While we do not often recognize the interesting to be of undeniable value, I hope that the analysis of the interesting that I have offered reveals the plausibility of this claim. Once we have isolated the interesting as a qualitative aspect of our experience, we see the value of it, a value backed up by our first-personal experiences of interesting experiences. We feel the pull of the interesting. We may feel it in different degrees and frequencies, but once we have felt it, we recognize its value to be undeniable.

The interesting and the pleasant, moreover, are both experiential values. They present the same kind of value in virtue of being qualities of our experience; that is, they are features of the experience itself, rather than products of

[9] I will return to discussion of attitudinal pleasures in the next section, where I argue against the move to reduce the interesting to a form of attitudinal pleasure.

[10] Katz, "Pleasure."

that experience.[11] In a similar vein, Bramble explains, "that the pleasantness and unpleasantness of experiences is right there in the experiences themselves" is commonsensical:

> if you are walking along a suburban street, and find yourself suddenly struck by a pleasant smell, say, of jasmine (or some other flower—take your pick) wafting from a passing garden, the experience you become aware of seems already to be pleasant, i.e., pleasant even before you have had a chance to take up any kind of attitude toward it.[12]

This is an important point: even though pleasing experiences typically generate positive attitudes, pleasure is the quality of our experience that generates those positive attitudes. We judge that our experiences are pleasant when they are pleasant; our judging them to be pleasant is not what makes them pleasant. The pleasure of eating a perfectly textured and rich chocolate mousse lies in the experience of eating it.

While the interesting does not just *strike* us as does the pleasant smell of jasmine, and often requires active engagement, it is nonetheless a *quality* of the experience in the same sense in which pleasure is a quality of experience—they are both qualitative aspects of our experience that are independent of the subsequent judgments or attitudes a subject may develop toward that experience.

Nor does the interesting derive from any preconceived judgments or attitudes we form prior to the experience, such as *being interested* in an upcoming activity. Experiencing something to be interesting is distinct from being interested in something. Often this attitude precedes a subject's engagement. We read a book because we are interested in it; we take a particular class because we are interested in it. While it is tempting to think that interesting experiences derive from a subject's sense of interest, and so derive from an antecedent attitude she has toward her activity, it takes only a quick reflection to realize that this is not true. There is an important difference between being interested in *X* and *X*'s being interesting. I might be interested in reading Kant's *Critique* but very well might not find the experience interesting. Whether or not I find the experience interesting depends upon more than just my attitude. One's attitudes contribute to the experience insofar as they help shape one's engagement in the activity and the synthesis arising from that engagement, but, as we have

[11] Ordinary language often identifies objects or activities themselves to be pleasant or interesting. It does so in a dispositional sense. A pleasant temperature is one that people tend to experience as pleasant; an interesting book is one that people will tend to experience as interesting to read. Both values are located within an experiencing subject and require an experiencing subject to be realized.

[12] Bramble, "The Distinctive Feeling Theory of Pleasure," 203.

seen, there is no formula to this synthesis. Being interested in something does not *make* the experience interesting. Sometimes I am interested in activities that do turn out to be interesting, and it is plausible to think that my attitude helped make the experience interesting, but it could very well go the other way. My prior interest in something might lead me to develop such high expectations that the experience cannot compare, and so would serve to detract from the interestingness of the experience.

Given that the interesting (and the pleasant) is a quality of our experience, claims to its value are (as with the value of the pleasant) compatible with the experience requirement, which holds that in order for something to be of *prudential* value for a given individual, it must factor into her experience.[13] The experience requirement is often invoked in the context of hedonism; indeed, Sumner describes it as "the important insight in classical hedonism."[14] The experience requirement appeals to those who worry about making claims that something is good for someone even if it does not impact her experientially. For instance, an oft-cited objection to desire theories of welfare concerns cases where desires are satisfied without the subject knowing it. Is it really plausible to maintain that her welfare has been improved when the determining factor (desire satisfaction) occurs without her being aware of it? Most agree that it is not.

While the experience requirement is intuitively plausible, we should notice that appreciating the value of the interesting does not *commit* one to the experience requirement, which is typically formulated as the claim that something *must* enter into your experience in order to contribute to your welfare. We can recognize that the interesting is an experiential value in the same sense in which the pleasant is an experiential value without having to also maintain that *all* prudential values must be like this.[15]

Finally, the interesting and the pleasant are similar in virtue of being prudential values. They benefit the person experiencing them and make her life go better for her. Goldstein makes this point especially well with respect to pleasure. He argues that we can understand the prudential value of pleasure through reflection on its reason-giving character. To say that pleasure is an intrinsic value is to make reference to its status as a self-justifying end. It affords "valid, intrinsic grounds for desire."[16] While I do not follow Goldstein in his claim that pleasure is the *only* intrinsic good, his construal of the value of

[13] Griffin, *Well-Being*.

[14] Sumner, *Welfare, Happiness, and Ethics*, 128.

[15] Compare to Bramble, who maintains that the plausibility of the experience requirement anchors hedonism ("A New Defense of Hedonism about Well-Being").

[16] Goldstein, "Pleasure and Pain," 275.

pleasure as intrinsic to its psychological occurrence, which on its own grounds desire, provides a good illustration of why we think pleasure is a prudential value. Pleasure is good for me—why? John Stuart Mill was not that far off in claiming that we know it is valuable because we desire it.[17] Whatever limitations this style of argument may have, that we desire something provides a good indication of *prudential* value.

The value of the interesting is like this. Its value is experienced from the inside. Kraut frames this idea in terms of "internal observation," arguing that

> some things are known by internal observation: this is how we know what pleasure and pain are like, what it is like to desire something, entertain doubts about something, find something intriguing, feel sadness, remorse, guilt, and so on. So, when we judge that an experience we are having or have had is immeasurably rich and worth having for itself—as when we are absorbed in a great work of art or surrounded by great natural beauty—we have some basis for valuing this experience precisely because it is our experience and we know it from the inside.[18]

When considered from the inside, we see that having an interesting experience taps into our desires: it makes us want more; it leads us to have positive attitudes toward that experience and to see more generally that the fact that something is interesting is a reason to engage in it. That the interesting has this reason-giving effect reveals it to be of prudential value. Yet, just as some people find the value of the pleasant to be *more* reason giving than others, some people find the value of the interesting to be *more* reason giving than others.

This variety of responsiveness to the values of the pleasing and the interesting should not make us question the value inherent to them. It is enough to establish intrinsic value to find that there *is* responsiveness to it, not that there are equal degrees of responsiveness to it across subjects. Railton argues that it seems "to capture an important feature of the concept of intrinsic value to say that what is intrinsically valuable for a person must have a connection with what he would in some degree find compelling or attractive, at least if he were rational and aware, but that it would be an intolerably alienated conception of someone's good to imagine that it might fail in any such way to engage him."[19]

[17] I refer to Mill's infamous proof of utilitarianism in which he argues that pleasure is valuable because we desire it (*Utilitarianism*, ch. 2). While I agree with those who point out that desiring something does not make it valuable, it also seems plausible that desiring something *prima facie* indicates that it is of value and that we can reflect on what we desire to help us identify what is valuable in itself.

[18] Kraut, *The Quality of Life*, 51.

[19] Railton, "Facts and Values," 9.

This seems right. In identifying something as an intrinsic value, we commit to saying that it will engage—or resonate with—most of us. We do not commit to saying that the degree to which people respond to its value will be consistent across subjects. Likewise, in identifying something as a prudential value, we commit to saying it is good for a subject. We do not commit to saying that it requires that people structure their lives around it. Thus, in claiming that the interesting is an intrinsic prudential value, my claim is simply that, where present, the interesting adds value to a subject's experiences and so benefits her. But there are lots of values like this, pleasure included, and it is within the individual's prerogative to choose which prudential values she prioritizes and structures her life by.[20]

The view I am putting forward here maintains that, as an intrinsic prudential value, the interesting (and the pleasant), when present, is always a valuable aspect of our experience. It is thus *pro tanto* reason giving. Its actual reason-giving force for any particular person, however, depends upon that person's responsiveness to its value. We all respond to the interesting—this is at root what it means for it to be an intrinsic value—but the degree to which we respond to its value informs the degree to which we see it as reason giving. Just as some take the fact that something is pleasant to be decisive, while others take it to simply count in favor of that experience, some will take the fact that something is interesting to be decisive, while others may not. Particularly because both are prudential values, their actual reason-giving force will vary between subjects, according to the degree to which they respond to its value.

This analysis of responsiveness is similar to what some have described in terms of "psychological resonance." Dorsey argues that the fact that something psychologically resonates with an agent indicates its relational intrinsic value.[21] What can it mean for something to be intrinsically good for a subject? Dorsey argues that the answer must be, at least in part, that it always resonates with her. This resonance is partly if not wholly what makes it good for her. We can extend this thought by recognizing the *degree*s in which something resonates within a particular subject. For example, most people will accept that the experience of pleasure resonates as an intrinsic value. But people differ widely in their reactions to the experience of pleasure. For some, this value resonates strongly and decisively; for others, this value resonates—but not very strongly, such that its value may not be decisively reason giving for them.

[20] The line of argument echoes Tiberius's theory of well-being as value fulfillment, which holds that a person's well-being depends on the degree to which she lives life according to her own values (*Well-Being as Value Fulfillment*).

[21] Dorsey, "Intrinsic Value and the Supervenience Principle."

We can look around at the people we know and recognize the different degrees in which something's being pleasant resonates with them; I think the same holds for the interesting. Something's being interesting to that person indicates some degree of psychological resonance (and it might very well be that something's being interesting is itself a particular manifestation of psychological resonance), but the degree to which it resonates varies, both between subjects and even within the same subject. It is a familiar occurrence to one day feel like pursuing interesting experiences, while the next day to feel like pursuing pleasant ones. Sometimes something just resonates more for us at a particular moment. Claiming that an experience has intrinsic prudential value involves making the claim that such experiences resonate with subjects, but it does not involve making a claim about the extent to which such experiences resonate.

4. THE VALUE OF THE INTERESTING: TOO MUCH LIKE PLEASURE?

I have argued that the pleasant and the interesting share the same kind of value: they are both prudential, intrinsic values that are experientially realized. This conclusion leads to the question of whether or not the interesting is *too much like pleasure,* and therein subject to some of the same objections often raised against the pleasant.

The first objection runs roughly as follows. The view of pleasure invoked by my argument, which takes pleasure to be characterized by its phenomenological feel, may be the ordinary sense of pleasure, but philosophical literature on pleasure invokes more sophisticated understandings of pleasure that focus not on its felt quality, but rather on its connection to attitudes. This is the position of attitudinal hedonism, according to which pleasure derives from the positive attitude a subject takes toward an activity, rather than the felt quality of her experience. Defenders of this move make it largely to avoid some of the counterintuitive implications that arise when hedonism combines with a phenomenological view of pleasure. While many of these implications are not relevant to the current analysis, one stands out as especially relevant. This is the concern that, absent a more sophisticated analysis of our attraction to pleasure (e.g., one that connects pleasure to attitude), experiencing pleasure is a contingent experience with a questionable value.

Noting its roots within Findley's critique of pleasure, Bramble describes this concern as "Findley's objection."[22] The basic worry is that the story I have presented thus far, which appeals solely to the phenomenological experience of pleasure to determine its value, cannot go far enough to establish the prudential

[22] Bramble, "A New Defense of Hedonism about Well-Being."

value of pleasure. If, as I have argued, the value of the interesting likewise has its roots solely within its phenomenology, it is open to the same concern. If we cannot explain why the pleasant (or the interesting) resonates with us, why should we think the fact that it does reveals intrinsic value? After all, we have all kinds of attractions and aversions to aspects of our experiences. That I am averse to the combination of seafood and cheese does not imply anything about its value. Why should we think our attraction to the pleasant (and the interesting) is different?

Bramble responds to this line of argument by emphasizing the nature of the claims we are making when we say something is pleasant. He argues that

> a pleasant experience, even if its subject has no notion that it is going on, still possesses the phenomenal feel characteristic of pleasures. This is why it is good. Why is it the involvement of "the pleasant feeling," rather than, say, the sound of Ella Fitzgerald's voice, the smell of jasmine, or yellow phenomenology, that is what makes an experience good? There is no answer to this question, but also no need for one.[23]

His point, I take it, is that it is the simple experience of *pleasure* that leads us to recognize an experience as valuable. There is nothing more to it; the features that contribute to our experience of pleasure (e.g., the smell of jasmine) are not the good-making feature of the experience—the pleasure is. Thus there is not a further account of *why* the pleasure is valuable and we do not need such an account to establish the experience as valuable.

This line of response works just as well, if not better, with respect to the interesting. We know the interesting has intrinsic value because of our experience of it. And especially because the current argument is that the interesting has intrinsic value that is neither exclusive nor decisive, there really is no need for further explanation.[24]

This focus on the *experiential* quality of the pleasant (and the interesting) takes us to a second objection frequently raised against hedonism, which calls into question whether *all* instantiations of pleasure have value. People find pleasure in all kinds of experiences, including experiences that harm others or violate

[23] Bramble, "A New Defense of Hedonism about Well-Being," 214.

[24] This is not to say that there *is* no further explanation possible of why we are so responsive to pleasure, or to interesting experiences. There very well may be one and I think it likely that in the case of the interesting, this explanation revolves around the nature of cognitive arousal. The point is that the further explanation is tangential to the questions of why it is valuable, for their value lies within the experience. Whatever explanations of these experiences we can offer will provide insight into their value, which lies solely with their phenomenology.

well-entrenched standards of morality: torturing bunnies, having sex with dead people, inflicting pain on others. It seems mistaken to claim that the pleasure in these contexts is valuable, even if it shares the same felt quality we recognize to be a value in other contexts. Surely, we might think, these kinds of examples suggest we have erred in locating intrinsic value within the experience of pleasure.

In a similar fashion, it seems possible that people find the interesting in all kinds of experiences, many of which we might hesitate to attach value to. Are we prepared to accept as valuable the interest one finds in reading about the gory details of Charles Manson's violent crimes? Or what about the interest one finds in staring at picture after picture of dead bodies? It is possible that people find these experiences to be interesting, yet it also seems counterintuitive to say that these experiences have intrinsic value for the subjects engaging in them.

In response to this line of criticism, I, like the hedonist, maintain that, despite the counterintuitive nature of these experiences, if there is a subject who finds them interesting (or pleasant), there is indeed prudential, intrinsic value in them—they are valuable for the subject. Yet we can assuage the counterintuitive impact of biting this bullet by acknowledging the existence of a plurality of intrinsic prudential values, as well as moral values that may constrain our pursuit of prudential value.[25] Recognizing a plurality of values allows us to accept that an experience's being interesting has intrinsic prudential value, even though the experience overall might have negative value for the subject. It might have negative prudential value, perhaps by stimulating within the subject desires that stand in tension with her aims, or it might have negative moral value, insofar as it displays a lack of respect for humanity.

My argument does entail that where an experience is interesting, it has intrinsic value for the subject, even if that experience is otherwise morally reprehensible. Acknowledging the existence of other forms of value (both prudential and moral) provides grounds to rationally criticize a person's engagement in these kinds of activities, while allowing that the fact that they are interesting

[25] It is not clear that hedonism can make this move, at least not as persuasively. Hedonism maintains that all and only pleasure has intrinsic value. Some forms of hedonism limit their claim to the singularity of pleasure to cover prudential value, while others cover moral value, but distinctive to hedonism is a claim about the singularity of pleasure as the only relevant value within a context, be it prudential, moral, or both. My understanding of experiential value carries with it no such singularity, and indeed embraces the notion that there is a plurality of experiential values, for it is certainly possible that there are more experiential values than the pleasant and the interesting. My analysis of experiential value moreover takes it to be one species of prudential value and claims neutrality with respect to moral value; therein it is compatible with the stipulation that there are moral values, some of which may override prudential values, including experiential prudential values, or that there are more pressing prudential values that outweigh experiential prudential values.

does deliver that experience some form of value, therein mitigating the counterintuitive nature of these examples.

A related concern arises with respect to locating value in meaningless experiences, such as the pleasure a subject might take in compulsive masturbation, or the interest one may develop in clicking on meme after meme on social media. In response to this concern, we first ought to seriously challenge whether they genuinely present experiences of the pleasant or the interesting, or whether these kinds of examples are just pseudo-experiences of these values that we label as pleasant or interesting. This strategy helps to put into perspective the examples: the person clicking away on Instagram probably is not having an interesting experience, just as the compulsive masturbator probably is not finding a lot of pleasure in her activities. A more accurate description of these examples is that the subject seeks out experiential value in activities that are not the best situated to deliver the experience they are looking for. But this strategy will not work to show that meaningless experiences *cannot* ever be experientially valuable. They can. And I think it important to recognize this: there can be experiential value in otherwise meaningless activities.

It may seem meaningless to find value, as I have claimed to do, in looking at trees, or, to address Rawls's example, in counting blades of grass.[26] But the fact that the interesting can be found in many sorts of otherwise meaningless activities is a *good* thing.[27] That I am able to find the interesting in looking at trees means I am able to transform otherwise mundane activities into valuable experiences. When I am sitting in the dentist's chair getting my teeth drilled and am able to look out at the trees outside the window and find doing so to be interesting, I have been able to insert some value into an otherwise painful experience. We can think of plenty of examples where being able to find an intrinsic value in an experience *transforms* that experience for the better. If a blade counter truly can find her activity to be a source of interest, that is representative of a critical skill that she can enrich her life by harnessing.

5. THE INTERESTING VS. THE PLEASANT

I have argued that the interesting is a prudential value in much the same sense in which the pleasant is a prudential value. Both present as intuitively valuable aspects of our experience that make our lives go better. Given the parallel form of value within each, at this point it is natural to question whether the

[26] Rawls, *A Theory of Justice*, 379.

[27] And to the extent that we can find pleasure in meaningless activities, that is a good thing too.

interesting is distinct from pleasure. Perhaps the interesting is another form of pleasure in disguise, in which case we have not departed from hedonism at all.

We find seeds for this line of argument within Bramble's defense of hedonism. Bramble argues that all experiential value is hedonic; that is, that as soon as we have embraced the experience requirement and take prudential value to be an aspect of our experiences, hedonism intuitively follows:

> There is a powerful reason, however, to accept the hedonist's view. This is what has come to be known as the experience requirement. The experience requirement says that for something to be good or bad for someone it must affect his experiences in some way—specifically, it must affect their phenomenology or ":what it is like for him to be having them. If the experience requirement is true, then hedonism is almost certainly true as well—indeed, it would be the reason why the experience requirement is true. There is little plausibility, after all, to the idea that any non-hedonic phenomenology (i.e., phenomenology that is neither pleasurable nor painful) is intrinsically relevant to well-being.[28]

Is it really so implausible to think that experiences could be not pleasant yet have intrinsic value? I do not see it. It seems that our experiences are multidimensional and have many different valences and degrees of intensity and being pleasurable is not the only way in which an experience can be of intrinsic value.[29]

Interesting experiences are sometimes positively valenced, but not always. They can be uncomfortable, as when we are reading the details of Charles Manson's life and empathizing with the young women who dedicated their lives to him, or when we stumble across a snake sunning itself in our path. Here the discomfort is part of what makes the experience so interesting; it does not detract from its value. When interesting experiences are positively valenced, such as an encounter with someone whose lifestyle could not be more different than your own, that generates fascination and intrigue but is also a pleasant conversational experience that makes you feel warm; the interesting aspect often carries the tone. We walk away from such an encounter thinking about how interesting it was, not just about how pleasing it was.

[28] Bramble, "The Role of Pleasure in Well-Being,"207.

[29] Kraut echoes this line of thought, writing, "our phenomenological world is a highly variegated matter, and pleasure is just one small aspect of it. If we consider in isolation from the riches of the other components of our experience, it remains something to which we are attracted, but we are (and ought to be) far more attracted to the complex phenomenological array of which it is a part. [Pleasure] is only a small part of a full account of what is good for us" (*The Quality of Life*, 39).

Defenders of hedonism, such as Bramble, must explain these kinds of experiences in terms of pleasure in order to maintain their commitment to hedonism's claim that all and only pleasure has value. Bramble would argue that the experiences I describe as interesting are really just pleasant ones, for he maintains that pleasures are more diverse than we often realize and a lack of positive affect does not automatically entail that the experience is not pleasant.[30] It might rather be the case, he argues, that we have limited introspective aspect to the positive valence of the experience, as there is increasing empirical evidence that calls into question our introspective capacities.

The felt-quality hedonist will thus strive to maintain that experiences of the interesting share the same felt quality as the pleasant, even if we may not be able to identify them as such. In contrast, an attitudinal hedonist will strive to explain the value we find in the interest in terms of the subject's positive attitudes toward that experience. Because the attitudinal hedonist defines pleasure in terms of attitudes, this move amounts to claiming that the value of the interesting is identical to that of the pleasant.

Heathwood, for example, defines pleasure exclusively in terms of a subject's attitude: if a subject wants something to occur (intrinsically, for its own sake) then it counts as pleasure for her.[31] Working from this framework, we might say that interesting experiences are valuable because we want them to occur, i.e., because they are pleasant. Even if we did not antecedently desire them, as long as we contemporaneously desire them the positive attitude we take toward them counts as pleasure and gives the experience its value.

It seems we are all working in the same ballpark. Both Bramble's and Heathwood's versions of hedonism recognize that there is more to prudential value than identification of the warm feelings associated with positive affect and both go very different ways—even from each other—to describe this aspect. But they both remain committed to hedonism and so move to develop more inclusive understandings of pleasure in order to accommodate these aspects of prudential experiential value. Here I cannot do justice to their sophisticated arguments, but I will simply ask: Why not just recognize the intrinsic value of the interesting? Why keep striving to reduce experiential value to the pleasant?

There are at least two pressing reasons why we should recognize interesting experiences as intrinsically valuable in virtue of being interesting and not just as another form of pleasure in disguise. First, given the rise in interdisciplinary research on prudential value, it is important to use of a sense of pleasure that crosses borders. In other disciplines where pleasure is studied, including

[30] Bramble, "The Distinctive Feeling Theory of Pleasure."

[31] Heathwood, "The Reduction of Sensory Pleasure to Desire."

psychology and economics, pleasure is interpreted in the sense I employ here as something that feels good and is defined by its positive valence. While I am sympathetic to the reasons why literature on hedonism has moved toward a more inclusive notion of pleasure, I worry about the limitations this move places on the applicability philosophical notions of pleasure have for empirical research.

The problem is multilayered. It concerns not just the likelihood that concepts developed on solely philosophical grounds will depart from the concepts as they are studied scientifically, but concerns also the additional threat that these concepts themselves are of limited use to scientific research insofar as they are less apt to inform constructs that can be measured and so used within research.[32] This problem is particularly acute in the case of pleasure, wherein the sense of pleasure invoked by ordinary discourse— i.e., taking "pleasure" to refer to a felt quality of positive affect—*can* be measured and *does* inform current research. Introducing (or continuing to use) theoretically developed concepts of pleasure in the face of these practices begets a gulf between philosophical theories and scientific research.

Second, and independently of the first, I think it is crucial to recognize the value of the interesting insofar as recognizing it as such not only validates those who choose to pursue the interesting, but also opens up a new dimension of value for others to pursue.[33] Part of the project of exploring experiential value is a normative one: our philosophical interest lies not solely in describing forms of value but in putting forward values as ones that are worthy of pursuing. Many people have interesting experiences yet may not recognize or appreciate the value this adds to their lives. Identifying the interesting as an intrinsic prudential values both validates the role these experiences play in people's lives and encourages people to seek out interesting experiences. It opens up a new dimension of the good life for them, one that might be better suited to them than the pursuit of pleasure. Redescribing the interesting in terms of pleasure threatens to erase a significant source of value in life.

[32] This line of thought echoes Alexandrova, who worries that, in the context of well-being, in particular, the theoretical definitions offered by philosophers are not capable of informing the constructs used in the science of well-being. The philosopher's theoretical goals, she argues, stand in tension with the goals of the scientist. The scientist needs to work with definitions of well-being that are "sensitive not only to the normative theories of the good life but also to the practical constraints of measurement and use of this knowledge" yet the "goals of theorizing about well-being in philosophy as it is currently practiced are not sensitive in this way" (Alexandrova, *A Philosophy for the Science of Well-Being*, xxxi).

[33] A conclusion I explore in more detail in Besser and Oishi, "The Psychologically Rich Life."

6. CONCLUSION

Recognizing the value of the interesting opens up a new framework from which to think about the phenomenal character of our experiences and the prudential value they offer. I have argued that the value of the interesting shares the same kind of value as sensory pleasure yet is distinct from it. This translates to a rejection of hedonism's claim that pleasure is the only intrinsic value, but it does not call into question the status of pleasure as an intrinsic value. Rather, I have suggested that there can be a plurality of intrinsic, prudential values that people respond to differently. We may not all *seek out* interesting experiences, but we find value in them. And so we probably should seek out interesting experiences more often.

Our daily lives consist in having experiences. While there is more to life than the sum of our experiences, and other forms of value than experiential value, finding value in the experiences we have is important. We do not always have to find the pleasant in our experiences for them to be valuable. Sometimes we just need to find them interesting. The reality is that it may be easier to find the interesting than the pleasant. To find the interesting, we might just need to open up our minds and engage in our activities with a sense of curiosity. Our minds might be able to transform an activity, allowing us to find value within an experience that might otherwise have simply been something we did on a Monday.

Middlebury College
lbesser@middlebury.edu

APPENDIX: BOREDOM

One question that arises with respect to the claim that the interesting has intrinsic value is what its opposite might be. The dichotomy between pleasure and pain is so clear, with one having positive value and one having negative value, that it is natural to assume that a parallel structure arises with other qualitative aspects of our experiences. The opposite of the interesting seems to be the boring. Boring experiences are dull, monotonous, uneventful. Rather than create a compelling spark within the individual, boring experiences generate anything from apathy to annoyance and aversion.

If the boring is the opposite of the interesting, and the interesting has intrinsic value, does the boring have intrinsic disvalue? The *prima facie* challenge is that some people do not mind boring experiences, and some people may even find the boring aspects of an experience appealing, so might even see it as a positive aspect of experience. After an extremely active or hectic time, some people seek out "boring" experiences as a way to balance themselves. If the boring can be choice worthy, that is an indication that it may not always be associated with negative value.

I worry that this line of reasoning embraces a misleading use of the term. When we seek out "boring" experiences, what we are really seeking out are *quiet* times, experiences that do not cognitively engage in the way that interesting experiences do. I do not think that this—a state in which the mind is at rest—is genuinely a state of boredom. States of boredom, rather, arise when our minds strive toward activation, yet are frustrated. We want to find something interesting, to be cognitively aroused, yet those wants go unfulfilled. This state of frustration more accurately tracks boredom—and notice that it *prima facie* presents as a negative value. Its negative value is explained by the individual striving to find something interesting and their failure to succeed. If so, then it is likely the intrinsic value of the interesting does, like pleasure, have an opposite, whose value, like pain, is negative.[34]

References

Alexandrova Anna. *A Philosophy for the Science of Well-Being*. Oxford: Oxford University Press, 2017.

Aristotle. *Nicomachean Ethics*. Edited by Roger Crisp. Cambridge: Cambridge University Press, 2014.

Besser, Lorraine L., and Shigehiro Oishi. "The Psychologically Rich Life." *Philosophical Psychology* 33, no. 8 (2020): 1053–71.

Bramble, Ben. "The Distinctive Feeling Theory of Pleasure." *Philosophical Studies* 162, no. 2 (January 2013): 201–17.

———. "A New Defense of Hedonism about Well-Being." *Ergo* 3, no. 4 (2016): 85–112.

———. "The Role of Pleasure in Well-Being." In *The Routledge Handbook of the Philosophy of Well-Being*, edited by Guy Fletcher, 199–208. London: Routledge Press, 2016.

Dorsey, Dale. "Intrinsic Value and the Supervenience Principle." *Philosophical Studies* 157, no. 2 (January 2012): 267–85.

[34] These ideas have emerged through many fruitful conversations, and I am grateful for the opportunities I have had to discuss this work with audiences at the Bled Ethics Conference, the Pacific Division of the American Philosophical Association, Washington University in St. Louis, and University of Vermont. My thanks go out in particular to Julia Annas, Anne Baril, Tyler Doggett, Dan Haybron, Kate Nolfi, Shige Oishi, Jason Raibley, and Vida Yao for their insightful feedback.

Goldstein, Irwin. "Pleasure and Pain: Unconditional, Intrinsic Values." *Philosophy and Phenomenological Research* 50, no. 2 (December 1989): 255–76.

Griffin, James. *Well-Being: Its Meaning, Measurement, and Moral Importance*. Oxford: Clarendon Press, 1986.

Grimm, Stephen. "What Is Interesting?" *Logos and Episteme* 2, no. 4 (2011): 515–42.

Heathwood, Chris. "The Reduction of Sensory Pleasure to Desire." *Philosophical Studies* 133, no. 1 (March 2007): 23–44.

Katz, Leonard D. "Pleasure." *Stanford Encyclopedia of Philosophy* (Winter 2016). http://plato.stanford.edu/archives/win2016/entries/pleasure.

Kraut, Richard. *The Quality of Life: Aristotle Revised*. New York: Oxford University Press, 2018.

Matthen, Mohan. "New Prospects for Aesthetic Hedonism." In *Social Aesthetics and Moral Judgment: Pleasure, Reflection and Accountability*, edited by Jennifer A. McMahon, 13–33. London: Routledge, 2018.

———. "The Pleasure of Art." *Australasian Philosophical Review* 1, no. 1 (2017): 6–28.

Mill, John Stuart. *Utilitarianism*. 2nd ed. Edited by George Sher. Indianapolis: Hackett Press, 2002.

Oishi, Shigehiro, Hyewon Choi, Nicholas Buttrick, Samantha J. Heintzelman, Kostadin Kushlev, Erin C. Westgate JaneTucker, Charles R. Ebersole, Jordan Axt, Elizabeth Gilbert, Brandon W. Ng, and Lorraine L. Besser. "The Psychologically Rich Questionnaire." *Journal of Research in Personality* 81 (August 2019): 257–70.

Railton, Peter. "Facts and Values." *Philosophical Topics* 14, no. 2 (Fall 1986): 5–31.

Rawls, John. *A Theory of Justice*. Rev. ed. Cambridge, MA: Harvard Univeristy Press, 1999.

Sumner, L. W. *Welfare, Happiness, and Ethics*. Oxford: Clarendon Press, 1996.

Tiberius, Valerie. *Well-Being as Value Fulfillment: How We Can Help Each Other to Live Well*. New York: Oxford University Press, 2018.

Journal of Ethics and Social Philosophy
vol. 24, no. 1 · february 2023

https://doi.org/10.26556/jesp.v24i1.2419

AGAINST BROOME'S "AGAINST DENIALISM"

Kabir S. Bakshi

Imagine that on a Sunday afternoon, I take a ride in my suv just to enjoy myself. I could have easily not taken the ride and done nothing. (I will refer to this case as *Joyride*.) Call those who defend *individual denialism* (id)—the claim that current humans (in some sense) do no wrong by not refraining from performing acts that emit insignificant or small amounts of greenhouse gases (call such acts *ghg acts*)—individual deniers.[1] An individual denier holds, for example, that I did no wrong in Joyride. Positions in the neighborhood of id have recently been defended by Walter Sinnott-Armstrong, Elizabeth Cripps, Aaron Maltais, and Ewan Kingston.[2] id has recently been argued against by John Broome.[3]

In this paper, I critically evaluate Broome's argument against id. He argues that the claim that individual emissions "do no harm is not true in general."[4] I proceed as follows. In section 1, I clarify, isolate, and present Broome's argument. Sections 2–4 contain three problems for Broome's argument. I argue that Broome's argument overgeneralizes (section 2), is in tension with his defense of carbon offsetting (section 3), and uses problematic assumptions (section 4). Section 5 closes.

Before I start, a preliminary. In the literature, id is expounded in various ways. Some use the language of obligations. For example, Aaron Maltais notes that "obligations to reduce one's greenhouse gas emissions appear to be difficult to justify."[5] Others use the language of (partial or group) causation. For example, Casey Rentmeester writes that "an individual drive does not itself cause climate change, but it is certainly a … factor."[6] Still others use the language

[1] "Small" is important in the formulation because even individual deniers will say that I do wrong if I commit acts that produce enormous amounts of emissions.

[2] Sinnott-Armstrong, "It's Not My Fault"; Cripps, *Climate Change and the Moral Agent*; Maltais, "Radically Non-Ideal Climate Politics and the Obligation to at Least Vote Green"; Kingston and Sinnott-Armstrong, "What's Wrong with Joyguzzling?"

[3] Broome, *Climate Matters* and "Against Denialism."

[4] Broome, "Against Denialism," 110.

[5] Maltais, "Radically Non-Ideal Climate Politics and the Obligation to at Least Vote Green," 589.

[6] Rentmeester, "Do No Harm," 16.

of contribution. For example, Melissa Lane notes that "every single emission contributes to the composite problem, which is made of trillions of tiny emissions."[7] Since my aim in this paper is to critically evaluate Broome's argument against ID, I grant him his formulation. According to Broome, ID is the thesis that "individual human beings do [no] harm by their emissions."[8] I think that Broome's argument against ID can be objected to on the ground that Broome's construal of ID is inaccurate, but this is a critique I will not develop further.

1. Broome's Argument

1.1. Two Mistakes

Broome, in his "Against Denialism," argues against ID by noting that the individual denier goes wrong in two respects. First, she fails to consider that what matters in issues of uncertainty is not the actual goodness (or value) of the outcomes of an action, but the expected goodness of the outcomes:

> In the face of uncertainty . . . what you ought to do depends, not on the goodness of actual results, which you cannot know, but instead on the goodness of the "prospect" that each of your alternative acts leads to. A prospect is a portfolio of all the various outcomes that might result from an act, each associated with its probability of happening.[9]

Broome cashes out the goodness of an act's prospect in terms of expected utility theory (EUT).[10] In EUT, the *ex ante* value of (the prospect of) an act is its expected value, where the expected value of an act is the sum of the product of the probability of each possible outcome of the act and the value associated with that outcome.[11] That is:

$$Exp(A) = \sum_j P(A \Box \rightarrow O_j)V(O_j).$$

[7] Lane, *Eco-Republic*, 59.

[8] Broome, "Against Denialism," 110.

[9] Broome, "Against Denialism," 114.

[10] There are many versions of EUT, including ones that give plausible results in infinitary scenarios (Easwaran, "Strong and Weak Expectations"; Thalos and Richardson, "Capitalization in the St. Petersburg Game") and those that permit different risk attitudes (Buchak, *Risk and Rationality*). For surveys, see Buchak, "Decision Theory"; Briggs, "Normative Theories of Rational Choice"; and Thoma, "Decision Theory."

[11] What follows is the causal decision theory version. One can also provide an evidential decision theory version (see Ahmed, *Evidential Decision Theory*, for a thorough treatment). The differences are inessential to my argument.

A consequentialist (like Broome) takes the further step and holds that an agent ought not to do an act iff there is an alternative that has a higher *ex ante* value.

Second, according to Broome, the individual denier errs in being ignorant of the chaotic nature of weather systems. Broome contends that "most writing on the ethics of climate change ignores the instability of the atmosphere."[12] "Chaotic" is a technical term and is usually used to describe nonlinear deterministic systems that are highly sensitive to initial conditions. Although a chaotic system is deterministic—that is, it follows a unique evolution fixed by the dynamics and the initial conditions—its final state is taken to be unpredictable. Importantly, a chaotic system (or any deterministic system for that matter) cannot evolve to become a nondeterministic system. The unpredictability of a chaotic system is not ontic: given initial conditions and dynamics, a chaotic system—*qua* a deterministic system—will after time *t* end up in a unique final state. The unpredictability is epistemic: given initial conditions and dynamics, it is not always possible to tell in which state a chaotic system will end up after *t*.[13]

I turn to presenting Broome's argument next.

1.2. Broome's Challenge

Broome founds his challenge to ID on the two mistakes mentioned above. Although in "Against Denialism," Broome is concerned with arguing against specific defenses of ID (particularly Kingston and Sinnott-Armstrong and Cripps), a general argument against ID can be isolated.[14] According to Broome, given the unstable nature of the atmosphere, the outcomes of a GHG act may result in a different weather condition than the condition that would have occurred had the GHG act not been committed. He writes, "Given the atmosphere's instability, we should expect global weather in a few decades' time to be entirely different if you go joyguzzling on Sunday from what it would have been had you stayed at home."[15] These different conditions, given the chaotic nature of the atmosphere, may result in states that are completely different in their goodness than other states:

[12] Broome, "Against Denialism," 113.

[13] This is an informal and rough gloss. For a more detailed and careful introduction of chaos, see Strogatz, *Nonlinear Dynamics and Chaos.* Batterman, "Defining Chaos"; Bishop, "Metaphysical and Epistemological Issues in Complex Systems"; and Bishop, "Chaos," provide good introductions to issues of chaos in philosophy of science.

[14] Kingston and Sinnott-Armstrong, "What's Wrong with Joyguzzling?"; Cripps, *Climate Change and the Moral Agent*.

[15] Broome, "Against Denialism," 113.

> Increasing emissions … will cause typhoons to form at quite different times and places, and it will lead to a completely different distribution of cholera outbreaks. Your Sunday drive will cause a completely different group of people to be exposed to cholera and other risks of death. Some who would have died will survive because of your drive, and others who would have survived will die.[16]

We do not know whether the state resulting from the outcomes of a GHG act will be good or bad: "When you consider whether or not to joyguzzle … you cannot know what good or harm will actually result from what you do. The result may be a typhoon or a child's death, or it may be good."[17] What we do know is that our GHG act will have some effect: "There is literally zero probability that [it] will do no harm and no good.… Also, there is about equal probability that it will do good as that it will do harm."[18] And the expected value of a GHG act will be lower than the expected value of not committing a GHG act because GHG acts cause harms in expectation. Broome bases the expected harm of a GHG act on the social cost of carbon: "Your joyguzzling on Sunday afternoon creates a prospect that has a positive expectation of harm to other people. Its expectation of harm is *given by the [social cost of carbon], which measures the average*, or expected, harm done by emissions of carbon dioxide" (emphasis mine). Since GHG acts result in expected harms, ID is incorrect: "Your act may or may not do harm, but it certainly creates an expectation of harm. Individual denialists [do] not claim merely that your emissions may not do harm, which is true. They claim they actually do no harm, which is not true in general."[19]

In premise-conclusion form and a bit more filled in, Broome's argument is:

P1. If the atmosphere is a chaotic system, then small changes in the state of the atmosphere at one time may lead to drastically different states of the atmosphere at a future time.

P2. *Unpredictability*: If the atmosphere is a chaotic system, we cannot know what the state of the atmosphere will be at a future time given a small change in the state of the atmosphere at a previous time.

P3. *Appropriate*: In decisions under uncertainty, we should appeal to expected utility theory.

[16] Broome, "Against Denialism," 113.

[17] Broome, "Against Denialism," 114.

[18] Broome, "Against Denialism," 113.

[19] Broome, "Against Denialism," 115. As I noted in the introduction, it is not obvious that individual deniers will accept Broome's characterization of ID.

P4. *Diff*: The drastically different states of the atmosphere resultant from small changes in the atmosphere correspond to states of affairs that may differ drastically in their goodness.[20]

P5. The atmospheric system is chaotic.

P6. GHG acts lead to small changes in the state of the atmosphere.

C1. GHG acts lead to drastically different states of the atmosphere that correspond to states of affairs that may differ drastically in their goodness. (from P1, P5, P6, and Diff)

C2. We cannot know what the state of the atmosphere will be at a future time given a GHG act at a previous time. (from Unpredictability and P5)

C3. In the decision of whether to commit a GHG act, we should appeal to expected utility theory. (from C2 and Appropriate)

P7. *Risk*: Under expected utility theory and given C1, GHG acts lead to a net expectation of harm.[21]

P8. ID denies that GHG acts lead to any harm.

C4. ID is incorrect. (from C3, Risk, P8)

I think that problems can be raised against many steps in this argument. For example, one may deny Appropriate by claiming that EUT (or at least standard EUT), which does not allow for differences in risk attitudes or discount for negligible probabilities, produces paradoxical results. One may maintain that in cases of uncertainty—especially when the probabilities associated are minuscule—we should eschew standard EUT. Monton considers infinite St. Petersburg paradox like cases and argues that it is irrational to take into account

[20] See the passage from Broome ("Increasing emissions ... will die"), quoted earlier in this section, where he seems committed to Diff. Indeed, it seems that something like Diff is indispensable to anyone who is sympathetic to Broome's argument. In other places in "Against Denialism" and in *Climate Matters* (see esp. ch. 7), Broome seems to assent to principles in the neighborhood of Diff.

[21] See the quote from Broome ("Your joyguzzling ... carbon dioxide") earlier in this section where Broome seems committed to Risk. As with Diff, Broome assents to Risk in *Climate Matters*. He writes:

> Expected value theory tells us that, in assessing the badness of climate change, we have to think in terms of expectations. The expectation of harm caused by a catastrophe is the badness of the catastrophe multiplied by the very small probability that it will happen. ... The most likely result of climate change is warming of a few degrees. But the view is that the possible catastrophe of a greater increase would be so bad that, even multiplied by its very small probability, its expected badness is more important than the harm that would be caused by this most likely result. (Broome, *Climate Matters*, 131)

small probabilities in our calculations.[22] Russell argues against Appropriate in the context of longtermism.[23] Decision-theoretic worries aside, one may deny Broome's argument on physical grounds. For example, recent works in the philosophy of physics problematize Unpredictability. In a series of papers, Werndl argues that the issue of unpredictability in climate science is much more nuanced than previously appreciated.[24] I think these points are—or at least can be converted into—powerful objections against Broome's argument. However, in this paper I will focus on Diff and Risk.

In the rest of the paper, I discuss reservations with Risk, arguing in section 2 that it leads to a problem of overgeneralization and that it is unstable, and arguing in section 3 that it is incompatible with Broome's appraisal of carbon offsetting. I discuss a reservation with Diff in section 4.

2. Overgeneralization and Instability

Here is a compelling principle:

> *Restrict*: Any account that denies ID must not be so restrictive that, on the account, an agent is required not to ϕ, when ϕ–ing is uncontroversially taken to be morally permissive.

An example of an uncontroversial act that is morally permissive is breathing continuously. I hope that this is uncontroversial: it would be incredibly strange to say that I ought not to breathe continuously if I can skip some breaths. But I submit that on Broome's argument, I do wrong when I do not skip a breath whenever I can; I do wrong when I exercise (other things being equal) because a human engaged in exercising produces (on average) up to eight times the CO_2 emissions of a sedentary human; I do wrong when I play the clarinet or perform any activity that increases my greenhouse gas emissions (other things being equal).[25]

[22] Monton, "How to Avoid Maximizing Expected Utility."

[23] Russell, "On Two Arguments for Fanaticism." See Wilkinson, "In Defence of Fanaticism," for a defense of fanaticism in the context of longtermism.

[24] Werndl, "What Are the New Implications of Chaos for Unpredictability," "On Defining Climate and Climate Change," and "Initial-Condition Dependence and Initial-Condition Uncertainty in Climate Science."

[25] Palmer, "Do We Exhale Carbon?" See also Sinnott-Armstrong, "It's Not My Fault," 301–2, where Sinnott-Armstrong makes a similar remark about exercise counting as a moral wrong if one denies ID. However, he does not take into account the chaotic and unpredictable nature of the atmosphere—one of Broome's major points in "Against Denialism."

Broome's argument is insensitive to the difference between the wrongness of a GHG act and the wrongness of a morally unproblematic act. What goes for a GHG act also goes for an act like breathing continuously. Consider:

> *BID-n*: In taking every *n*th breath, an individual does no expected harm.

Steps after P6 in Broome's argument can be suitably modified to give an argument requiring an individual to skip, say, every *n*th breath:

P6*. Every *n*th breath I take leads to small changes in the state of the atmosphere.
C1*. Every *n*th breath I take leads to drastically different states of the atmosphere that correspond to states of affairs that may differ drastically in their goodness.
C2*. We cannot know what the state of the atmosphere will be at a future time given that I breathe at a previous time.
C3*. In the decision of whether I should take every *n*th breath, we should appeal to expected utility theory.
P7*. *Risk**: Under expected utility theory and given C1*, every *n*th breath I take leads to a net expectation of harm.
P8*. BID-*n* denies that breathing acts lead to any expected harm.
C4*. BID-*n* is incorrect.

This, I think, is a highly undesirable consequence of Broome's argument. In focusing only on the expected harm and arguing that an individual ought not to perform an act only because of its expected harm, Broome's account fails to satisfy Restrict.

Climate scientists and policymakers usually explain why breathing does not contribute to climate change by appealing to the fact that respiration is part of a closed-loop cycle—that is, respiration is part of a cycle that, on net, is (approximately) carbon neutral. The closed loop includes CO_2 absorption by plants. Humans—like other animals—are a carbon sequestration machine, albeit a very slight one.[26] But this reply is not available to Broome. The material facts of an act of emission are of no consequence in Broome's view. All that matters, as I noted above, is whether the act of emission has expected harm.

For Broome, an act of emission has expected harm even if the emissions are insignificant. In arguing against the insignificance of a particular GHG act, Broome argues both that an act that produces insignificant harms cannot have zero expected harms because "only zeros add up to zero" and that a proponent of

[26] Schwarcz, "Why Isn't the Carbon Dioxide from Breathing a Concern for Global Warming?"; Withers, "How Much Does Human Breathing Contribute to Climate Change?"

ID "must say that . . . emissions do no harm at all."[27] But if the only thing that matters is that an act has expected harm of zero, then we should refrain from committing any number of activities—including breathing—because even though the expected harm of the CO_2 emitted when I breathe is extremely small, it is not zero.

It is estimated that every day an average human emits about one kilogram of CO_2.[28] Numbers are inessential to Broome's argument. Suppose that I breathe *m* times in a day and that I can—without any effect on me—skip taking a breath once every two days. I will not contribute 1/2*m* kilograms of CO_2 into the atmosphere per day. Since Broome maintains that I cannot know whether this particular emission will "trigger a jump" to an extremely bad state of affairs that includes floods, famines, and so on, I must base my decision of whether I should exhale this particular breath on the expected harm related to this act. As I noted above, there is some (maybe extremely tiny) expected harm associated with my exhaling this particular breath. I should, Broome must say, refrain from breathing once every two days. This is an extreme consequence, and I submit that any account that supports such a consequence must be dismissed.

But this is not all. Broome's account has another distasteful consequence: it is *unstable*. A Broome-style argument can be used to show that the action Broome's account recommends is itself an action that leads to expected harms. That is, a Broome-style argument can be given to show that in refraining from committing a GHG act, an agent does harm in expectation. Consider

> *NID*: In not committing a GHG act, an agent does no expected harm.

Call the omission of a GHG act a *non-GHG act*.[29] To be sure, by a "non-GHG act" and by "not committing a GHG act," I mean the non-doing of a GHG act. For example, if instead of taking the ride in my SUV in Joyride, I do not take the ride, I commit a non-GHG act. A Broome-style argument can then be run to show that NID is incorrect.[30] Steps P6 onward in Broome's argument against ID can be suitably modified as:

[27] Broome, "Against Denialism," 122.

[28] Palmer, "Do We Exhale Carbon?"

[29] This should not be read as any endorsement of whether an omission is an act or whether an omission can cause something else to be the case. For positions arguing against causation by omission, see Beebee, "Causing and Nothingness"; and Moore, *Causation and Responsibility*. For positions arguing for causation by omission, see Lewis, "Causation as Influence"; Lewis, "Void and Object"; Schaffer, "Causation by Disconnection"; and Schaffer, "Causes Need Not Be Physically Connected to Their Effects." And for positions arguing that omissions can cause something to be the case but that omissions have an inferior causal status, see Dowe, *Physical Causation*; and Hall, "Two Concepts of Causation."

[30] Thanks to an anonymous reviewer for advising me to expand this point. The reviewer also raised a worry that if "non-GHG acts" is read this way, P6** may be contentious, especially

P6**. Every non-GHG act leads to small changes in the state of the atmosphere.

C1**. Non-GHG acts lead to drastically different states of the atmosphere that correspond to states of affairs that may differ drastically in their goodness.

C2**. We cannot know what the state of the atmosphere will be at a future time given a non-GHG act at a previous time.

C3**. In the decision of whether to commit a non-GHG act, we should appeal to expected utility theory.

P7**. Under expected utility theory and given C1**, non-GHG acts lead to a net expectation of harm.

P8**. NID denies that non-GHG acts lead to any harm.

C4**. NID is incorrect.

This makes Broome's account critically unstable. If we are to agree with Broome, we (do harm in expectation and hence) do wrong by both committing and not committing a GHG act. I submit that this consequence is sufficient to dismiss Broome's argument.

3. Tension with Offsetting

I now turn to another problem with Risk. Risk is in tension with Broome's defense of offsetting our carbon emissions. In *Climate Matters*, Broome notes that the "most effective way of reducing your emissions to zero is to cancel or offset the emissions" and that "offsetting is the way you can fulfil your duty of justice."[31] He takes offsetting to be any action that removes greenhouse gases

if "lead to" is read in causal terms. The point I want to make in P6** is not that the absence of a GHG act will lead to (or cause, partially or otherwise) changes in the atmosphere. That will commit me to some form of causation by omission. What I want to say in P6** is a bit more modest: whatever action an agent does instead of performing the GHG act will inevitably have some impact on the distribution of greenhouse gases and hence will lead to (or cause, partially or otherwise) changes in the atmosphere. It is in this sense that every non-GHG act leads to changes in the state of the atmosphere. Consider Joyride. As I set it up in the introduction, I commit a GHG act in Joyride because I take a ride in my SUV. Now consider the modified version of Joyride I mentioned in my discussion of non-GHG acts. If I refrain from taking the ride in my SUV, I commit a non-GHG act. But not taking a ride (i.e., a non-GHG act) will lead to (or cause, partial or otherwise) changes in the state of the atmosphere. For example, instead of taking the SUV ride, I might take the bus or walk or decide to stay in my room. Whatever I do (e.g., when I breathe), I change the state of the atmosphere in some way. It is in this way I want to read P6**. And it seems to me that read this way, P6** is not very contentious.

[31] Broome, *Climate Matters*, 80.

from the atmosphere. According to Broome, an individual does “no harm by emissions” if they “successfully offset all” of their emissions, although he considers some concerns about the idea of offsetting. He writes: “I am not recommending offsetting to governments; I am recommending it only to individuals as a way of acting justly.… Private offsetting is a means by which each person can avoid causing harm to others.”[32] He notes in “Against Denialism” that “once carbon dioxide is in the atmosphere, some fraction of it remains there in effect forever,” and he says in *Climate Matters* that “once you have put a tonne of carbon dioxide molecules into the atmosphere, those molecules will wreak their damage.”[33] But in justifying offsetting, he further claims:

> If at the same time you remove the same number of other carbon dioxide molecules, you prevent those ones from wreaking damage. Your overall effect is zero. As far as the climate is concerned, emitting a tonne of carbon dioxide and offsetting it is exactly as good as not emitting it in the first place, providing the offset is genuine.[34]

But this seems at odds with his argument against ID, especially Risk.[35] Broome’s argument is insensitive to any kind of offsetting. Once a GHG act has been committed and some amount of greenhouse gases have been added to the atmosphere, the expected harm of the act cannot be changed. Even if offsetting removes the same (or indeed even if it removes a greater) amount of greenhouse gases, Broome must say that the GHG act ought not to be committed. Broome’s argument against ID makes offsetting or any kind of carbon-canceling principle otiose. The only way for offsetting (of a GHG act *G*) not to be in tension with Broome’s account is for offsetting to guarantee that it instantly takes out from the atmosphere the particular greenhouse gas molecules emitted because of *G*. This is impossible. Broome’s argument may block ID, but it makes his own views untenable.

The force of my objection against Broome can be appreciated in another way. In justifying and defending offsetting, Broome is appealing to a principle along the following lines:

> *Aggregate*: The overall effect due to a GHG act *G* is fully grounded on facts about the change in the total amount of greenhouse gases in the atmosphere due to *G*.

[32] Broome, *Climate Matters*, 94–95.

[33] Broome, “Against Denialism,” 118.

[34] Broome, *Climate Matters*, 118.

[35] See Campbell, “Offsetting, Denialism, and Risk,” for a similar observation but a different line of argument to show the tension between offsetting and Broome’s argument against ID.

When Broome writes that "if . . . you remove the same number of other carbon dioxide molecules, you prevent those ones [i.e., the ones from a GHG act] from wreaking damage," Broome appeals to Aggregate. However, it seems that Aggregate stands in stark tension to his argument that due to the chaotic and unstable nature of the atmosphere (and Diff), a GHG act will produce expected harm, however one compensates for it. Consider Joyride. When you drive your gas-guzzler for fun, your emissions of greenhouse gases will create atmospheric disturbances that may cause significant changes in large-scale meteorological events, some of which may result in harm that would not otherwise have occurred. Offsetting your emissions probably would not undo these effects. Insofar as offsetting would also cause (sufficiently large) atmospheric disturbances, offsetting would probably also inflict harm that would not otherwise have occurred.[36] Here again, instability looms.[37]

Contra Broome, it is difficult to claim, given his own argument against ID, that adding greenhouse gases and then removing the same amount of greenhouse gases (and not the same greenhouse gases) from the atmosphere does not result in any harm.

4. AGAINST DIFF

Recall Diff. In the case of GHG acts and given P6, Diff gives us:

> *Diff-G*: The drastically different states of the atmosphere resultant from GHG acts correspond to states of affairs that may differ drastically in their goodness.

Given the atmosphere's chaotic nature, Diff-*G* seems like a reasonable premise. After all, since a chaotic system is sensitive to initial conditions, the minutest change in the composition of the environment may lead to completely divergent final states. However, I think that this is a bit quick. I present two reasons for my trepidation.[38]

[36] As an anonymous reviewer helpfully points out, offsetting itself requires acts such as planting trees and rainwater harvesting that will need one to commit GHG acts.

[37] Elliott Thornley raises the point that offsetting might lower the expected harm of my life overall and that it might also cause benefits that would not otherwise have occurred. Agreed. But these points can also be made in favor of GHG acts. A GHG act may also cause benefits that would not have occurred, and due to the chaotic nature of the atmosphere, it might also lower the expected harm of my life overall. Broome designs his argument to deny individual deniers this strategy. But in doing so, he also denies helping himself to this strategy to defend offsetting.

[38] Diff—at least as it stands—can also be challenged on the grounds of the existence of carbon reservoirs. One can argue that not all the emissions from a GHG act are absorbed

First, observe that Broome's argument has nothing to do with climate change *per se*. Broome's argument is independent of the issues of climate change or global warming. Broome's argument only turns on decision-making under uncertainty. In our case, it is decision-making about committing a GHG act under the instability of the weather. This feature makes the scope of Broome's argument quite wide. It seems that his argument against ID is insensitive to particular issues of climate change. I leave it to the reader to judge how bad such a consequence is for Broome's account.

My second reason against Diff is more direct. I think Diff-*G* is question begging. According to Broome, the expected goodness of different states due to a GHG act is measured by the social cost of carbon (SCC), which represents "the present discounted value of the additional social costs (or the marginal social damage) that an extra tonne of carbon released now would impose on the current and future society."[39] Putting SCC and Diff-*G* together gives:

> GHG acts at time *t* lead to drastically different states of the atmosphere that correspond to states of affairs that may differ in their goodness, where the goodness of a state of affairs can be measured (or represented) by a function of the amount of CO_2 in the atmosphere at the state and the SCC at *t*.

According to Broome, whether an agent should commit a GHG act then boils down to whether the SCC associated with the act is positive. If it is positive, then the act is associated with expected harm and thus the agent should refrain from performing it:

by the atmosphere; rather, the (extremely) vast majority of these emissions are absorbed by deep oceans. Indeed, environmental economists working on calculating the social cost of carbon make this assumption. For example, the influential integrated assessment model Dynamic Integrated Climate-Economy (DICE) developed by William Nordhaus explicitly models the deep oceans as a carbon reservoir (see Nordhaus, *A Question of Balance*; and Nordhaus and Boyer, *Warming the World*). More recently, Golosov et al. write:

> The stock of carbon in the deep oceans is very large compared to the amount in the atmosphere and also relative to the total amount of fossil fuel yet to be extracted. This means that, of every unit of carbon emitted now, only a very small fraction will eventually end up in the atmosphere. Thus, the linear model [such as DICE] predicts that even heavy use of fossil fuel will not lead to high rates of atmospheric CO_2 concentration in the long run. ("Optimal Taxes on Fossil Fuel in General Equilibrium," 64)

I will, however, not engage further with this strand of opposition, in part because Broome maintains that any tiny amount of emissions absorbed by the atmosphere, given the atmosphere's chaotic nature, creates expected harm.

[39] Hope and Newbery, "Calculating the Social Cost of Carbon," 10.

> The total benefit of cancelling the emission [i.e., not performing a GHG act] is the integral over all future times of the reduction in harm at each time. This integral is what is measured by the social cost of carbon.... Its total benefit is therefore the integral over an infinite time of a positive amount.... [But] economists who estimate the SCC estimate this integral as finite, because they discount future benefits exponentially. Exponential discounting leads to a convergent integral.... My point in this paper is only that the integral is not zero.[40]

Broome's explication of the SCC as an "integral over all future times of the reduction in harm at each time" is, I think, unclear at best. Rather the SCC (at a time *t*) is a function of the difference between the harms associated with large-scale emissions activity committed at *t* and the harms associated with no such large-scale activity. For example, in explaining their Policy Analysis of the Greenhouse Effect 2002 (PAGE) integrated assessment model (IAM), Hope and Newbery write:

> The PAGE model calculates the social cost of carbon (SCC) by finding the difference in the discounted economic cost of climate change impacts between two emission scenarios that are identical except for the emission of an extra one billion tonnes of carbon as CO_2 in 2001 for one of the scenarios. The difference in impacts is divided by one billion to obtain the SCC.[41]

Similarly, Nordhaus's influential DICE IAM defines the SCC at time *t* as:

$$\mathrm{SCC}(t) := \frac{\dfrac{\partial W}{\partial E(t)}}{\dfrac{\partial W}{\partial C(t)}} = \frac{\partial C(t)}{\partial E(t)},$$

where *W* is a welfare function (which depends on, among other things, population, per capita consumption, and time discount factor), *C*(*t*) is a consumption

[40] Broome, "Against Denialism," 118.

[41] Hope and Newbery, "Calculating the Social Cost of Carbon," 14. Influential IAMs include DICE (Nordhaus, *A Question of Balance*), CETA (Peck and Teisberg, "CETA"), PAGE (Hope, "The Marginal Impact of CO_2 from PAGE2002," "Optimal Carbon Emissions and the Social Cost of Carbon over Time under Uncertainty," and "Discount Rates, Equity Weights and the Social Cost of Carbon"), MERGE (Manne and Richels, "Merge"), FUND (Tol, "On the Optimal Control of Carbon Dioxide Emissions"), and MIT ISGM (Webster et al., "Uncertainty Analysis of Climate Change and Policy Response").

function at t, and $E(t)$ is the total carbon emission function.[42] $E(t)$ is in turn defined as:

$$E(t) := \sigma(t)[1 - \mu(t)]Y(t) + EL(t),$$

where $\sigma(t)$ is the "carbon intensity" due to uncontrolled industrial CO_2 emissions, $\mu(t)$ is the emissions reduction rate, and $EL(t)$ is the exogenous land emissions. Again, as in the case of PAGE, we see that the SCC is calculated, not by considering GHG acts, but by using a unit of emissions due to industrial or large-scale disturbances.

A proponent of ID might already get off the bus because to appeal to the SCC to determine the expected harm of a GHG act seems to beg the question. The individual denier is not necessarily a collective denier, so she will be happy to accept that the SCC provides a measure of the harms caused by humans as a collective. But the individual denier, rightly in my opinion, will deny that it follows from the SCC that a GHG act causes any expected harm.

Moreover, even if one is not antecedently committed to ID, Broome's use of the SCC to argue against ID is problematic. First, his argument commits what Zimmerman and Kaiserman label the pie fallacy, "according to which there is some fixed 'quantity' of responsibility available for every outcome, to be distributed among all those, if any, who are responsible for it."[43] It is not obvious—and Broome must provide additional arguments to support the claim—that there is a fixed amount of responsibility associated with the harms of climate change that should be distributed among individuals. Second, in taking the SCC as the basis for proportioning expected harms among individuals committing GHG acts, Broome is implicitly assuming that moral facts about groups and collectives reduce to or supervene on moral facts about individuals. This is a highly controversial position.[44] Indeed, I think that a central issue in the debate about ID is whether such a reduction is even possible. The proponents of ID may deny it.[45] Broome in using the SCC is begging the question.

[42] Nordhaus, "Revisiting the Social Cost of Carbon," 1521.

[43] Zimmerman, "Sharing Responsibility"; and Kaiserman, "Responsibility and the 'Pie Fallacy,'" 3598.

[44] Recent work in deontic logic proves that given plausible assumptions about group membership, moral facts about groups are not logically reducible to moral facts about individuals. See Tamminga and Duijf, "Collective Obligations, Group Plans and Individual Actions"; Tamminga and Hindriks, "The Irreducibility of Collective Obligations"; and Duijf, Tamminga, and Van De Putte, "An Impossibility Result on Methodological Individualism."

[45] Relatedly, a proponent of ID may deny that there is any connection between the phenomenon of climate change due to the aggregate greenhouse gases in the atmosphere and the greenhouse gas molecules that make up the aggregate. For example, Kingston and

Objection: In raising problems against Broome's use of the SCC to evaluate the expected harms of GHG acts, I have concentrated only on the calculation of the SCC by IAMs. Models by their design only provide an approximate measure of harm. Maybe an analytic expression rather than numeric calculation of the SCC will not face the same problems.

Reply: Analytic expressions for the SCC are—like analytic expressions for any complex problem—nonexistent or extremely difficult. In recent years, there has been a growing literature on providing analytic expressions for the SCC, under suitable assumptions.[46] The most prominent expressions are by Golosov et al. and van den Bijgaart, Gerlagha, and Liski.[47] Given certain plausible assumptions, Golosov et al. derive the SCC as:

$$SCC(t) = Y(t)[Exp(\Sigma_{j=0}^{\infty} B^j \gamma_{t+j}(1 - d_j))].$$

The details of the expressions are not important to my point. What is important is the fact that analytic derivations proceed in the same way as the IAMs. They take harm done at a large scale and then calculate the harm done per capita. In the Golosov et al. expression, $(1 - d_j)$ represents "the amount of carbon that is left in the atmosphere" j time steps in the future.[48] Similarly, van den Bihgaart, Gerlagha, and Liski use the global CO_2 stock, defined "over and above the pre-industrial level of CO_2."[49] The problem I raised for IAMs still stands for (at least the current) analytic expressions of the SCC.

Moreover, every model or analytic procedure to calculate the SCC makes use of global temperature patterns. The damage function that "describes the economic impacts or damages of climate change" used in DICE is defined as:

$$\Omega(t) := \frac{D(t)}{1 + D(t)},$$

where

$$D(t) := \psi_1 TAT(t) + \psi_2 [TAT(t)]^2.$$

Sinnott-Armstrong argue that it is plausible to think that climate change is an emergent property ("What's Wrong with Joyguzzling?" 175–76).

[46] Namely that (i) utility is a logarithmic function of consumption, (ii) current climate damages are proportional to output and are a function of the current atmospheric carbon concentration with a constant elasticity (a relationship that is allowed to vary over time/be random), (iii) the stock of carbon in the atmosphere is linear in past and current emissions, and (iv) the saving rate is constant.

[47] Golosov et al., "Optimal Taxes on Fossil Fuel in General Equilibrium"; and van den Bijgaart, Gerlagha, and Liski, "A Simple Formula for the Social Cost of Carbon."

[48] Golosov et al., "Optimal Taxes on Fossil Fuel in General Equilibrium," 51.

[49] Van den Bijgaart, Gerlagha, and Liski, "A Simple Formula for the Social Cost of Carbon," 77.

and where *TAT* is the global average temperature.[50] Nordhaus writes that "the DICE-2016R model takes globally averaged temperature change (TAT) as a sufficient statistic for damages."[51] Parallelly, Golosov et al. follow Nordhaus in taking the damage function to be (in the first step) "the mapping from carbon concentration to climate (usually represented by global mean temperature)."[52] The individual denier is well within her rights to deny Broome's argument just by pointing out that in using the SCC to argue against ID, Broome is begging the question.

5. Conclusion

My aim in this paper has been to critically evaluate Broome's recent argument against individual denialism. By clearly presenting Broome's argument in section 1, I made the target of my criticism clear. In sections 2–4, I raised three problems for Broome's argument. In particular, I showed that Broome's use of Risk overgeneralizes and categorizes even innocuous activities in the same basket as GHG acts. Furthermore, Risk makes Broome's account unstable, making an agent powerless since Broome's reasoning applies equally well to the omission of GHG acts. Risk is also problematic—or so I argued—because it is in stark tension with Broome's defense of offsetting. I also argued against Diff and Broome's use of the SCC by showing that Broome's use of the SCC employs some problematic assumptions.

I close by noting the upshot of my argument on wider issues. The problem of individual denialism is a collective action problem involving tipping points.[53] There are many more problems with a similar structure: voting for a responsible electoral candidate, consuming factory-farmed meat, and checking one's microaggressions, to name a few. Broome's mistake—I suggest—is in neglecting the collective dimension of the problem of individual denialism. By focusing only on individual acts and their (expected) harms, Broome misses what makes the problem of individual denialism puzzling: the complicated interaction between an agent *qua* an individual and the agent *qua* a member

[50] Nordhaus, "Revisiting the Social Cost of Carbon," 1519.

[51] Nordhaus, "Revisiting the Social Cost of Carbon," 1519.

[52] Golosov et al., "Optimal Taxes on Fossil Fuel in General Equilibrium," 50.

[53] See Lenton et al., "Tipping Elements in the Earth's Climate System," for a thorough discussion on the tipping points associated with climate change. They list fifteen policy-relevant tipping elements in the climate, where tipping elements are "subsystems of the Earth system that are at least subcontinental in scale and can be switched—under certain circumstances—into a qualitatively different state by small perturbations" (1786). The switch or transition point is defined as the "tipping point."

of a collective. I think that any solution to the problems of voting, individual denialism, consumption of factory-farmed meat, and microaggressions that only appeals to the individual will face analogous problems to the ones I have raised in this paper. It is only in taking the collective dimension of these problems seriously that we can make progress in solving these difficult issues.[54]

University of Pittsburgh
ksb75@pitt.edu

References

Ahmed, Arif. *Evidential Decision Theory*. Cambridge: Cambridge University Press, 2021.

Batterman, Robert W. "Defining Chaos." *Philosophy of Science* 60, no. 1 (March 1993): 43–66.

Beebee, Helen. "Causing and Nothingness." In Collins, Hall, and Paul, *Causation and Counterfactuals*, 291–308.

Bishop, Robert C. "Chaos." *Stanford Encyclopedia of Philosophy* (Spring 2017). https://plato.stanford.edu/entries/chaos/.

———. "Metaphysical and Epistemological Issues in Complex Systems." In *Philosophy of Complex Systems*, edited by Cliff Hooker, 119–50. Amsterdam: North Holland, 2010.

Bowman, Paul. "The Relevance of Motivations to Wrongdoing for Contributing to Climate Change." In *Studies on Climate Ethics and Future Generations*, vol. 3, edited by Joe Roussos and Paul Bowman, 137–61. Stockholm: Institute for Futures Studies, 2021.

Briggs, R. A. "Normative Theories of Rational Choice: Expected Utility." *Stanford Encyclopedia of Philosophy* (Fall 2019). https://plato.stanford.edu/entries/rationality-normative-utility/.

Broome, John. "Against Denialism." *Monist* 102, no. 1 (January 2019): 110–29.

———. *Climate Matters: Ethics in a Warming World*. New York: Norton, 2012.

Buchak, Lara. "Decision Theory." In *Oxford Handbook of Probability and*

[54] See Hormio, "Can Corporations Have (Moral) Responsibility Regarding Climate Change Mitigation?"; and Bowman, "The Relevance of Motivations to Wrongdoing for Contributing to Climate Change," for recent attempts to solve the general problem of ID by taking seriously the collective dimension of ID.

Many thanks to two anonymous reviewers of *JESP*, Professor Ian Rumfitt, and Elliot Thornley for many helpful comments on earlier drafts of this paper. Thanks also to the Felix Scholarship Trustees for supporting my studies on the BPhil at the University of Oxford.

Philosophy, edited by Christopher Hitchcock and Alan Hájek, 789–814. Oxford: Oxford University Press, 2016.

———. *Risk and Rationality*. Oxford: Oxford University Press, 2013.

Campbell, Tim. "Offsetting, Denialism, and Risk." In *Studies on Climate Ethics and Future Generations*, vol. 3, edited by Joe Roussos and Paul Bowman, 125–36. Stockholm: Institute for Futures Studies, 2021.

Collins, John, Ned Hall, and L. A. Paul, eds. *Causation and Counterfactuals*. Cambridge, MA: MIT Press, 2004

Cripps, Elizabeth. *Climate Change and the Moral Agent: Individual Duties in an Interdependent World*. Oxford: Oxford University Press, 2013.

Dowe, Phil. *Physical Causation*. Cambridge: Cambridge University Press, 2000.

Duijf, Hein, Allard Tamminga, and Frederik Van De Putte. "An Impossibility Result on Methodological Individualism." *Philosophical Studies* 178, no. 12 (December 2021): 4165–85.

Easwaran, Kenny. "Strong and Weak Expectations." *Mind* 117, no. 467 (July 2008): 633–41.

Golosov, Mikhail, John Hassler, Per Krusell, and Aleh Tsyvinski. "Optimal Taxes on Fossil Fuel in General Equilibrium." *Econometrica* 82, no. 1 (January 2014): 41–88.

Hall, Ned. "Two Concepts of Causation." In Collins, Hall, and Paul, *Causation and Counterfactuals*, 225–76.

Hayward, Clare, and Dominic Rosers. *Climate Justice in a Non-Ideal World*. Oxford: Oxford University Press, 2016.

Hope, Chris. "Discount Rates, Equity Weights and the Social Cost of Carbon." *Energy Economics* 30, no. 3 (May 2008): 1011–19.

———. "The Marginal Impact of CO_2 from PAGE2002: An Integrated Assessment Model Incorporating the IPCC's Five Reasons for Concern." *Integrated Assessment* 6, no. 1 (2006): 19–56.

———. "Optimal Carbon Emissions and the Social Cost of Carbon over Time under Uncertainty." *Integrated Assessment* 8, no. 1 (2008): 107–22.

Hope, Chris, and David Newbery. "Calculating the Social Cost of Carbon." Cambridge Working Papers in Economics EPRG Working Paper, University of Cambridge, 2007. https://doi.org/10.17863/CAM.5125.

Hormio, Säde. "Can Corporations Have (Moral) Responsibility Regarding Climate Change Mitigation?" *Ethics, Policy and Environment* 20, no. 3 (2017): 314–32.

Kaiserman, Alex. "Responsibility and the 'Pie Fallacy.'" *Philosophical Studies* 178, no. 11 (November 2021): 3597–616.

Kingston, Ewan, and Walter Sinnott-Armstrong. "What's Wrong with Joyguzzling?" *Ethical Theory and Moral Practice* 21, no. 1 (February 2018): 169–86.

Lane, Melissa. *Eco-Republic: What the Ancients Can Teach Us about Ethics, Virtue, and Sustainable Living*. Princeton, NJ: Princeton University Press, 2012.

Lenton, Timothy M., Hermann Held, Elmar Kriegler, Jim W. Hall, Wolfgang Lucht, Stefan Rahmstorf, and Hans J. Schellnhuber. "Tipping Elements in the Earth's Climate System." *Proceedings of the National Academy of Sciences* 105, no. 6 (February 2008): 1786–93.

Lewis, David. "Causation as Influence." *Journal of Philosophy* 97, no. 4 (April 2000): 182–97.

———. "Void and Object." In Collins, Hall, and Paul, *Causation and Counterfactuals*, 277–90.

Maltais, Aaron. "Radically Non-Ideal Climate Politics and the Obligation to at Least Vote Green." *Environmental Values* 22, no. 5 (October 2013): 589–608.

Manne, Alan S., and Richard G. Richels. "Merge: An Integrated Assessment Model for Global Climate Change." In *Energy and Environment*, edited by Richard Loulou, Jean-Philippe Waaub, and Georges Zaccour, 175–89. Boston: Springer, 2005.

Monton, Bradley. "How to Avoid Maximizing Expected Utility." *Philosophers' Imprint* 19, no. 18 (June 2019): 1–25.

Moore, Michael S. *Causation and Responsibility: An Essay in Law, Morals, and Metaphysics.* Oxford: Oxford University Press, 2009.

Nordhaus, William D. *A Question of Balance: Weighing the Options on Global Warming Policies.* New Haven: Yale University Press, 2008.

———. "Revisiting the Social Cost of Carbon." *Proceedings of the National Academy of Sciences* 114, no. 7 (January 2017): 1518–23.

Nordhaus, William D., and Joseph Boyer. *Warming the World: Economic Modeling of Global Warming.* Cambridge, MA: MIT Press, 2000.

Palmer, Brian. "Do We Exhale Carbon?" Natural Resources Defense Council. May 19, 2015. https://www.nrdc.org/stories/do-we-exhale-carbon.

Peck, Stephen C., and Thomas J. Teisberg. "CETA: A Model for Carbon Emissions Trajectory Assessment." *Energy Journal* 13, no. 1 (1992): 55–78.

Posner, Eric A., and David Weisbach. *Climate Change Justice.* Princeton, NJ: Princeton University Press, 2010.

Rentmeester, Casey. "Do No Harm: A Cross-Disciplinary, Cross-Cultural Climate Ethics." *De Ethica* 1, no. 2 (2014): 5–22.

Russell, Jeffrey S. "On Two Arguments for Fanaticism." GPI Working Paper No. 17-2021, Global Priorities Institute, October 2021. https://globalprioritiesinstitute.org/on-two-arguments-for-fanaticism-jeff-sanford-russell-university-of-southern-california/.

Schaffer, Jonathan. "Causation by Disconnection." *Philosophy of Science* 67, no.2

(June 2000): 285–300.

———. “Causes Need Not Be Physically Connected to Their Effects.” In *Contemporary Debates in Philosophy of Science*, edited by Christopher Hitchcock, 197–216. Oxford: Basil Blackwell, 2004.

Schwarcz, Joe. “Why Isn’t the Carbon Dioxide from Breathing a Concern for Global Warming?” Office for Science and Society, McGill University. March 20, 2017. https://www.mcgill.ca/oss/article/environment-quirky-science-you-asked/humans-and-animals-exhale-carbon-dioxide-every-breath-why-not-considered-be-problem-far-global.

Sinnott-Armstrong, Walter. “It’s Not My Fault: Global Warming and Individual Obligations.” In *Perspectives on Climate Change: Science, Economics, Politics, Ethics*, edited by Walter Sinnott-Armstrong and Richard B. Howarth, 285–307. Oxford: Elsevier, 2005.

Strogatz, Steven. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. 2nd ed. Boulder: Westview Press, 2014.

Tamminga, Allard, and Hein Duijf. “Collective Obligations, Group Plans and Individual Actions.” *Economics and Philosophy* 33, no.2 (July 2017): 187–214.

Tamminga, Allard, and Frank Hindriks. “The Irreducibility of Collective Obligations.” *Philosophical Studies* 177, no. 4 (April 2020): 1085–109.

Thalos, Mariam, and Oliver Richardson. “Capitalization in the St. Petersburg Game: Why Statistical Distributions Matter.” *Politics, Philosophy and Economics* 13, no.3 (August 2014): 292–313.

Thoma, Johanna. “Decision Theory.” In *The Open Handbook of Formal Epistemology*, edited by Richard Pettigrew and Jonathan Weisberg, 57–107. PhilPapers Foundation, 2019.

Tol, Richard. “On the Optimal Control of Carbon Dioxide Emissions: An Application of FUND.” *Environmental Modeling and Assessment* 2, no. 3 (October 1997): 151–63.

van den Bijgaart, Inge, Reyer Gerlagha, and Matti Liski. “A Simple Formula for the Social Cost of Carbon.” *Journal of Environmental Economics and Management* 77, no. 1 (May 2016): 75–94.

Webster, Mortm, Chris Forest, John Reilly, Mustafa Babiker, David Kicklighter, Monika Mayer, Ronald Prinn et al. “Uncertainty Analysis of Climate Change and Policy Response.” *Climatic Change* 61, no. 3 (December 2003): 295–320.

Werndl, Charlotte. “Initial-Condition Dependence and Initial-Condition Uncertainty in Climate Science.” *British Journal for the Philosophy of Science* 70, no. 4 (December 2019): 953–76.

———. “On Defining Climate and Climate Change.” *British Journal for the Philosophy of Science* 67, no. 2 (June 2016): 337–64.

———. "What Are the New Implications of Chaos for Unpredictability?" *British Journal for the Philosophy of Science* 60, no. 1 (March 2009): 195–220.

Wilkinson, Hayden. "In Defence of Fanaticism." GPI Working Paper No. 4-2020. Global Priorities Institute, September 2020 (updated January 2021). https://globalprioritiesinstitute.org/hayden-wilkinson-in-defence-of-fanaticism/.

Withers, Nikki. "How Much Does Human Breathing Contribute to Climate Change?" BBC *Science Focus*. Accessed February 2, 2022. https://www.sciencefocus.com/planet-earth/how-much-does-human-breathing-contribute-to-climate-change/.

Zimmerman, Michael J. "Sharing Responsibility." *American Philosophical Quarterly* 22, no. 2 (April 1985): 115–22.

Journal of Ethics and Social Philosophy
VOL. 24, NO. 1 · FEBRUARY 2023

https://doi.org/10.26556/jesp.v24i1.1922

PRUDENTIAL PARITY OBJECTIONS TO THE MORAL ERROR THEORY

François Jaquet

As opposed to a success theory, an error theory states that all judgments of a specific domain are false (or untrue) because they entail (or presuppose) the existence of something that actually does not exist. John Mackie famously held such a view about judgments in the moral domain.[1] On his account, all moral judgments are false because they entail the existence of "objectively prescriptive" facts when there are really no such things. While this theory was not always treated as a serious alternative to expressivism, subjectivism, naturalism, and nonnaturalism, its popularity has increased considerably in the last two decades. It is now considered one of the main contenders in metaethics.

Until recently, most error theorists were *local* error theorists: they targeted moral judgments only. Concerning prudential judgments, they accepted a success theory—they held some such judgments to be true. This combination of a moral error theory with a prudential success theory has now come under attack. Advocates of the "prudential parity claim" maintain that the arguments supporting a moral error theory also support a prudential error theory.[2] In their view, if all moral judgments are false, then so are all prudential judgments. Not that parity claimers agree on which lesson to draw from this, far from it. Some conclude that moral judgments happen to be true; others, that both moral and prudential judgments are always false. All nonetheless converge on the claim that moral error theorists are committed to a prudential error theory.

This paper defends a prominent local error theory—the categoricity-based error theory—against such prudential parity objections. In section 1, I distinguish this theory from another account—the irreducibility-based error theory. In the remaining sections, I discuss three arguments that have been put forward in support of the parity claim. If I am correct, these objections fail when

[1] Mackie, *Ethics*.

[2] Bedke, "Might All Normativity Be Queer?"; Cline, "The Tale of a Moderate Normative Skeptic"; Fletcher, "Pain for the Moral Error Theory?"

targeted at the categoricity-based error theory, and they fail because they conflate it with the irreducibility-based error theory.

1. The Categoricity-Based Moral Error Theory

Whatever the phrase "*the* moral error theory" implies, there are many moral error theories. But most share a basic structure: they combine the conceptual claim that moral judgments entail the existence of certain entities with the ontological claim that these entities do not exist. Where they diverge is around the entities in question. It has thus been argued that moral judgments are false because they entail the existence of free will, explanatorily dispensable facts, irreducibly normative facts, and nonconventional categorical reasons.[3]

Let me elaborate on the latter view, as this is the theory I will defend. On this account, moral judgments are false because they entail the existence of reasons that would be both nonconventional (i.e., independent from any institution or convention) and categorical (i.e., independent from their bearer's ends or desires), but all the reasons we have are either conventional or hypothetical. You may well have a conventional categorical reason not to speak with your mouth full (derived from a local norm of etiquette) or a nonconventional hypothetical reason to quit smoking (derived from your desire not to get cancer), but you cannot have a nonconventional categorical reason not to set a cat on fire. Yet, this is precisely the reason whose existence is entailed by the moral judgment that you should not set a cat on fire. Hence, this judgment is false. Let me motivate this view somewhat.

Start with the conceptual claim, and consider again the judgment that you ought not to set a cat on fire. This judgment entails that you have a reason not to set a cat on fire—it would hardly make any sense to say that you ought not to set a cat on fire but have no corresponding reason. What kind of reason is that? By contrast with your reason not to speak with your mouth full, your reason not to set a cat on fire does not seem to depend on your partaking in some institution. By contrast with your reason to quit smoking, it does not seem to depend on your ends: you should not set a cat on fire whatever you happen to desire. Hence, the judgment that you ought not to set a cat on fire entails that you have a reason that is both nonconventional and categorical.

[3] Haji, *Deontic Morality and Control*; Hinckfuss, *The Moral Society*; Olson, *Moral Error Theory*; Streumer, *Unbelievable Errors*; Mackie, *Ethics*; Garner, "On the Genuine Queerness of Moral Properties and Facts"; Joyce, *The Myth of Morality*; Kalf, *Moral Error Theory*.

As for the ontological claim, it stems from the combination of two theses.[4] On the one hand is an instrumentalist theory of nonconventional reasons: a subject *S* has a nonconventional reason to *ϕ* iff *S*+—an idealized counterpart who differs from *S* only in that she is fully informed and deliberates flawlessly—would advise *S* to *ϕ*. On the other hand is a constraint on *S*+'s advice for *S*: *S*+'s advice for *S* necessarily depends on *S*'s desires. Of course, they need not always coincide. It may be that *S* wants to *ϕ* because she makes a mistake in her deliberation, in which case *S*+ would not advise *S* to *ϕ*. Yet, being *S*'s counterpart, what she would advise *S* to do depends on what *S* desires. Assuming that Jim and Pam have very different aspirations, Jim+ would advise Jim to do things that Pam+ would advise Pam not to do. From this combination of claims, it follows that all the nonconventional reasons we have depend on our desires and, *a fortiori*, that there are no nonconventional categorical reasons.

Why accept instrumentalism about nonconventional reasons? Here is a suggestion.[5] There is something odd about the question "A rational version of myself would advise me to *ϕ*, but so what?" I would certainly have a reason to perform an act such a version of myself would recommend to me. The question "So what?" would be off the mark. By contrast, there seems to be nothing odd about the question "An impartial observer would advise me to *ϕ*, but so what?" or about the question "*ϕ*-ing would maximize aggregate pleasure, but so what?" Intuitively, it makes rational sense to ignore the advice of an impartial observer or the amount of pleasure in the world, not so much to ignore the advice of an ideally rational version of oneself. *That* would be plainly irrational. Instrumentalism accounts for this fact.

Now, as I said, this categoricity-based error theory is often conflated with another one. On this alternative view, moral judgments are uniformly false not because they entail the existence of nonconventional categorical reasons but because they entail the existence of irreducibly normative facts—that is, normative facts that do not reduce to ordinary, empirically accessible, natural facts.[6] The purported fact that you ought not to set a cat on fire, for instance, is distinct from

[4] For the sake of illustration, I use Richard Joyce's case for the ontological claim (see *The Myth of Morality*, 68–76). This is the most accessible version of an argument shared with other proponents of this error theory.

[5] Joyce, *The Myth of Morality*, 81–85.

[6] Streumer, *Unbelievable Errors*. When it comes to normative facts, I use the predicates "nonnatural" and "irreducibly normative" interchangeably. This is a simplification—strictly speaking, not all nonnatural facts are irreducibly normative. A normative fact that would be reducible to a supernatural fact could not be properly described as irreducibly normative. But this simplification is harmless, since we can assume that supernatural facts are ontologically dubious.

the fact that setting a cat on fire would cause the cat unnecessary pain—or from any other natural fact, for that matter. And that is what makes it ontologically suspicious. Call this other error theory the "irreducibility-based error theory."

Although these theories are often lumped together, they are distinct. One can consistently accept the categoricity-based error theory's ontological claim—namely, that there are no categorical reasons—and yet reject the irreducibility-based error theory's ontological claim—namely, that there are no irreducibly normative facts. This is because one can consistently believe in the existence of nonnatural facts and yet accept instrumentalism about reasons. This possibility is obfuscated by a common mistake, that of equating instrumentalism with the following meta-normative account: the fact that *S* has a reason to ϕ is identical to the fact that *S*+ would advise *S* to ϕ. This meta-normative view entails that there are no irreducibly normative reasons, and presumably no irreducibly normative facts more generally.

This equation is misguided, for instrumentalism is not a meta-normative view, as can be seen through an analogy with utilitarianism. While utilitarians maintain that an act is right if and only if it maximizes pleasure, they are not committed to the metaethical view that moral facts are facts about pleasure. Likewise, instrumentalists maintain that *S* has a reason to ϕ if and only if *S*+ would advise *S* to ϕ, but they are not committed to the meta-normative view that facts about reasons are ultimately facts about rational counterparts. Just as utilitarianism is a first-order view that remains silent about the reduction of moral facts to facts about pleasure, instrumentalism is a first-order view that remains silent about the reduction of facts about reasons to facts about rational counterparts.[7]

As a result, categoricity-based error theorists may ban nonconventional categorical reasons from their ontology and yet let nonnatural facts in—and thereby reject the irreducibility-based error theory's ontological claim. Maybe this will feel like an unstable position to some philosophers. After all, a key motivation behind instrumentalism lies in its ontological elegance: the view is appealing, those philosophers would say, mainly because it postulates the existence of nothing beyond natural entities. Why, then, would one accept both nonnaturalism and instrumentalism?

[7] Dorsey, "Idealization and the Heart of Subjectivism," 217. One might object that a first-order version of instrumentalism cannot yield the needed claim that reasons depend on desires because such dependence claims belong to meta-normative theory rather than first-order normativity. I disagree. The first-order view that an act is right if and only if it maximizes pleasure entails that an act's rightness depends on its effects on overall pleasure. In the same way, the first-order view that *S* has a reason to ϕ if and only if *S*+ would advise *S* to ϕ entails that an agent's reasons depend on their desires. Because utilitarianism and instrumentalism provide us with criteria for rightness and reasons, they have implications regarding the grounds of rightness and reasons. This does not make them meta-normative views.

Here is why. One might buy into the above case for instrumentalism and yet be convinced of the existence of nonnatural facts by the following companions-in-guilt argument. Some philosophical judgments are true—the judgment that personal identity consists in psychological continuity (or that it consists in something else), the judgment that there are universals (or that there are no such things), and the judgment that moral properties are irreducibly normative (or that they are not). But these judgments do not state natural facts—the facts they state are not empirically accessible. Hence, true philosophical judgments are made true by nonnatural facts.[8] Someone sensitive to this companions-in-guilt argument will reject the irreducibility-based error theory. She will nevertheless subscribe to the categoricity-based error theory insofar as she accepts instrumentalism.

Turning to prudence, most proponents of the categoricity-based error theory hold that prudential judgments involve no commitment to nonconventional categorical reasons. The judgment that you should quit smoking entails that you have a reason to quit smoking, and this reason transcends all conventions—prudentially speaking, you should quit smoking, whichever institutional practices you partake in. But this reason arguably applies only to the extent that you do not want to get cancer (or would not if you deliberated flawlessly). If getting cancer was your plan, then you should keep smoking. Thus, the reasons entailed by prudential judgments seem to be hypothetical and therefore conform to instrumentalism about nonconventional reasons. This suggests that the categoricity-based error theory does not generalize to prudential judgments.[9]

Some philosophers deny that. In order to criticize local error theories, they rely on the following parity claim: if any moral error theory is true, then an analogous prudential error theory will be true. Notice how general this parity claim is. Instead of targeting a specific local error theory, it is meant to apply to *all* local error theories—or at least to all *plausible* local error theories. Hence, it should apply to the categoricity-based local error theory. In the remaining sections, I will argue that it does not. Three arguments have been advanced in favor of the prudential parity claim. The first is based on the alleged irreducibility of prudential facts; the second, on the lack of a story about the normativity of hypothetical reasons; the third, on the very nature of reasons. I shall argue

[8] Of course, these facts are not irreducibly normative—they are not normative in the first place. However, if irreducibly normative facts are queer, that must be because they do not reduce to ordinary natural facts.

[9] For the sake of presentation, I will set aside conventional reasons from now on. This will be innocuous since all the parties in this debate agree that moral and prudential reasons are nonconventional.

that these arguments fail against the categoricity-based local error theory. As will become apparent in the process, they seem to work against this error theory only so long as one confuses it with the irreducibility-based error theory.

2. Fletcher and Irreducible Normativity

Guy Fletcher uses the prudential parity claim as part of a companions-in-guilt defense of a moral success theory. This is his reasoning. If a moral error theory were true, then an analogous prudential error theory would be true. But all prudential error theories are false for, surely, there are prudential truths: "it seems undeniable that some things are good or bad for people, that some people lead lives that go well for them (or vice versa), and that some outcomes are better (or worse) than others for someone."[10] By way of consequence, all moral error theories are false.[11]

Fletcher's parity claim rests on the contention that prudential judgments resemble moral judgments in that they entail the existence of irreducibly normative facts and properties.[12] In support of this contention, Fletcher appeals to a thought experiment inspired by Terrence Horgan and Mark Timmons' Moral Twin Earth scenario.[13] So, let us start with a quick reminder of Horgan and Timmons's scenario. Suppose that moral judgments on Earth are generally in line with hedonistic utilitarianism: Earthlings will say that an act is right just in case that act maximizes pleasure. Next, imagine another planet, Twin Earth, similar to Earth in all respects except that its (seemingly) moral judgments align with Kantian deontology: Twin Earthlings will say that an act is right just in case its maxim is universalizable. Apart from that, they use the term "right" just as we do—namely, to evaluate actions.

This description of both planets invites the following observation. If moral judgments ascribed natural properties, then the term "right" would refer to the

[10] Fletcher, "Pain for the Moral Error Theory?" 478.

[11] Fletcher phrases his parity claim in terms of various arguments for the moral error theory rather than various moral error theories: if an argument suffices to establish the moral error theory, then an analogous argument will suffice to establish the prudential error theory. This is because he uses the label "moral error theory" to refer to the claim that all moral judgments are false. By contrast, I take a moral error theory to consist of this claim together with an argument in its support. Since this difference is purely terminological, I take the liberty to rephrase Fletcher's point in my own vocabulary.

[12] On Fletcher's characterization, a property is irreducibly normative if it "cannot be identified with any ontologically innocent, natural, property" ("Pain for the Moral Error Theory?" 475). Accordingly, a fact or property is irreducibly normative if and only if it is nonnatural. This is in line with my use of "irreducibly normative" as a synonym for "nonnatural."

[13] Horgan and Timmons, "New Wave Moral Realism Meets Moral Twin Earth."

property of maximizing pleasure on Earth and to the property of falling under a universalizable maxim on Twin Earth. But then Earthlings and Twin Earthlings would not disagree about moral value; they would talk past each other. This implication is counterintuitive—Earthlings and Twin Earthlings very much seem to disagree about morality. All this suggests that moral judgments do not ascribe natural properties but nonnatural properties.

Fletcher constructs a similar thought experiment about prudence.[14] Suppose that prudential judgments on Earth are generally in line with hedonism: Earthlings will say that a state of affairs is good for a subject just in case that state of affairs maximizes her pleasure. Next, imagine another planet, Twin Earth, that is similar in all respects except that its (seemingly) prudential judgments align with the desire satisfaction theory: Twin Earthlings will say that a state of affairs is good for a subject just in case that state of affairs best satisfies that subject's informed preferences. Apart from that, they use the term "good" just as we do—namely, to evaluate states of affairs.

Fletcher makes the following observation: "It seems plausible that agents in both worlds make prudential judgments and that they manifest a disagreement in conceptions of prudential value.… This suggests that we cannot reduce prudential properties to hedonic properties in particular, and it seems like the argument can be run with equal plausibility for other ontologically innocent properties."[15] If prudential judgments ascribed natural properties, then Earthlings and Twin Earthlings would not disagree about prudential value; they would talk past each other. Yet, they do disagree about prudential value. Hence, prudential judgments do not ascribe natural properties. Just like moral judgments, they ascribe nonnatural properties.

While this thought experiment indicates that prudential properties are nonnatural, it does not support Fletcher's general prudential parity claim. Surely, it supports *a* prudential parity claim: if moral judgments are uniformly false because they presuppose the existence of nonnatural facts, then prudential judgments are uniformly false too. Fletcher's thought experiment is evidence that those moral error theorists who locate moral queerness in irreducible normativity are committed to a prudential error theory. However, as we saw earlier, not all moral error theorists locate moral queerness in irreducible normativity. In particular, those error theorists who locate moral queerness in categoricity are immune to this objection. They can grant that prudential facts are irreducible to natural facts and yet welcome them in their ontology, provided that

[14] Fletcher, "Pain for the Moral Error Theory?" 479.

[15] Fletcher, "Pain for the Moral Error Theory?" 480.

these facts generate only hypothetical reasons. And this seems sufficient for Fletcher's parity claim to be toothless against them.

Not so fast, Fletcher would respond. In his opinion, local error theorists cannot appeal to this strategy, for *all* moral error theorists must locate queerness in irreducible normativity; *all* moral error theorists must deny the existence of nonnatural facts and properties. This is his argument:

> The property commonly focused upon in discussion of error theory is the property of being a categorical reason for action.... However, the moral error theorist is not only worried about moral *reasons*. The scope of Mackie's error theory is clear when he writes: "There are no objective values.... The claim that values are not objective, are not part of the fabric of the world, is meant to include not only moral goodness ... but also other things that could be more loosely called moral values or disvalues—rightness and wrongness, duty, obligation." According to error theorists, moral discourse is committed to an *array* of normative properties that are *sui generis*, irreducibly normative, features of actions, etc. In holding that the properties ascribed within moral discourse are *sui generis* irreducibly normative features of actions, the error theorist contends that they cannot be identified with any ontologically innocent, natural, property.[16]

Since all moral error theorists are concerned not only with moral reasons but also with moral values, duties, and obligations, they *must* say that these entities are queer too. But these entities cannot be queer because they are categorical reasons—they are not. If they are queer at all, this must be because they are irreducibly normative. Accordingly, all moral error theorists must locate queerness in irreducible normativity. And all moral error theorists must say that prudential facts and properties are queer too.

What should we make of this argument? Certainly, moral error theorists are not concerned solely with moral reasons; they deny the existence of moral facts and properties generally. That they take these facts and properties to be nonnatural is generally true as well. But it does not follow that they must object to the existence of moral facts and properties *on the grounds that these entities would be nonnatural*. They can coherently object to the existence of these facts and properties *on the grounds that they would generate categorical reasons*. On this view, the fact that you should not set a cat on fire is queer, and it may well be nonnatural. Nevertheless, it is not queer *because* it is nonnatural. It is queer because it would provide you with a queer, categorical reason not to set a cat on

[16] Fletcher, "Pain for the Moral Error Theory?" 475.

fire. In order to reject moral values, duties, and obligations, moral error theorists need therefore not say that these entities are queer because they are nonnatural.

Let me sum up the present section. Fletcher might be correct to the extent that both moral and prudential facts and properties are nonnatural. But this does not entail that both moral and prudential reasons are categorical—prudential reasons might be hypothetical even though prudential facts and properties are nonnatural. Since the error theory under scrutiny locates moral queerness in categoricity, it is therefore not committed to the queerness of prudential reasons, nor does it entail a prudential error theory. Insofar as it is carefully distinguished from an irreducibility-based error theory, it evades Fletcher's prudential parity objection.

3. Cline and the Normativity of Hypothetical Reasons

Fletcher uses the prudential parity claim to build a companions-in-guilt argument against moral error theorists. The philosopher whose view we are going to discuss now instantiates an opposite trend. Brendan Cline appeals to the prudential parity claim in defense of a global error theory. Another difference is that he specifically targets categoricity-based error theorists.

Here is his argument. Categoricity-based local error theorists criticize moral success theorists for lacking an account of the normativity of categorical reasons. Yet, they lack an account of the normativity of hypothetical reasons. As Cline puts it, "there is simply no story offered about how the prescriptive force of desire works," "no positive story about the normativity of desire."[17] But, the argument continues, if the absence of a story of the former kind suffices to establish that all moral judgments are false—and indeed it does—then the absence of a story of the latter kind suffices to establish that all prudential judgments are false. By their own lights, categoricity-based error theorists should apply their criticism to prudential judgments.[18]

[17] Cline, "The Tale of a Moderate Normative Skeptic," 155, 156.

[18] This is actually only one of two arguments Cline opposes to a local error theory. The argument I leave aside relies on a general parity claim: any moral error theory entails a global error theory. In support of this claim, Cline points out that normative cognition is a unified system rather than a set of diversified mechanisms corresponding to moral norms, prudential norms, epistemic norms, and so on. However, this case rests on Stephen Stich's account of normative psychology, which is highly contentious and does not clearly entail that all normative reasons are of the same kind (see Kumar, "Moral Judgment as a Natural Kind"; Joyce, "Replies"). There surely are similarities between all normative thoughts—the opposite would be surprising. But nothing indicates that all normative thoughts are similar in that they entail the existence of categorical reasons. Yet, this is what Cline would need to show to back up his parity claim.

Cline's parity claim hinges on two claims. First, categoricity-based error theorists deny the existence of moral facts on the grounds that we lack an account of the normative significance of categorical reasons. Second, they lack an account of the normative significance of hypothetical reasons. What should we make of these two claims? I am reluctant to give this question a clear-cut answer as it is unclear to me what is meant by the phrase "a story about the normativity of." Nonetheless, my impression is that (at least) one of these two claims will be false no matter what interpretation is correct.

Does any argument to the effect that reasons depend on desires constitute such a story? If so, then Cline's second claim is false, for categoricity-based error theorists do supply an account of the normativity of desires in this sense. As we saw above, they put forward the following argument: a subject *S* has a reason to *ϕ* if and only if *S*+ would advise *S* to *ϕ*, but *S*+'s advice for *S* necessarily depends on *S*'s desires, so *S*'s reasons necessarily depend on *S*'s desires. These error theorists even deliver a defense of instrumentalism: there is something very odd about the question, "A version of myself who would be fully informed and reason flawlessly would advise me to *ϕ*, but so what?"

Suppose alternatively that this is not the kind of story Cline is after. Maybe he is rather demanding an account of *how* facts about desires could have that kind of normative force, or *how* facts about desires could give us reasons. The concern would be that one hardly understands how hypothetical reasons could have genuine normative force if they were identical to natural facts about desires. Categoricity-based error theorists might be compelled to appeal to irreducibly normative facts to account for the normativity of hypothetical reasons. Notice, however, that the issue would then pertain to meta-normative theory, a domain in which categoricity-based error theorists remain silent. These philosophers do not deny the existence of moral facts because we lack a meta-normative conception of the relation between natural facts and moral facts that would account for the significance of categorical reasons; they deny the existence of moral facts because these would generate categorical reasons that we do not have according to instrumentalism.

To see the point more clearly, recall that these error theorists can combine their instrumentalist take on prudential reasons with a nonnaturalist account of prudential facts. If this is their position, they will quietly concede that they lack a positive story about the normativity of desires and yet insist that instrumentalism is true; they will concede that they cannot explain *how* desires ground reasons and yet insist that they do. This should come as no surprise. After all, if both prudential instrumentalism and meta-prudential nonnaturalism are true, then it is a brute fact, an ungrounded metaphysical truth, that we have a prudential reason to do something just in case our rational counterparts would advise us to. Compare: if both hedonistic utilitarianism and metaethical nonnaturalism are true, then it is a brute fact, an ungrounded metaphysical truth, that an act is right just in case it maximizes pleasure. If ethical nonnaturalism is true, utilitarians need not explain why facts about pleasure have normative force. The same is true of instrumentalists if prudential nonnaturalism is true.

To recap, what we have is a dilemma for Cline's parity claim. Either a story about the normativity of *X* amounts to a substantive argument to the effect that a certain normative property depends on *X*, in which case categoricity-based error theorists do supply us with a story about the normativity of desires. Or a story about the normativity of *X* amounts to a meta-normative account of the relation between *X* and normative facts, in which case categoricity-based error theorists do not reject moral facts because we lack a story about the normativity of categorical reasons. Either way, no analog to their argument will suffice to establish that all prudential judgments are false, and Cline's prudential parity claim collapses. Just like Fletcher's, it appears plausible only so long as one conflates the categoricity-based error theory with the irreducibility-based error theory.

4. Bedke and the Nature of Reasons

In contrast with both Fletcher and Cline, Mathew Bedke defends neither a moral success theory nor a prudential error theory. Although he relies on the prudential parity claim, he remains neutral as to which conclusion to draw from it. His point is merely that prudential reasons are just as queer as moral reasons. Whether both kinds are queer enough to warrant a ban from our ontology is an issue he leaves for another day.

Bedke's defense of the parity claim is straightforward.[19] At the bottom, reasons refer to relations. They may be represented formally as predicates of the form $R(F, S, \phi)$ in *C*. A subject *S* has a reason to ϕ in conditions *C* if and only if, in *C*, there is a fact *F* that counts in favor of *S*'s ϕ-ing. Pam has a moral reason to pay her taxes given her living standard if and only if, given her living standard, the fact that taxes help fund public schools counts in favor of Pam's paying her taxes. It follows from this account of reasons that if there is anything queer about moral reasons, then it must be one of these elements: the conditions *C*, the fact *F*, the subject *S*, the action ϕ, or the relation *counting in favor of*. But *C*, *F*, *S*, and ϕ are very unlikely to be queer: Pam's living standard, the fact that taxes help fund public schools, Pam herself, and the act of paying one's taxes are ontologically banal entities. The only thing that could be objectionable about moral reasons is the relation *counting in favor of*.

[19] Bedke, "Might All Normativity Be Queer?" 48–51.

But then, Bedke proceeds, prudential reasons must be queer too since they involve the same favoring relation. Just like moral reasons, prudential reasons are predicates of the form $R(F, S, \phi)$ in C. Jim has a reason to jog given that his bones are fragile if and only if, given that his bones are fragile, the fact that jogging strengthens bones counts in favor of Jim's jogging. Just like Pam's moral reason, Jim's prudential reason involves the relation *counting in favor of*. If one is queer, so is the other. Hence a general version of the prudential parity claim: if any queerness-based argument establishes that all moral judgments are false, then an analogous argument will establish that all prudential judgments are false.

Central to Bedke's demonstration is the claim that error theorists must locate moral queerness in the relation *counting in favor of*. In particular, assuming as he does that the facts that do the favoring are ordinary natural facts, there cannot be anything suspicious about them: the fact that taxes help fund public schools and the fact that jogging strengthens bones are ontologically irreproachable. But one might reject this assumption and argue instead that the facts that do the favoring must be evaluative. For example, the fact that paying taxes is morally good might count in favor of Pam's paying her taxes. Likewise, the fact that jogging is good for Jim might count in favor of Jim's jogging. Call this the "value-first view."

On this account of the nature of reasons, one might argue that the fact that does the favoring is queer in the moral case (where it involves objective, desire-independent goodness) while it is not in the prudential case (where it involves only subjective, desire-dependent goodness). There is something queer about the moral fact that Pam's paying her taxes is good regardless of her desires; there is nothing queer, by contrast, about the prudential fact that jogging is good for Jim in light of his desires. Accordingly, the parity claim would be false: one argument that suffices to establish the falsity of all moral judgments would have no analog sufficient to establish the falsity of all prudential judgments. This sounds like a natural line of objection for categoricity-based error theorists.

In response to this objection, Bedke maintains that prudential reasons would be queer even if the favoring facts were evaluative: "the value-first view does not do away with reason relations; it simply relocates them between evaluative considerations and actions rather than non-evaluative considerations and actions. In addition to that, the view introduces another metaphysical entity subject to queerness."[20] In other words, what counts in favor of Jim's jogging may be the natural fact that jogging strengthens bones or the evaluative fact that jogging is good for Jim. In either case, the relation *counts in favor of* will appear somewhere in the equation. Since this relation is queer, prudential reasons will also involve a queer element, even on the value-first view. The parity claim still stands.

[20] Bedke, "Might All Normativity Be Queer?" 51.

This rejoinder is unconvincing. Why should the value-first theorist admit that the favoring relation is queer at all? Once one buys into evaluative facts, *counting in favor of* appears to come as a free bonus. Assuming that paying taxes is good, it is pretty clear that this fact counts in favor of paying one's taxes. In what sense would paying taxes be good if that did not count in favor of doing it? Local error theorists might therefore just as well contend that only the favoring evaluative fact (objective, desire-independent goodness) is queer in the moral case. Since the favoring evaluative fact (subjective, desire-dependent goodness) is not queer in the prudential case, prudential reasons would not be queer, and the parity claim would turn out false.[21]

But this is not the main point I want to stress, so let us grant Bedke that the value-first view is flawed for the sake of argument. A more serious concern with his demonstration has to do with the assumption that the queerness of moral reasons must be traced to the queerness of one of their components. Admittedly, this assumption is plausible enough as long as queerness consists in irreducibility. If moral reasons are queer because they are nonnatural, then they must inherit their queerness from the conditions *C*, the fact *F*, the subject *S*, the action *ϕ*, or the relation *counting in favor of*—moral reasons could not be nonnatural if all their components were reducible to natural entities. Bedke's objection might therefore work against an irreducibility-based local error theory.

Be that as it may, it is powerless against the categoricity-based error theory. The idea that moral reasons must inherit their queerness from one component of the moral-reason relation is much less plausible if queerness consists in categoricity. On that assumption, there need be nothing queer about the conditions *C*, the fact *F*, the subject *S*, the action *ϕ*, and the relation *counting in favor of*. Reasons will be queer only when they combine some such elements—only when a fact that obtains regardless of the agent's desires is supposed to count in favor of an act. Since this would happen in the case of (categorical) moral reasons, moral reasons are queer. However, this does not happen in the case of (hypothetical) prudential reasons. As a result, these are ontologically respectable. Again, this sounds like a perfectly natural line of thinking for proponents of a local error theory based on categoricity.

[21] Bedke levels two independent objections against the value-first view: natural facts are all it takes to account for practical reasons, and the idea that both a natural fact and an evaluative fact count in favor of every rational act involves a problematic kind of double counting ("Might All Normativity Be Queer?" 52). As they stand, I find both objections unpersuasive. The first appears to beg the question against the value-first theorist, who presumably believes that evaluative facts *are* needed to account for reasons—why would she not? The second commits a strawman fallacy: according to the value-first theorist, only the evaluative fact does the favoring. No double counting is involved there.

Bedke anticipates something like this objection. On the view he discusses, we should distinguish between an objective favoring relation (where *F* is unrelated to *S*'s desires) and a subjective favoring relation (where *F* depends on *S*'s desires). In response, he argues that this distinction is confused: "I do not know what it would mean to say that some of these considerations favour objectively versus favour subjectively. They either favour or they do not."[22] So far, we are in agreement. I do not mean to suggest that we are dealing with different favoring relations depending on the nature of the favoring fact. But Bedke goes further and concludes that, if anything is queer, it has to be the favoring relation itself. He supports this assertion with an analogy:

> Suppose that someone claims that certain witches are not metaphysically queer, viz., witches who can cast spells on themselves only. These witches are "subjectively magical," the claim goes, and so the queerness objection does not apply to them. The appropriate reply is that the queerness resides in spell-casting quite generally. It does not matter what the spells affect. I think we can say the same kind of thing to those who claim that favouring relations flow only from ends. The queerness objection applies to the favouring relations quite generally, and it does not matter wherefrom the favouring flows. If you recognize one kind of witch, objections to other witches cannot be based on metaphysical queerness; if you recognize one kind of reason, objections to other kinds of reasons cannot be based on metaphysical queerness.[23]

This is where we part ways. Bedke seems to think that whenever a relational fact is queer, this must be in virtue of one of its components: either the relation or a relatum. I disagree. We certainly want to say that facts about spells derive their queerness from one of their components. But this is because we know that one of their components is queer—namely, the relation *casting a spell on*. Now, this is an accidental feature of this example. Other relational facts appear to be queer even though all of their components are individually fine. For instance, there would be something odd about a future event causing a past event, yet there is nothing particularly odd about causation, future events, or past events. In cases like this, it is the combination of a certain relation with a certain relatum that makes the resulting relational fact suspicious.

According to the present objection, this is what happens with moral reasons. Individually, none of the components of the reason $R(F, S, \phi)$ in C is queer. Still, that reason is queer because it combines these elements in a problematic

[22] Bedke, "Might All Normativity Be Queer?" 54.

[23] Bedke, "Might All Normativity Be Queer?" 54.

way—a fact that is unrelated to the agent's desires features as a relatum of the relation *counting in favor of*. Again, this line of reasoning is very congenial to categoricity-based error theorists. For these philosophers, the relation *counting in favor of* is akin to the relation *causing* more than it is to the relation *casting a spell on*. You may admit one kind of causal relation and yet object to others on grounds of metaphysical queerness; you may admit one kind of reason and yet object to others on grounds of metaphysical queerness.

At the end of the day, Bedke's argument rests on a false dilemma. The local error theorist need not pick a queer element in the set $\{C, F, S, \phi, R\}$ in order to say that $R(F, S, \phi)$ in C is queer in the moral case. Even though C, F, S, and ϕ are perfectly fine, she is therefore not committed to the view that R is queer. As a result, she can concede that prudential reasons have the relation *counting in favor of* as one of their components while maintaining that these reasons are ontologically unproblematic. Just like Fletcher's and Cline's, Bedke's prudential parity claim seems fairly plausible if queerness has to do with irreducibility, not so much if queerness is located in categoricity.

5. Conclusion

Local error theories combine a moral error theory with a prudential success theory: all moral judgments are false, but some prudential judgments are true. According to prudential parity objections, all such theories are flawed: for any argument to the effect that all moral judgments are false, there is an equally good argument to the effect that all prudential judgments are false. Notice one last time how general these objections are: they are meant to impair *all* local error theories. Consequently, they will be refuted if we can find *one* moral error theory that does not entail a prudential error theory. This is what I have attempted to do in this paper. More precisely, I argued that the categoricity-based local error theory is immune to the main three prudential parity objections on the market.

Guy Fletcher's objection succeeds to the extent that it establishes *a* parity claim: if all moral judgments are false because they state irreducibly normative facts, then all prudential judgments are false too. Unfortunately, this objection is powerless against the categoricity-based local error theory, which is consistent with the existence of irreducibly normative facts. As we have seen, Fletcher believes that all moral error theorists are committed to an irreducibility-based error theory. But we have also seen that this belief is unfounded.

Brendan Cline's parity objection targets proponents of the categoricity-based error theory specifically. Here is his parity claim: if all moral judgments are false because we lack a story about the normativity of categorical reasons, then all prudential judgments are false too, for we also lack a story about the normativity of hypothetical reasons. Alas, this parity claim is either irrelevant or false, depending on what is meant by "a story about the normativity of *X*." The parity claim is irrelevant if this means a meta-normative account of *how X* grounds normative facts, for categoricity-based error theorists do not object to moral truths on the grounds that we lack such an account. The parity claim is false if this means a substantive argument to the effect *that X* grounds a normative property, for categoricity-based error theorists do provide us with such an argument in the case of prudential reasons.

For Mathew Bedke, all moral error theorists are committed to saying that moral judgments are false because they entail the existence of the relation *counting in favor of*, but prudential reasons entail the existence of this relation too. As a consequence, if all moral judgments are false, then all prudential judgments are false too. This objection does not succeed, however, because moral error theorists need not deny the existence of the relation *counting in favor of*. Proponents of the categoricity-based error theory will simply insist that the fact that does the favoring cannot be unrelated to the agent's desires—such a combination would be queer. Ultimately, this is just another way of saying that moral reasons would be queer because they would be categorical.

In the end, all the extant defenses of the prudential parity claim seem to collapse as far as the categoricity-based error theory is concerned. This is a major worry for three reasons. First, considering that the categoricity-based error theory has at least as many defenders as the irreducibility-based error theory, it is not the kind of view that one can dismiss as unimportant. Prudential parity objections are much less interesting if they do not affect it. Second, most proponents of the irreducibility-based error theory are global error theorists, which suggests that prudential parity objections should focus primarily on the categoricity-based local error theory.[24] Finally, prudential parity claimers explicitly target leading categoricity-based error theorists. Pending a better objection, this combination of a moral error theory and a prudential success theory remains a live option.[25]

Université de Strasbourg
fjaquet@unistra.fr

[24] The main two irreducibility-based error theorists are Bart Streumer and Jonas Olson. While Streumer explicitly argues that all prudential judgments are false, Olson is very much open to that idea.

[25] This work was supported by Montreal's Center for Research on Ethics and the AHP-PREST (UMR 7117—Université de Strasbourg, Université de Lorraine, CNRS). For helpful feedback on previous drafts, I would like to thank Charles Côté-Bouchard, Florian Cova, Nils Franzén, Martin Gibert, Valéry Giroux, Jessica Isserow, Florence Larocque, Stephanie

REFERENCES

Bedke, Matthew S. "Might All Normativity Be Queer?" *Australasian Journal of Philosophy* 88, no. 1 (2010): 41–58.

Cline, Brendan. "The Tale of a Moderate Normative Skeptic." *Philosophical Studies* 175, no. 1 (January 2018): 141–61.

Dorsey, Dale. "Idealization and the Heart of Subjectivism." *Noûs* 51, no. 1 (March 2017): 196–217.

Fletcher, Guy. "Pain for the Moral Error Theory? A New Companions-in-Guilt Argument." *Australasian Journal of Philosophy* 96, no. 3 (2018): 474–82.

Garner, Richard T. "On the Genuine Queerness of Moral Properties and Facts." *Australasian Journal of Philosophy* 68 (1990): 137–46.

Haji, Ishtiyaque. *Deontic Morality and Control*. Cambridge: Cambridge University Press, 2003.

Hinckfuss, Ian. *The Moral Society: Its Structure and Effects*. Canberra, Australia: Australian National University, 1987.

Horgan, Terrence, and Mark Timmons. "New Wave Moral Realism Meets Moral Twin Earth." *Journal of Philosophical Research* 16 (1991): 447–65.

Joyce, Richard. *The Myth of Morality*. Cambridge: Cambridge University Press, 2001.

———. "Replies." *Philosophy and Phenomenological Research* 77, no. 1 (July 2008): 245–67.

Kalf, Wouter F. *Moral Error Theory*. London: Palgrave Macmillan, 2018.

Kumar, Victor. "Moral Judgment as a Natural Kind." *Philosophical Studies* 172, no. 11 (November 2015): 2887–2910.

Mackie, John L. *Ethics: Inventing Right and Wrong*. New York: Penguin Books, 1977.

Olson, Jonas. *Moral Error Theory: History, Critique, Defence*. Oxford: Oxford University Press, 2014.

Streumer, Bart. *Unbelievable Errors: An Error Theory about All Normative Judgements*. Oxford: Oxford University Press, 2017.

Leary, Stéphane Lemaire, Don Loeb, Victor Moberger, Hichem Naar, Jonas Olson, Nicolas Olsson Yaouzis, Olle Risberg, Raffaele Rodogno, Sophia Rousseau-Mermans, Juliette Roussin, Christine Tappolet, and Silvan Wittwer.

Journal of Ethics and Social Philosophy
VOL. 24, NO. 1 · FEBRUARY 2023

https://doi.org/10.26556/jesp.v24i1.1659

CAN WE HAVE MORAL STATUS FOR ROBOTS ON THE CHEAP?

Sebastian Köhler

Should artificial agents (such as robots) be granted *moral status*? This seems like an important question to resolve, bearing in mind that we are very likely to encounter a growing number of increasingly sophisticated artificial agents in the not too distant future. Given that moral status is the property an entity has "if and only if *it* or *its interests* morally matter to some degree for the entity's own sake," without a clear answer about artificial agents' moral status, we risk either doing significant wrong in our interactions with such agents, or wasting significant moral concern that might be better allocated elsewhere.[1]

At the same time, many will think that before we can even start to tackle questions about the moral status of artificial agents, we first need to solve further, equally tricky issues in the philosophy of mind.[2] The "orthodox view" about moral status explains why:

> *Orthodoxy*: (Necessarily for all entities *e*) *e* has moral status only if it is (or belongs to a kind) capable of having mental states.[3]

If we accept Orthodoxy, we will think that for moral status considerations, what "goes on 'on the inside' matters greatly,"[4] as Nyholm and Frank put it. More precisely, we will believe that an entity such as an artificial agent will be eligible for moral status only if it has a mental life. As should be obvious, though, whether an entity has a mental life raises extremely controversial questions in

[1] Jaworska and Tannenbaum, "The Grounds of Moral Status," emphasis added.

[2] For an overview of the different prominent positions about moral status, see Jaworska and Tannenbaum, "The Grounds of Moral Status."

[3] Note that the prominent views disagree about what kinds of mental states are relevant for moral status. What is relevant might be consciousness, the possession of preferences, or the possession higher level cognitive capacities. However, all of these views are committed at least minimally to Orthodoxy, which is why this paper focuses on this shared assumption.

[4] Nyholm and Frank, "From Sex Robots to Love Robots," 223.

the philosophy of mind. Most importantly, this concerns the question of what it takes to have a mental life in the first place. Equally importantly, it also includes the question to what extent it is possible to have *evidence* that entities have a mental life. And this, to return to our specific context of artificial agents, raises the crucial question of whether what we *normally* take to be signs of a mind can also be taken as evidence for a mind across the board, including the *unnormal case* of artificial agents. As a result, if we are to follow Orthodoxy, there is no way around addressing such tricky questions in the philosophy of mind before settling the moral status of artificial agents, such as robots.

Given as much, one might hope that we could give a plausible account of moral status that evades these kinds of issues in the philosophy of mind that Orthodoxy imposes on us. Specifically, one might hope that there are plausible conditions on the basis of which we can grant entities moral status *without* having to undertake controversial commitments in the philosophy of mind. Let us call views of this kind "minimalism." Forms of minimalism are all views according to which we can give sufficient conditions on the basis of which we should grant entities moral status, without having to undertake commitments in the philosophy of mind. Note that minimalist views would have to satisfy at least two conditions to be plausible. First, they need to offer *sufficient* conditions for granting moral status to an entity that we can know to be satisfied without knowing whether that entity has a mind. Second, the plausibility of this criterion applied to concrete cases should not require us to combine the view with commitments in the philosophy of mind.

In recent years, a number of authors concerned with the moral status of artificial agents have suggested views that can, plausibly, be read in a minimalist vein.[5] This paper is concerned with the prospects of such views. However, the focus of the paper is not on minimalist views *generally*. Rather, the paper operates under a specific constraint. The paper brackets forms of minimalism about the moral status of artificial agents (such as those based on Luciano Floridi's *Information Ethics*, for example) that seem inherently overgeneralizing in the following sense: they imply that many more entities than we might think have moral status and, in particular, that many entities have moral status that are paradigmatic instances of entities lacking such status.[6] So, the paper focuses on views that are potentially the least revisionary forms of minimalism. Here, the paper looks at two recently suggested views that can plausibly be read

[5] See, e.g., Danaher, "Welcoming Robots into the Moral Circle"; Coeckelbergh, "Growing Moral Relations"; Coeckelbergh and Gunkel, "Facing Animals"; Gunkel, "Robot Rights" and "The Other Question"; Floridi, "Information Ethics"; Tavani, "Can Social Robots Qualify for Moral Consideration?"

[6] See Mosakas, "On the Moral Status of Social Robots," for arguments along those lines.

as views of this kind (and that have been getting some traction in the debate about the moral status of artificial agents), Ethical Behaviourism and Ethical Relationism.[7] These are here taken to be the *prima facie* most promising ways of fleshing out a minimalist view of this kind. This is so because—as will become more transparent in the discussion—they tie moral status to something it *is* paradigmatically related to, and, hence, are probably our best bet when it comes to developing forms of minimalism that are not radically revisionary (in what follows, I will use the label "minimalism" just to refer to minimalist views that are not revisionary in this sense).

With regards to these views, this paper argues that we should be pessimistic about the prospects of finding a plausible version of this kind of minimalism. Specifically, it argues that we have good reason to think that with regards to the two conditions on plausible minimalist views outlined above, views that satisfy the first condition will have to violate the second condition: views that satisfy the first condition are only plausible insofar as and because, in order to be applicable to concrete cases, they need to draw on additional assumptions in the philosophy of mind.[8] This is so, because (as I will try to show) our intuitions about moral status tend to *follow* our verdicts about what entities have a mind, and such views can accommodate this connection only by making assumptions about the mind. Hence, rather than avoiding controversies in the philosophy of mind, these views put themselves squarely within these controversies. As such, I conclude that there is no way of getting around the thorny questions in the philosophy of mind thought to plague Orthodoxy, when one is concerned with the moral status of artificial agents, unless one aims to *radically* revise the way we think about what entities have moral status.

The paper proceeds as follows: The first section discusses Ethical Behaviourism. The second Section discusses Ethical Relationism. The final section draws general lessons from the discussion and concludes.

1. ETHICAL BEHAVIOURISM

Ethical Behaviourism has recently been suggested by John Danaher.[9] He uses the position to argue that certain sorts of robots should be granted significant

[7] For Ethical Behaviourism, see Danaher, "Welcoming Robots into the Moral Circle." For Ethical Relationism, see, e.g., Coeckelbergh, "Growing Moral Relations"; Coeckelbergh and Gunkel, "Facing Animals"; Gunkel, "Robot Rights" and "The Other Question."

[8] Note that I am not arguing that these views or their proponents do *in fact* already make such assumptions. This is a matter of interpretation that I do not want to get into here. Rather, the point is that these views have to be combined with such assumptions to be plausible.

[9] Danaher, "Welcoming Robots into the Moral Circle."

moral status. The view's core tenet is this: "If an entity *A* is *roughly performatively equivalent* to another entity *B* whom, it is widely agreed, has significant moral status, then it is right and proper to afford *A* the same moral status as *B*."[10] Danaher claims that Ethical Behaviourism is only an epistemic thesis: it is supposed to be a thesis about when we have *evidence* that an entity has moral status, not a thesis about when an entity has, in fact, moral status.[11] In fact, he claims that Ethical Behaviourism is, *in principle*, compatible with the thesis that moral status is *ultimately grounded* in having mental states.

It is important to note, however, that Danaher does not simply hold that performative equivalency is, e.g., *defeasible* evidence for an entity having moral status—which is something that a proponent of Orthodoxy could accept as well.[12] Rather, he holds a much stronger view:

> This article … argues that what's going on 'on the inside' *does not matter* from an ethical perspective. Performative artifice, by itself, can be sufficient to ground a claim of moral status as long as the artifice results in *rough performative equivalency* between a robot and another entity to whom we afford moral status.[13]

That is, on Danaher's view performative equivalency is a *sufficient* condition for it to be the case that we should grant an entity a certain moral status: any entity that is performatively equivalent to another entity to which we grant moral status is to be granted the same moral status, no matter what other differences there might be between these entities. Hence, Ethical Behaviourism *starts* with a set of entities that are paradigmatic instances of entities with moral status—not considering in virtue of what they have moral status—and then assumes that we should expand the circle of which entities we attribute moral status to on the basis of performative equivalence, without considering any other differences there might be between these entities.

It should be clear that Ethical Behaviorism satisfies the first condition of minimalism, given that it provides us with a sufficient condition on the basis of

[10] Danaher, "Welcoming Robots into the Moral Circle," 2025.

[11] See Smids, "Danaher's Ethical Behaviourism," for issues with the epistemic reading of Ethical Behaviourism.

[12] Note that, in fact, the view that performative equivalency is *defeasible* evidence for moral status is not a minimalist view at all. The best explanation as to why performative equivalency is defeasible evidence for moral status is that performative equivalency is defeasible evidence for mental equivalency. However, this position takes on distinctive commitments in the philosophy of mind, e.g., that what is evidence in normal cases for a mental life (e.g., other humans, non-human animals) is also such evidence in unnormal cases (such as artificial agents).

[13] Danaher, "Welcoming Robots into the Moral Circle," 2025.

which we should grant moral status, and given that we can know whether this condition is satisfied without knowing whether the entities in question have a mind. After all, while all entities that are *paradigmatic* instances of entities with moral status are entities with a mind (namely humans and, maybe, some non-human animals), it is by no means clear that an entity that is performatively equivalent to an entity with a mind also has a mind. Furthermore, Danaher supports Ethical Behaviourism explicitly with the idea that it allows us to resolve tricky questions about whether we should grant moral status to robots without having to resolve tricky metaphysical and epistemological questions, e.g., in the philosophy of mind.[14] So, on the reading of Ethical Behaviorism suggested here, Ethical Behaviorism seems to aspire to being a form of minimalism.

At first sight, the criterion offered by Ethical Behaviourism is not implausible, at least in the sense that it is likely to give us intuitively correct verdicts with regards to moral status for a variety of normal cases. However, I will now argue that Ethical Behaviourism fails to offer a plausible version of minimalism, because the sufficient condition for granting moral status that it offers is only plausible to the extent that we combine it with certain assumptions in the philosophy of mind. If we drop these assumptions, the plausibility of Ethical Behaviourism disappears. Rather than avoiding controversial assumptions in the philosophy of mind, Ethical Behaviourism has to be combined with very robust such assumptions.

I will argue for this conclusion as follows: First, I will argue that on first sight Ethical Behaviourism faces a counterexample that is parallel to a counterexample to behaviourism in the philosophy of mind. I will then observe that Ethical Behaviourism escapes *this* counterexample only by making a move that is similar to a move made by those who adopt *functionalism* about the mind. This makes Ethical Behaviourism's verdicts about what entities have moral status co-extensional with functionalism's verdicts about what entities have a mind. I will then use a counter-example to *functionalism* about the mind to argue that this co-extensionality is *not* accidental, because the plausibility of Ethical Behaviourism depends on the plausibility of functionalism's verdicts. I will support this further by highlighting that if we adopt another view about the mind, teleo-functionalism, the plausibility of Ethical Behaviourism vanishes. Hence, to be plausible Ethical Behaviourism must be tied to functionalist assumptions in the philosophy of mind. While the argumentation here might seem more complicated than it needs to be to establish a problem for Ethical Behaviourism as a minimalist view, let me note that it takes this structure to draw out an

[14] E.g., Danaher, "Welcoming Robots into the Moral Circle," 2028.

important further point for the discussion. This is that our intuitions about moral status tend to follow our verdicts about what entities have a mind.

As I said, my argumentation starts with a case from the philosophy of mind. This case is "Blockhead."[15] To start, consider the observation that "at any point in a creature's life there are only finitely many discriminately distinct possible inputs and outputs at its periphery."[16] If this is true, we could potentially draw a finite list of all the possible inputs and outputs in the life of any human. Let us call this list that human's "game of life." Plausibly, given any human's game of life, we could build a creature that is at first sight very similar to the human, but has a chip inside of it on which the human's game of life is inscribed, where this chip fully controls the behavior of the creature. Let us call such a creature the "Blockhead twin" of the human. Plausibly, any human has a possible Blockhead twin.

Blockhead twins are an objection to *behaviorism* in the philosophy of mind. Behaviorism in the philosophy of mind is the view that entities have mental states solely in virtue of certain behavioral dispositions. For example, on such a view an entity experiences pain just in case it displays the kind of *behavior* that is typically associated with a painful experience. The reason why Blockhead twins are relevant in the context of this paper is this: Blockhead twins *are* roughly performatively equivalent to human beings. Yet Blockhead twins clearly do not have a mind—behaviorism is false. And, they also should not be granted the moral status that a human being has. Assume, for example, that you could either save a human or their Blockhead twin from being crushed by a falling boulder. Here, it is intuitively very clear that you should save the human and not the Blockhead twin. In fact, Blockhead twins quite plausibly have *no* moral status—there is no question *at all* about whom to save in the described situation. And the best explanation for this is that Blockhead twins do not have a mind. Hence, Blockhead twins also seem to constitute a counterexample to Ethical Behaviourism.

At this point, Danaher has a response available, because on his view "the concept of 'behaviour' should be interpreted broadly. It is not limited to external physical behaviours (i.e., the movement of limbs and lips); it includes all external observable patterns, including functional operations of the brain."[17] Hence, on his view Blockhead twins are not problematic, because while Blockhead twins' external physical behavior is performatively equivalent to that of humans, there is nothing in the twins that is performatively equivalent to the

[15] See Block, "Psychologism and Behaviourism"; here I draw on the presentation in Braddon-Mitchell and Jackson, "The Philosophy of Mind and Cognition," 114–19.

[16] Braddon-Mitchell and Jackson, "The Philosophy of Mind and Cognition," 116.

[17] Danaher, "Welcoming Robots into the Moral Circle," 2028.

functional operations of the brain. And this explains why Blockhead twins should not be granted the moral status that humans have.

However, in making this move, Danaher makes the verdicts about what entities should be granted moral status co-extensional with what entities have mental states according to *functionalists* about the mind.[18] According to functionalists, mental states are characterized by their relational properties, namely by their relations to inputs to and outputs from the cognitive system (what would matter on a purely behaviorist picture) and their relations amongst each other. That is, "a functionalist theory of mind specifies mental states in terms of three kinds of clauses: input clauses that say which conditions typically give rise to which mental states; output clauses that say which mental states typically give rise to which behavioural responses; and interaction clauses which say how mental states typically interact."[19] Hence, for a functionalist Blockhead twins do not have a mind, because they lack the relevant *internal* functional architecture, while anything that has the relevant functional architecture and inputs and outputs does have a mind. So, anything that should be granted moral status on the sophisticated version of Ethical Behaviourism is something that has a mind according to functionalism.

Co-extensionality by itself is not a problem. However, I will now argue that it is not a coincidence, because the plausibility of Ethical Behaviourism *depends* on the plausibility of functionalism's verdicts. To illustrate this, I will use another case from the philosophy of mind, the Global Brain.[20] Assume that there is a computer program that mimics the functional operation of a human brain at neuron by neuron level. This program is run as follows: First, each adult living on Earth is assigned the job of just one neuron (this is a highly unrealistic assumption, as even the Earth's population is not large enough). Each earthling gets a phone for this specific purpose and what they have to do is to call certain numbers, if they are called by other numbers. To tell them what to do, they have a specific list of instructions that "exactly models what their assigned neuron does, and the inputs to and outputs from their phones are connected up so as to run the program."[21] Furthermore, the network is connected to a robot body, which receives inputs from its environment in a similar way these inputs come

[18] For an introduction and overview, see Braddon-Mitchell and Jackson, "The Philosophy of Mind and Cognition," 45–64, 84–94, and 107–28; or Levin, "Functionalism."

[19] Braddon-Mitchell and Jackson, "The Philosophy of Mind and Cognition," 47.

[20] The example originates in Block, "Troubles with Functionalism." Here I draw on the updated version in Braddon-Mitchell and Jackson, "The Philosophy of Mind and Cognition," 107-8. Note that the original example is called the "China Brain." I've modified it to avoid problematic stereotyping or exoticizing that might be present in that version.

[21] Braddon-Mitchell and Jackson, "The Philosophy of Mind and Cognition," 107.

to us. The network also produces outputs that go from the robot body that, then, in actual and counterfactual circumstances behaves very similarly to a human. "Thus, the android will behave in the various situations that confront it very much as we do, despite the fact that the processing of the environmental inputs into final behavioural outputs goes via a highly organized set of [earthlings] rather than a brain."[22]

Let us call the system that consists in the robot plus the population of Earth the "Global Brain." The Global Brain is a counter-example to functionalism, because functionalism implies that Global Brain has the same mental states as a human, while it seems intuitively that it does not have mental states at all. Assume for the moment with functionalism's opponents that it is, in fact, intuitively plausible that Global Brain does not have mental states. Then it *also* seems plausible that it does not have moral status: Global Brain should *not* be granted the same moral status that is granted to a human. For example, if we face a choice between killing a human or shutting down Global Brain, then it is intuitively very clear that you should shut down Global Brain. In fact, Global Brain quite plausibly has *no* moral status, as there is no question *at all* about whom to save in this situation. Hence, the Global Brain *also* constitutes a counterexample to Ethical Behaviourism. But why does it not seem plausible that Global Brain has moral status? As with the Blockhead twin, the best explanation for this seems to be that the Global Brain does not have a mind.

Here is a way to substantiate that this is the best explanation: functionalists will deny the intuitions about Global Brain. They will try to argue that if we only think about the case in the right way, we will think that Global Brain *does* have the same mental states as a human.

> The source of the intuition that [Global Brain] lacks mental states like ours seems to be the fact that it would be so *very* much bigger than we are. We cannot imagine "seeing" it as a cohesive parcel of matter.... A highly intelligent microbe-sized being moving through our flesh and blood brains might have the same problem. It would see a whole mass of widely spaced entities interacting with each other in a way that made no sense to it.[23]

But notice what happens if we submit to this resistance by functionalists, and come to believe that Global Brain has a mind just like any human does. In this case, our verdicts about moral status will tend to align with the verdicts of Ethical Behaviourism: we will likely think that Global Brain *should* be afforded the

[22] Braddon-Mitchell and Jackson, "The Philosophy of Mind and Cognition," 108.

[23] Braddon-Mitchell and Jackson, "The Philosophy of Mind and Cognition," 109.

same moral status that is afforded to humans, making the case described above a real moral dilemma. So, even if our intuitions about Global Brain having a mind are, in the end, incorrect, this case and the Blockhead case still spell trouble for Ethical Behaviourism as a form of minimalism: What these cases reveal is that the plausibility of Ethical Behaviourism seems to hang on the commitment to some sort view in the philosophy of mind, namely one that entails that the mental lives of two *roughly performative equivalent* beings are identical. Specifically, it seems as if Ethical Behaviourism stands or falls with the plausibility of such a view about the mind. So, Ethical Behaviorism seems to violate the second condition for a plausible version of minimalism.

We can bring this point home by considering views about the mind which do not entail that the mental lives of two *roughly performative equivalent* beings are identical. Consider, for example, teleo-functionalist views.[24] According to these views, in biological beings like us etiological biological functions are (at least partially) constitutive of having mental states, where the etiological function of something is what explains why it exists. On such views, we have mental states *because* of our biological history—*because* us having such states was selected by evolution. But note that on such a view, a being with a different causal history from ours could lack mental states, even if it is *roughly performative equivalent* to us. So, for example, if such a view was true, even a highly sophisticated android that is by all plausible performative standards roughly equivalent to a human—such as Data from *Star Trek*—would not have *any* mental states.[25]

Suppose now that we believe teleo-functionalism to be true, and so concede that Data does not have mental states. In this case, we would likely also think that the robot should not be granted the same moral status that we possess.[26] After all, despite all appearances, this robot would not experience pain, have preferences or interests, or anything else for which one needs a mental life. And in this case, it seems intuitively quite clear that none of the moral reasons

[24] For an introduction and overview of teleo-functionalist views, see Neander and Schulte, "Teleological Theories of Mental Content."

[25] See Hofmann, "Could Robots Be Phenomenally Conscious."

[26] Danaher considers the objection that the efficient cause of an entity might matter (Danaher, "Welcoming Robots to the Moral Circle," 2032-3). Specifically, he considers whether it might be relevant that humans (or animals) are produced by evolution, while robots are manufactured. He rejects this objection by pointing out that humans or animals could also be manufactured. However, this response would miss the central point of the objection pressed here, which is that an entity with a different kind of causal history lacks mental states. A human that is the product of selective breeding still stands in the relevant evolutionary chain to make it such that she has mental states, but a robot that is manufactured stands in no such chain.

that derive from moral status are present. However, what this means is that if a teleo-functional view about the mind is correct, Ethical Behaviourism must be false. And that means that to be plausible, Ethical Behaviourism must be combined with a commitment to the idea that teleo-functionalism is false, and something like the functionalism presented further above is true. Hence, rather than avoiding controversy in the philosophy of mind, Ethical Behaviourism is hostage to fortune with regards to such controversies.

What all of this shows is this: Ethical Behaviourism is plausible as a thesis about what entities should be granted moral status only to the extent that we combine it with certain assumptions in the philosophy of mind. That is, Ethical Behaviourism is only plausible to the extent that what it identifies as a sufficient condition on the basis of which we should grant moral status is also a sufficient condition for assuming that an entity having a mind. So, rather than it not mattering what goes on "on the inside," this actually matters a great deal for the plausibility of Ethical Behaviourism. It turns out, therefore, that Ethical Behaviourism is only really plausible if it is paired with very robust assumptions in the philosophy of mind. This means that Ethical Behaviourism does not really allow us to avoid the tricky questions in the philosophy of mind that plague Orthodoxy. At most, it assumes these questions away. So, Ethical Behaviourism cannot offer a plausible form of minimalism.

Furthermore, there is a more general lesson to take away from this discussion, namely this: as cases like Blockhead twins and Global Brain illustrate, our verdicts about what entities should or should not be afforded moral status, and about what entities have a mind tend to go hand in hand. When we judge an entity to not have a mind, we will tend to judge that it does not have moral status. A plausible view about moral status needs to be able to accommodate this. It is unclear how forms of minimalism can accommodate this, however, unless they are combined with assumptions about the mind that make their verdicts about moral status coincide in the relevant cases with what entities have a mind. If this is the case, though, these positions would not actually offer progress over Orthodoxy, at least when it comes to evading tricky issues in the philosophy of mind. But Ethical Behaviourism is not the only theory to which this applies. It holds for Ethical Relationism too, as I will argue next.

2. Ethical Relationism

In recent work, Mark Coeckelberg and David Gunkel have suggested what they call a "relational, other-oriented approach to moral standing."[27] The position

[27] This is, in fact, the subtitle of Coeckelbergh and Gunkel, "Facing Animals."

they develop (in co-authored and individual work) is very rich and complex, and it will not be possible to do justice to all of the details here.[28] What I will do is present what I take to be a plausible version of the view, focusing on the details that are central for the discussion here.

The core idea of Ethical Relationism is this:

> As we encounter and interact with others—whether they be other human persons, an animal, the natural environment, or a social robot—this other entity is first and foremost situated in relationship to us. Consequently, the question of social and moral status does not necessarily depend on what the other is in its essence but on how she/he/it … supervenes before us and how we decide, in "the face of the other" (to use Levinasian terminology), to respond.[29]

How should we understand this? First, the basic idea is that entities have moral status in virtue of the *relations* in which we stand to them. What kinds of relations? The general idea seems to be the following: in our interactions with certain other entities, these entities *present* themselves to us in certain ways. When such an entity presents itself to us as "having face," it has moral status.

This terminology of something "presenting itself to us" or "having face," needs a little unpacking. As I understand it, these notions are supposed to be graspable by the distinctive *phenomenology* of coming to regard an entity as worthy of moral consideration. This is the phenomenology of, e.g., experiencing discomfort at the entity being treated in certain ways, feeling affection for the entity, regarding the entity with respect, etc. that might arise in us due to our interactions with an entity.

To make the phenomenology vivid, think about the Robovie experiment that tested how children relate to social robots.[30] In this experiment, children ages nine to fifteen interacted with a humanoid robot for roughly fifteen minutes, including making small talk with the robot, playing a game, and the robot asking the child for a hug. At end of the session an adult comes and orders the robot to go into a closet. The robot protests vehemently, asking not to be put in the closet, which it claims is dark and lonely. Now put yourself in the position of a child having interacted with Robovie in this way. It seems plausible to assume that you will feel a certain bond with Robovie, and that you will feel distress at Robovie's reaction to being put in the closet. And, if you

[28] They have developed and defended the view in various places, e.g., Coeckelbergh, "Growing Moral Relations"; Coeckelbergh and Gunkel, "Facing Animals"; Gunkel, "Robot Rights." The view draws on the work of Emmanuel Levinas. See Levinas, "Totality and Infinity."

[29] Gunkel, "Robot Rights," 96.

[30] Kahn et al., "Robovie, You'll Have to Go into the Closet Now."

watch videos of the experiment you will likely find this reaction compelling—an emotional reaction that lingers on, even when you know that Robovie is nothing more than a puppet remotely controlled by a human in another room.[31] As I understand Coeckelbergh and Gunkel, these sorts of phenomenological responses are what determine moral status. Specifically, it is when these phenomenological responses make us experience the entity as "having face" that it should be afforded moral status—where I assume the phenomenology of experiencing something as "having face" is what we *paradigmatically* experience when we come to regard something as worthy of moral consideration. Hence, for Coeckelbergh and Gunkel the phenomenological responses of coming to regard something to be worthy of moral consideration in our interactions with it have *explanatory priority* to being of moral consideration. This is a reversal of the ordinary order of explanation: normally we would think that something being of moral consideration is what the relevant sorts of responses that constitute moral regard are *sensitive* to, while here being of moral consideration is determined or constructed by these responses.

Cockelbergh's and Gunkel's own presentation of the account raises many questions that I do not have room to investigate here. The way I read it, though, a helpful analogy for understanding this sort of view can be found in secondary quality accounts of colors.[32] On such accounts, something having a certain color crucially depends on our own responses to the thing. For example, on such accounts, things are red in virtue of appearing red to certain sorts of observers in certain sorts of conditions. Similarly, according to Ethical Relationism an entity has moral status in virtue of triggering the kinds of responses that constitute moral regard in certain sorts of observers in certain sorts of conditions. This might not be Coeckelbergh's and Gunkel's official account.[33] However, this is the reading I will presuppose in what follows to sharpen the discussion.

Ethical Relationism seems to be a form of minimalism. After all, we can *in principle* come to experience the distinctive phenomenology of moral regard toward an entity in our interactions with it without knowing whether it has a mind. The Robovie case illustrates this clearly: Robovie does not have a mind, but it presents itself in a way that gives rise to the right kinds of phenomenological responses. So, Robovie could, *in principle*, present as having moral status,

[31] For illustration, see, e.g., http://depts.washington.edu/hints/video8.shtml.

[32] For a secondary quality account of color, see, e.g., Johnston, "How to Speak of the Colors." Another helpful analogy might be the Strawsonian account of responsibility (see Strawson, "Freedom and Resentment") on which being responsible just *is* being the proper target for our reactive attitudes.

[33] Coeckelbergh alludes to this analogy (see "Growing Moral Relations," 207), though it seems that he does not fully embrace it.

even though it has no mind. And, Coeckelbergh and Gunkel explicitly argue that a primary motivation for Ethical Relationism is that it avoids the tricky questions in the philosophy of mind that plague Orthodoxy.[34] However, using the lessons from the discussion of Ethical Behaviourism, I will now argue that Ethical Relationism also does not provide a plausible version of minimalism, because Ethical Relationism is only plausible if it is combined with additional assumptions in the philosophy of mind. Again, therefore, the second condition that plausible versions of minimalism need to satisfy is violated.

The way to argue for this is, simply, by alluding to the cases we've discussed before, and by asking what best explains what is going on in these cases—which is something any plausible view about moral status should capture. Think about Blockhead twins or the Global Brain. What the discussion of Ethical Behaviorism revealed is that, in these cases, our intuitions about whether the entities in question should be afforded moral status follow our verdicts about whether they have a mind. If we meet an entity like Blockhead Twin or Global Brain, we might at first be inclined to attribute moral status to it, but we will retract this verdict upon realizing that the entity, plausibly, does not have mental states. In fact, it seems that Robovie also supports this. Our initial responses to Robovie upon interacting with it are, plausibly, the kinds of responses that characterize regarding a being as worthy of moral consideration. But, when someone shows us the remote control and explains that Robovie is, actually, just a puppet, we will judge that Robovie is not actually worthy of moral consideration.

How can Ethical Relationists accommodate these cases, and the way our intuitions about moral status here follow our verdicts about who has a mind? First, what the cases show is that the kinds of responses that characterize seeing a being as worthy of moral consideration *by themselves* cannot be sufficient for it to be the case that we *should* afford the being moral consideration. After all, it would be implausible to hold that Blockhead Twins, Global Brains, or Robovie do have moral status as long as we are mistaken about their mental lives, and they lose their moral status when we come to think that nothing is going on "on the inside."

However, Ethical Relationism as I understand it here already accommodates this. After all, on the suggested version of Ethical Relationism, an entity has moral status in virtue of triggering moral regard in *certain sorts of observers, in certain sorts of conditions*. So, with the right kind of view about what sorts of observers and what sorts of conditions are relevant here, Ethical Relationism might be able to explain how and why our verdicts about moral status and mental life tend to go together, and to imply the correct verdicts about the moral status of entities such as Blockhead twins, Global Brain, and Robovie.

[34] See, e.g., Coeckelbergh and Gunkel, "Facing Animals," 718–20.

What sorts of conditions and observers might an Ethical Relationist invoke at this point, though? Here, I will offer what seems to me the most plausible and straightforward response to this question.

Let me start with an observation about the responses that characterize moral regard. Irrespective of what we think about the relationship between these responses and something having moral status, it is, *prima facie*, very plausible that there is a close connection between many of the most central kinds of responses that constitute moral regard and our inclinations to attribute mental states. A core part of moral regard is to *feel with* and *feel for* the entity in question, to put ourselves in that entities' shoes, and so on. Hence, there seems to be a tight connection between our mindreading capacities and some of the most central responses that constitute moral regard. Specifically, it seems that many of the core responses that constitute moral regard are partially constituted by or are the result of our mindreading capacities.

However, our mindreading capacities sometimes *misfire*. For example, it is likely that this will happen when we interact with Blockhead twins. In fact, this is what happens with Robovie. And it is unsurprising that our mindreading capacities misfire in these sorts of situations. After all, while our mindreading capacities probably work very well in normal conditions where the relevant sorts of moral regard are good evidence for the presence of minds, the relevant sorts of cases are *not* normal: they are not situations for which our epistemic capacities for gaining knowledge of other minds have evolved.

These observations allow us to suggest a diagnosis as to what is going on in the cases in question: plausibly, the kind of moral regard we experience in reaction to Blockhead twins, Global brain, or Robovie is exactly the kind triggered or partially constituted by our mindreading capacities. But, when we realize that these relevant capacities *misfired* in this case, we retract our inclination to attribute moral status. If this explanation is plausible (and I cannot see a better explanation), this, in turn, allows us to spell out further the *conditions* in which something has to appear to certain observers as having moral status in a way that allows us to account for the intuitions in question. This suggests that part of the relevant conditions for granting moral status is just this: that our epistemic capacities for gaining knowledge of other minds do *not* misfire.

If we include this sort of condition in Ethical Relationism, we can explain why our intuitions about who has a mind and our intuitions about who should be afforded moral status tend to go together: those responses of moral regard that result from our mindreading capacities are amongst our most central ones, and their sensitivity to these capacities explains the connection. We can offer this explanation only, however, by making assumptions about when and why our mindreading capacities misfire—i.e., by making assumptions about when entities have a mind and when we have knowledge of their minds. Hence, in including this sort of condition, we've now uncovered that to deal with the challenge on the table, Ethical Relationism has to violate the second condition for a plausible form of minimalism: to make plausible verdicts about certain sorts of cases, Ethical Relationism must be combined with certain robust assumptions in the philosophy of mind. But, this puts the account squarely back into thorny philosophical issues in the philosophy of mind regarding these capacities. Hence, rather than avoiding such problems, we are back to facing these problems.

At the end of section one it was observed that a plausible view about when we should grant entities moral status needs to be able to accommodate the fact that our intuitions about moral status and about who has a mind tend go together. It was suggested that it is unclear how forms of minimalism can accommodate this unless they make assumptions about the mind. The discussion of Ethical Relationism in this section substantiates this suspicion. Before I go on to draw some general lessons from the discussion, though, let me respond to two possible worries about my argument against Ethical Relationism.

First, we should address whether the argumentation against Ethical Relationism is question-begging. Does the argument here implicitly *presuppose* that only entities with mental states have moral status?[35] Two features of my argumentation might raise this worry. First, above I have tied our mindreading capacities to moral regard. The way I did this might give the impression that the argumentation already presupposes that *proper* moral regard is only triggered by entities with a mind. Second, the discussion did not even consider any "mind-less" entities that might conceivably have moral status, such as plants, forests, mountains, or the planet itself, that explicitly figure in some of Coeckelbergh's and Gunkel's work.[36] This might give the impression that the cases I've focused on smuggle in, again, the assumption that only entities with mental states have moral status into the argumentation.

However, my argumentation does not rely on such an illicit implicit assumption. First, the challenge that I raised for Ethical Relationism does not itself rest on such an assumption. What the challenge does is ask Ethical Relationism to explain why our intuitions about what entities have a mind and what entities have moral status tend to go together in the way suggested by cases like Blockhead twins, Global brain, or Robovie. This challenge would stand, even if there *were* clear cases in which our verdicts about moral status are not influenced by our judgements about minds. So, the challenge itself

[35] Thanks to an anonymous reviewer for *JESP* for suggesting that I address this worry.

[36] See, e.g., Coeckelbergh, "What Do We Mean by a Relational Ethics?" and "Environmental Skill"; Gunkel, "Robot Rights."

is compatible with the possibility that mindless things have moral status: to raise the challenge, we do not have to implicitly assume that only entities with mental states have moral status.

Second, the response to the challenge that I offered on behalf of Ethical Relationists is also not tied to a question-begging assumption. I suggested that Ethical Relationists can deal with the challenge by imposing a condition on the kinds of responses that matter for determining moral status: that the reactions that matter are formed by observers in conditions in which our mindreading capacities do not misfire. However, this condition is *compatible* with people reacting with moral regard to at least certain things that do not have a mind. Whether there are such cases depends on what kinds of moral regard there are, and what they are triggered by. Regarding this, the explanation only takes the following stance: First, there is a close connection between many responses that constitute moral regard and our inclinations to attribute mental states. Second, these responses are *core* to what constitutes moral regard. Third, focus on these responses is sufficient to explain what is going on in the cases relevant for the discussed challenge. All of these claims are very plausible, but none of them rules out the possibility of other kinds of moral regard triggered by mindless entities. Specifically, the explanation does not need the assumption that *all* types of moral regard are tied to our attributions of mental states. It only needs the assumption that this is true for *some* types, and that appeal to those is sufficient to deal with the challenge. The explanation is compatible, for example, with there being a distinctive type of *moral* regard we experience in relation with a magnificent oak. Note, though, that even if there *are* such responses, the condition I suggested for dealing with the challenge is not problematic. These responses are not going to be influenced by our mindreading capacities either way, so requiring that moral status be determined by the responses of people in conditions in which these capacities do not misfire is not going to change what moral status is attributed on the basis of these responses. Hence, the argument against Ethical Relationism does not rely on question-begging assumptions about the connection between minds and moral regard.

Before I move on to the second worry, let me end the discussion here by flagging that even if there *are* types of moral regard properly triggered by mindless entities, they are not going to help *minimalist* Ethical Relationism. First, these responses do not change the fact that the best way for Ethical Relationists to deal with the challenge is to add the condition I suggested. But, adding this condition to Ethical Relationism itself defeats the minimalist aspirations of the view. Second, even if we *only* focus on the case of artificial agents, the relevant types of responses are not going to be of help. This is so because the kind of moral regard that is triggered by such agents is, very plausibly, exactly the type of moral regard that is partially constituted by, or the result of, our mindreading capacities. This is suggested, for example, by reflection on the cases discussed in this paper, and on what best explains how our intuitions about these cases are influenced by our verdicts about the presence of minds.

Let me now discuss a second potential worry. This is the worry about whether I've discussed the best version of Ethical Relationism here.[37] The view I've discussed draws on work by Cockelbergh and Gunkel, but their actual view is more complex than the one I suggest. Most relevantly, Coeckelbergh's and Gunkel's work highlights certain sorts of factors that influence moral status attributions, but which have not been considered as part of the account I suggest. The main factors they focus on are the language we use to talk about entities, our relationships with those entities, and social norms concerning those entities. Coeckelbergh and Gunkel note, for example, that it matters a lot for how one feels about an animal whether one tends to think of it "as an animal," or whether one has given it a name. Similarly, social acceptance of eating meat can influence how we feel for farm animals in the way we engage with them. This raises the question of whether these additional factors can help Ethical Relationists with the challenge at hand.

No. First, let us be clear how exactly these observations should feed into the account. Coeckelbergh's and Gunkel's observations are about our actual responses of moral regard, and such factors do indeed, very plausibly, influence our actual responses of moral regard in our interactions with entities. However, what should matter on the Ethical Relationist account are not our *actual* responses of moral regard in response to our interactions with an entity: this would be very implausibly relativistic *and* highly revisionary.[38] It would imply, implausibly, for example, that if slave owners harden themselves to the plight of the enslaved, slaves do not have moral status, at least relative to slave owners. Rather, it must be the responses of people who interact with the entity under certain *ideal* conditions that determine moral status. I assume that this should mean at least that one reacts to the entity as a result of *fair* engagement:

[37] Thanks to an anonymous reviewer for *JESP* for raising this objection.

[38] In fact, it is important to be very clear here that the only thing that matters is the influence these factors have on *moral* regard, not on other emotional reactions. After all, these factors influence a variety of ways we feel about entities. For example, the relationship I have with my wedding ring makes it such that I care a lot about it—much more than I care about qualitatively identical other rings. However, the way my relationship makes me feel about my wedding ring should not count for moral status. This would be *highly* revisionary: my wedding ring matters *to me*, but it does not matter for its own sake. Of course, this raises another issue for Ethical Relationism that I have not talked about, which is how we individuate specifically *moral* regard. But this is an issue that goes beyond this paper. Here we should rely on the intuitive difference between different types of responses.

engagement that is honest and freed from certain pre- and mis-conceptions that unduly interfere with the responses that constitute moral regard. I think that this sort of emphasis on fair engagement is a plausible reading of how Coeckelbergh and Gunkel themselves would understand their suggestions.

The factors they highlight as influencing moral status attributions can then be understood as feeding into the conditions in which moral regard has to be formed to determine moral status. This is plausible: if our naming practices interfere with the moral regard we feel for animals, for example, then only responses should count that are not unduly influenced by such practices. Of course, figuring out how exactly to spell out undue interference by these sorts of factors in a way that is non-circular is going to be quite a challenge for Ethical Relationists, but this is not an issue that should concern us here.

What is important here is only that we can now respond the second worry raised above, as we can now see that these additional factors are not going to help with the challenge. For cases like Blockhead Twins, Global Brain, or Robovie, it does not seem plausible that social norms, relationships, or the language we use interfere unduly with our moral regard. In fact, if there is *anything* that interferes in these cases with our responses, it is our overreactive tendency to attribute minds to mind-less entities. A good example for this, I think, is the case of the social humanoid robot Sophia.[39] Sophia *does* have a name and if we consider the way people tend to interact with it, it seems quite plausible that Sophia triggers in them at least some of the responses we associate with moral regard. However, when we understand the actual workings of the robot and that it is little more than a "chat-bot with a face," our inclination to attribute moral status immediately disappears.[40] It would be implausible, though, to attribute this retraction to an interference of the language we use or the relationships we have with the robot. Rather, what explains the retraction best, quite simply, is that our earlier attribution was based on our now corrected misconception that Sophia has a mental life.[41] One might suggest that we are unduly influenced here by a preconception that Sophia would need to have a mind to have moral status, but now *this* would be question-begging without further argument that it is an *illicit* preconception.

[39] See Hanson Robotics, "Sophia."

[40] See, e.g,. Gershgorn, "Inside the Mechanical Brain of the World's First Robot Citizen"; Ghosh, "Facebook's AI Boss Described Sophia the Robot as 'Complete B------t.'"

[41] This impression is further suggested by, e.g., the way David Hanson (the founder of Hanson Robotics which created Sophia) tends to (mis-)represent Sophia in a way that suggests the existence of a mental life (see Vincent, "Sophia the Robot's Co-Creator Says the Bot May Not Be True AI"; Hanson Robotics marketing material (see Hanson Robotics, "Sophia") is also very suggestive in this regard).

In any case, it does not seem as if the further conditions that we might derive from Coeckelbergh and Gunkel would help Ethical Relationism deal with the challenge that I posed. Rather, it still seems like the best way to accommodate the way our intuitions about moral status tend to follow our verdicts about who has a mind is to build the condition I suggested into Ethical Relationism. But this condition saddles Ethical Relationism with commitments in the philosophy of mind.

3. Conclusion: General Lessons about Minimalism

The discussion of this paper has yielded two important lessons. First, that our intuitions about what entities should be afforded moral status tend to go hand in hand with our intuitions about whether an entity has a mind. In a sense, this is unsurprising, given how strong Orthodoxy's following is. It is a significant result in the face of the promises of minimalism, however, because with this result on the table, one can formulate a challenge to any form of minimalism: it needs to be able to accommodate how our intuitions go together in this way, and to explain why those intuitions tend to go together.

The second important lesson is then about the prospects of minimalism meeting this challenge. Here the discussion provides evidence that views about moral status that do not explicitly subscribe to a version of Orthodoxy can accommodate these intuitions about moral status only by themselves being combined with controversial assumptions in the philosophy of mind.[42] And this means that such views violate the second condition for plausible versions of minimalism. Rather than avoiding controversies in the philosophy of mind, such views put themselves squarely within those controversies.

In this paper we have only looked at views that try to give sufficient conditions for granting moral status in terms of two sorts of alternatives to mental states: outward behavior and moral responses. While this set of alternatives is restricted, these are also the *prima facie* most promising alternatives, as they tie

[42] I am proceeding on the assumption, of course, that we take the intuitions to be legitimate evidence for moral status. Another option for the minimalist to respond to the challenge is to debunk those intuitions, e.g., by suggesting they are based in some sort of bias. Obviously, though, minimalists would have to give us good reasons to think that these intuitions need to be debunked—our moral intuitions are amongst our best guides to the correct answers to moral questions and we should not just give them up just because they conflict with minimalism. I don't see any good reasons to give these particular intuitions up (in particular because there are many ways to explain why there might be moral requirements to interact in certain ways with entities without a mind, even if we concede they lack moral status. Things that do not matter in themselves might still matter because people care about them, after all).

moral status to something it *is* paradigmatically related to: both are *normally* very good evidence for moral status.[43] If even these kinds of views do not succeed—which are probably our best bet when it comes to developing forms of minimalism that are not radically revisionary—this is strong evidence that when thinking about moral status, there is no way around the issues in the philosophy of mind that plague Orthodoxy (at least not without radical revision). Hence, these sorts of issues do not seem to provide a good motivation to move away from Orthodoxy.[44]

Frankfurt School of Finance and Management
s.koehler@fs.de

References

Block, Ned. "Troubles with Functionalism." *Minnesota Studies in the Philosophy of Science* 9 (1978): 261–325.

Block, Ned. "Psychologism and Behaviourism." *The Philosophical Review* 90, no. 1 (January 1981): 5–43.

Braddon-Mitchell, David, and Frank Jackson. *The Philosophy of Mind and Cognition*. 2nd ed. Oxford: Blackwell, 2007.

Coeckelbergh, Mark. *Growing Moral Relations. Critique of Moral Status Ascriptions*. New York: Palgrave MacMillan, 2012.

Coeckelbergh, Mark. *Environmental Skill. Motivation, Knowledge and the Possibility of a Non-Romantic Environmental Ethics*. New York: Routledge, 2015.

Coeckelbergh, Mark. "What Do We Mean by a Relational Ethics? Growing a Relational Approach to the Moral Standing of Plants, Robots and Other Non-Humans." In *Plant Ethics. Concepts and Applications,* edited by Angela Kallhoff, Marcello Di Paola, and Maria Schörgenhumer, 98–109. New York: Routledge, 2018

Coeckelbergh, Mark, and David Gunkel. "Facing Animals: A Relational,

[43] Of course, the discussion here indicates that the primary reason why these are *normally* good evidence for moral status just is because they are *normally* good evidence for a mental life. The problem with artificial agents, however, is that it is not clear whether what *normally* is good evidence for a mental life is also evidence for a mental life in conditions that are not normal.

[44] For helpful feedback I would like to thank Johannes Himmelreich, Leo Menges, Christine Tiefensee, as well as two anonymous referees for this journal. I would also like to thank the audience of the FS Philosophy Forum, where the idea for this paper was born. Research for this paper was conducted while I was a principal investigator of the project group *Regulatory Theories of* AI of the *Centre Responsible Digitality* (ZEVEDI).

Other-Oriented Approach to Moral Standing." *Journal of Agricultural and Environmental Ethics* 27, no. 5 (October 2014): 715–33.

Danaher, John. "Welcoming Robots into the Moral Circle: A Defence of Ethical Behavourism." *Science and Engineering Ethics* 26, no. 4 (August 2020): 2023–49.

Floridi, Luciano. "Information Ethics: On the Philosophical Foundation of Computer Ethics." *Ethics and Information Technology* 1, no 1 (March 1999): 33–52.

Gershgorn, Dave. "Inside the Mechanical Brain of the World's First Robot Citizen." Quartz. November 12, 2017. https://qz.com/1121547/how-smart-is-the-first-robot-citizen/.

Ghosh, Shona. "Facebook's AI Boss Described Sophia the Robot as 'Complete B------t' and 'Wizard-of-Oz AI.'" Business Insider. January 6, 2018. http://www.businessinsider.com/facebook-ai-yann-lecun-sophia-robot-bullshit-2018-1.

Gunkel, David. *Robot Rights*. Cambridge MA: MIT Press, 2018

Gunkel, David. "The Other Question: Can and Should Robots Have Rights?" *Ethics and Information Technology* 20, no. 2 (June 2018): 87–99.

Hanson Robotics. "Sophia." https://www.hansonrobotics.com/sophia/.

Hofmann, Frank. "Could Robots Be Phenomenally Conscious." *Phenomenology and the Cognitive Sciences* 17, no. 3 (July 2017): 579–90.

Jaworska, Agnieszka, and Julie Tannenbaum. "The Grounds of Moral Status." *The Stanford Encyclopedia of Philosophy* (Spring 2021). https://plato.stanford.edu/archives/spr2021/entries/grounds-moral-status/.

Johnston, Mark. "How to Speak of the Colors." *Philosophical Studies* 68, no. 3 (December 1992): 221–63.

Kahn, Peter H., Takayuki Kanda, Hiroshi Ishiguro, Nathan G. Freier, Rachel L. Severson, Brian T. Gill, Jolina H. Ruckert, and Solace Shen. "'Robovie, You'll Have to Go into the Closet Now': Children's Social and Moral Relationships With a Humanoid Robot." *Developmental Psychology* 48, no. 2 (March 2012): 303–14.

Levin, Janet. "Functionalism." *The Stanford Encyclopedia of Philosophy* (Winter 2021). https://plato.stanford.edu/archives/win2021/entries/functionalism/.

Levinas, Emmanuel. *Totality and Infinity: An Essay on Exteriority*. Pittsburgh: Duquesne University Press, 1969.

Mosakas, Kestutis. "On the Moral Status of Social Robots: Considering the Consciousness Criterion." *AI and Society* 36, no. 2 (June 2021): 429–43.

Neander, Karen and Peter Schulte. "Teleological Theories of Mental Content." *The Stanford Encyclopedia of Philosophy* (Spring 2021). https://plato.stanford.edu/archives/spr2021/entries/content-teleological/.

Nyholm, Sven, and L E. Frank. "From Sex Robots to Love Robots: Is Mutual Love with a Robot Possible?" In *Robot Sex: Social and Ethical Implications,* edited by John Danaher and Neil McArthur, 219-244. Cambridge MA: MIT Press, 2017.

Smids, Jilles. "Danaher's Ethical Behaviourism: An Adequate Guide to Assessing the Moral Status of a Robot?" *Science and Engineering Ethics* 26, no. 5 (October 2020): 2849–66.

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

Tavani, Herman T. "Can Social Robots Qualify for Moral Consideration? Reframing the Question about Robot Rights." *Information* 9, no. 4 (April 2018): 73.

Vincent, James. "Sophia the Robot's Co-Creator Says the Bot May Not Be True AI, but It Is a Work of Art." *The Verge*. (November 2017). https://www.theverge.com/2017/11/10/16617092/sophia-the-robot-citizen-ai-hanson-robotics-ben-goertzel.

Journal of Ethics and Social Philosophy
vol. 24, no. 1 · february 2023

https://doi.org/10.26556/jesp.v24i1.1972

ACTUAL GUIDANCE IS ENOUGH

Stefan Fischer

In a recent paper, Nate Sharadin and Rob van Someren Greve cast into doubt the seemingly self-evident idea that deontic evaluation—the application of deontic concepts to action—is capable of guiding action.[1] The authors' skeptical endeavor is provocative, given that many philosophers would instantly grant that applying the concept wrong to an action can guide the agent in question with respect to whether or not to perform it.[2] In this critical note, I argue that Sharadin and van Someren Greve's skeptical endeavor is unwarranted.

The authors' starting point is an argument schema, instances of which appear throughout the literature on deontic concepts.[3] The point of arguments of this kind is to learn something interesting and substantive about deontic concepts by asking what these concepts would have to be like in order to serve their *function*: to guide action.[4]

Schematic Argument:

p1. The function of deontic evaluation is to guide action (Guidance Function).

p2. If deontic concepts have feature *X*, then deontic evaluation would not be able to guide action (Disabling Condition).

c. So, it is not the case that deontic concepts have feature *X* (Substantive Result).

[1] Sharadin and van Someren Greve, "Is Deontic Evaluation Capable of What It Is For?" Parenthetical citations refer to pages in this article and, for reasons of brevity, I will often refer to "the authors."

[2] Small caps denote concepts.

[3] They mention Smith, "Moral Realism, Moral Conflict, and Compound Acts"; Carlson, "Deliberation, Foreknowledge, and Morality as a Guide to Action"; Wiland, "How Indirect Can Indirect Utilitarianism Be?"; Bykvist, "Violations of Normative Invariance"; and Copp, "'Ought' Implies 'Can' and the Derivation of the Principle of Alternate Possibilities."

[4] The authors explicitly accept this function (205).

Throughout their paper, the authors focus on Holly Smith's instance of the schematic argument, as will I:

> *Smith's Argument*:
> 1. *Guidance Function*: The function of deontic evaluation is to guide action.
> 2. *Disabling Condition*: If the concepts RIGHT and WRONG ever apply to the same action, then deontic evaluation in terms of RIGHT and WRONG would not be able to function.[5]
> 3. *Substantive Result*: So, it is not the case that the concepts RIGHT and WRONG ever apply to the same [action].[6]

Smith's argument serves to illustrate the theoretical appeal of arguments of this kind: it supposedly generates an interesting and substantive insight—that it is a mistake to apply RIGHT and WRONG to the same action—on the basis of the idea that the point of deontic evaluation is to guide action.

However, Sharadin and van Someren Greve's main claim is that we cannot really learn anything interesting and substantive from arguments of this kind

[5] An anonymous referee points out that P2 is implausible. If the concepts RIGHT and WRONG applied to an action in one case, this would only show that deontic evaluation is *not always* able to function, but not that it *never* is. In Smith's defense, I do not think this worry is justified. Actions are wrong just when they are not-to-be-done; they are right when they are not not-to-be-done. If it were possible that RIGHT and WRONG apply to one and the same action, then the thought "ϕ is wrong (not-to-be-done)" would not preclude the thought "ϕ is right (not not-to-be-done)." And then the thought that ϕ is wrong *alone* would be practically worthless (because thinking that an action is not-to-be-done only guides action, it seems to me, if it precludes that the action is not not-to-be-done). Therefore, Smith's premise strikes me as plausible. But, the anonymous referee maintains, there are deontic evaluations that *still* fulfill their function, even if RIGHT and WRONG sometimes apply to the same action—namely, deontic evaluations of the form "ϕ is right and not wrong." These evaluations, the referee insists, could still provide guidance. I agree. Deontic evaluations of this and *only this* form would still be able to function. But, crucially—and here I stomp my foot—evaluations of the form "ϕ is wrong/right" would *not* be able to function. So, my suggestion is this: the most charitable reading of P2 states that deontic evaluations in terms of *only one* of the two concepts would not be able to function. Thus understood, P2 is plausible. Most importantly, however, I take it that Sharadin and van Someren Greve (could easily) agree with my reading of Smith. I see no problem with respect to the issues under dispute between the authors and myself. Either the anonymous referee is right and all three of us make the "mistake" of ignoring a problem in Smith's argument, or the referee is wrong and none of us make a mistake. But even if the authors and myself make a mistake here, all that follows is that we need to find another instance of the schematic argument to get our discussion going. (And, as Sharadin and van Someren Greve show, there are many versions to be found.) So, in the context of the current paper, not much depends on this dispute about Smith's second premise.

[6] I slightly changed Smith's formulation of P2 and the conclusion. Nothing of substance is lost.

because they have a hidden premise that, as it turns out, cannot be defended in a satisfying way. We have already seen it:

> *Capable:* Deontic evaluation is capable of guiding action.

The authors are right: the schematic argument does presuppose that deontic evaluation *can*, in principle, guide action. (Capable is "hidden" in Guidance Function: calling *F* the function of *A* implies that there are at least some instances of *A* that "fulfill" *F*.) If deontic evaluation was, in principle, incapable of guiding action, the argument clearly would not work. And so it turns out that the plausibility of the schematic argument—and whether or not it can lead to interesting, substantive results—depends on the truth of Capable. Consequently, Sharadin and van Someren Greve investigate how Capable might be defended. After discussing several possible defenses, the authors ultimately conclude that none of them are satisfying (227).

In this discussion note, my focus lies on two of the discussed defenses. The first one is the platitude argument.[7] It states that Capable is a platitude that need not be defended. The second is the pragmatist argument. It states that deontic concepts are *defined* as the concepts that are capable of guiding action, in which case there is no room to doubt the truth of Capable.[8] Note that these two arguments are not independent of each other. The pragmatist strategy *explains* why Capable is a platitude: if being capable of guiding action were a definitional feature of deontic concepts, it would be platitudinously true that they can do so. Nevertheless, I am going to treat both arguments separately. Even though, personally, I am most sympathetic to the pragmatist argument, I think the platitude argument can be defended against the authors' worries without it.

So, in the following, I am going to argue that the authors' criticism of the platitude and the pragmatist argument is unfounded. If I am right, the attempt to kick off their eponymous controversy about deontic evaluation fails. My two main worries with the authors' position concerns their (I argue) implausibly broad understanding of what it means to be actually guided by a deontic evaluation and a certain unclarity about what is to count as an "interesting and substantive" conclusion produced by the schematic argument. I ultimately provide a positive outlook: we may remain optimistic about the prospects of learning something interesting and substantive about deontic concepts by investigating their function—providing actual guidance, properly understood.

[7] See Sharadin and van Someren Greve, "Is Deontic Evaluation Capable of What It Is For?" secs 2.1, 2.2.

[8] See Sharadin and van Someren Greve, "Is Deontic Evaluation Capable of What It Is For?" sec. 3.

1. The Platitude Argument

Before we ask ourselves whether Capable is a platitude, let us have a closer look at what it says. The authors present two interpretations that are based on two different understandings of action guidance—namely "actual guidance" (AG) and "correct guidance" (CG).

> *Capable*$_{AG}$: A concept is capable of *actually guiding* when an agent's thinking about something in terms of that concept helps settle, for the agent, what they shall do … and (at least in part) motivates the agent toward doing that thing. (211)
>
> *Capable*$_{CG}$: A concept is capable of *correctly guiding* when an agent's thinking about something in terms of that concept helps *correctly settle*, for the agent, what they shall do … and (at least in part) motivates the agent toward doing that thing. (211)

There is a further ambiguity in Capable that needs to be addressed. It concerns the involved quantifiers. Capable might be understood as saying that, *for every* action alternative ϕ, a deontic evaluation of ϕ helps settle the agent what to do. But this would be too strong for a simple reason: morality is silent on many practical issues.[9] If I ask myself whether I should watch a movie or play a video game tonight, a deontic evaluation of both action alternatives (as, say, permitted) does not help me settle the issue. Therefore, Capable is best understood as stating that *there are some* action alternatives such that deontic evaluations of them help the agents in question settle what to do.

Having explained this, it is quite clear that Capable$_{AG}$ is a platitude.[10] As soon as we have a single example for an agent who applies the concept wrong to an action such that this helps them settle whether to perform the action, the agent has been actually guided by their deontic evaluation. Then, Capable$_{AG}$ turns out true. It will be very easy to find one such example.

However, note that Capable$_{CG}$ is not obviously true. For, whether an agent is correctly guided by applying the concept wrong to an action trivially depends on whether or not the action is, in fact, wrong. (One can falsely believe an action is wrong.) In other words, while the platitude argument can help us defend Capable$_{AG}$, it cannot help us defend Capable$_{CG}$.

The next step is crucial. The authors now claim that we *must* understand *Capable* in terms of correct guidance. Why? Because, they say, Smith's

[9] Thanks to an anonymous referee for pointing this out.

[10] The authors agree; see Sharadin and van Someren Greve, "Is Deontic Evaluation Capable of What It Is For?" 211.

argument (as well as the schematic argument) has two features, and it could only have these features if we understood action guidance as correct guidance: Smith's argument is (1) *sound* and (2) *interesting and substantive*. So, in short, the authors' case against the platitude argument is that the interpretation of *Capable* it defends is simply *not* the interpretation we need to get Smith's argument to work. Thus, the platitude argument fails.

I will proceed to discuss the authors' following two claims in turn:

1. Capable$_{\text{AG}}$ renders Smith's argument unsound.
2. Capable$_{\text{AG}}$ renders Smith's argument uninteresting and non-substantive.

My point of attack is this: I do not think that Smith's argument (and, thus, the schematic argument) could only be sound, interesting, and substantive if we understood Capable in terms of correct guidance. I will outline a natural and coherent sense of "actual guidance"—the mundane sense—that, while falling well short of correct guidance, nevertheless renders Smith's argument sound, interesting, and substantive. Moreover, I will suggest that the platitude argument can be used to defend Capable in my "mundane" understanding of actual guidance.

1.1. Actual Guidance Does Not Render Smith's Argument Unsound

Here is Smith's argument in the *Actual Guidance* interpretation.

Smith's Argument$_{\text{AG}}$:

1. The function of deontic evaluation is to actually guide action.
2. *Capable*$_{\text{AG}}$: Deontic evaluation is capable of actually guiding action.
3. If the concepts RIGHT and WRONG ever apply to the same action, then deontic evaluation in terms of right and wrong would not be able to function.
4. So, it is not the case that the concepts RIGHT and WRONG ever apply to the same action.

The authors suggest that this argument is unsound because it has a false premise.[11] Since they grant the truth of the first two premises (207–8), the only available point of attack is the third:

[11] I am not sure whether they additionally claim that the argument is invalid in the sense that true premises would not guarantee a true conclusion. The paper oscillates between the terms "invalid," "unsound," and "far less plausible." There are two points in the paper where the authors explicitly say that the argument is invalid. They claim that correct guidance is needed for the argument schema "to be valid" (215) then claim that the argument schema is invalid without Capable$_{\text{CG}}$ (217). However, I do not think this is correct. Consider a formal version of the argument:

> [Smith's argument in terms of actual guidance] lacks any . . . plausibility if the point of deontic evaluation is simply to *actually* guide action. . . . Why should we think that the fact that the concepts WRONG and RIGHT apply to the same action would somehow interfere with this capacity? At most what is true is that *thinking* these two concepts apply to the same action would make it more difficult in practice—though not at all *impossible* in practice—for deontic evaluation to actually guide action. . . . Notice that an agent's ability to *actually* be guided by deontic evaluation in her action is not even affected by the fact that these concepts apply to the same action unless, as a matter of fact, she thinks both concepts apply. (213)

I have two remarks about this passage. First, in our current context it seems unhelpful to differentiate between the *fact* that a deontic concept applies to an action and the *thought* that it applies. Deontic evaluation is the application of deontic concepts to actions (211)—which, of course, takes place in practical *thought*. Therefore, the question we are interested in at this point in the dialectic is whether or not *thinking* that a deontic concept applies to an action actually helps the agent in question settle whether to perform it.

Second, and more importantly, I am a bit startled that the authors do not seem to worry that a classification of an action as both right *and* wrong might cause actual guidance issues. Suppose Fatima ponders whether she should ϕ and concludes that ϕ-ing is right *and* wrong. To my mind, it seems rather obvious that this deontic evaluation did not "help settle" Fatima's question of whether or not to ϕ. But the authors make it very clear that deontic evaluations of this kind do guide action:

> What will happen is that the agent will find herself being actually guided towards the action (since it is right) and actually guided away from it (since it is wrong), with the result that, in the end, she will either perform it or not. Whatever she does, as a matter of fact *she will have been actually guided.* (214)

This, it seems to me, is an implausibly broad understanding of "being actually guided." I believe that deontic evaluations of this kind do not "help settle" the

1. The function of *A* is *F*.
2. *A* is capable of *F*.
3. If *A* had property *X*, then it would not be able to function.
4. So, *A* does not have property *X*.

This seems to be a valid argument (even if *F* means "providing actual guidance"). However, in the following, I neglect validity and only focus on the charge that the argument is unsound in virtue of having a false premise.

issue in a way that deserves this label. In contrast, they are entirely practically unhelpful. Let me explain.

In practical deliberations, we try to figure out what to do. A deontic evaluation is part of this process; so, the practical issue at the heart of this evaluation is figuring out whether or not to ϕ. And, intuitively, if a deontic evaluation of ϕ does not bring an agent any "closer " to deciding whether or not to ϕ, we would not say that the evaluation has helped her figure out what to do.[12]

Now, it is clear that we can figure out what to do *correctly* and *incorrectly*. Let us say that we do so correctly just when the result of our deontic evaluation of ϕ represents the *objective* deontic properties of ϕ, that is, the deontic properties ϕ actually has. And let us say that we do so incorrectly if it does not. We may distinguish these objective deontic properties from the *subjective* deontic properties, that is, the ones *we believe* ϕ has. In other words, subjective deontic properties are the deontic properties as they present themselves from our own epistemic outlook. Since we can falsely believe an action to have a property, subjective and objective deontic properties can come apart. Then, as I am going to say, our *subjective deontic landscape* and the *objective deontic landscape* come apart. However, importantly, only the objective deontic landscape determines the correctness of deontic evaluations. I take it that Sharadin and van Someren Greve accept all of this.

This distinction between objective and subjective deontic properties allows us to make a further distinction between two ways in which a deontic evaluation can "help settle" an agent whether or not to ϕ.

1. *Correct*: A deontic evaluation helps an agent to *correctly* settle what to do just when it brings her subjective deontic landscape closer to the objective one and (at least in part) motivates the agent toward performing the action under consideration.[13]
2. *Mundane*: A deontic evaluation helps an agent to *mundanely* settle what to do just when it changes the agent's subjective deontic landscape and, thereby, changes the agent's motivation toward the action under consideration.[14]

These, I take it, are two natural and coherent senses of "helping to settle" an agent what to do. While it seems clear that the first sense is much more ambitious and, typically, the one we strive for, it is similarly clear that something could help us

[12] I appeal to the reader's intuition here, but my ultimate point will not rely on an intuition.

[13] This sense of "help settle" corresponds to what the authors' call *Correct Capable* (216).

[14] Note that, according to Mundane, it is possible that an agent's deontic evaluation helps her settle a practical question incorrectly because, here, only the agent's own epistemic outlook is relevant, irrespective of the objective deontic properties.

settle what to do only in the second, more mundane, sense. For example, my friend's comment that the bus fares are outrageously expensive could help me settle to dodge the fare by making me believe that it might not be wrong to do so once a month (and by changing my motivation respectively). Then, my friend's comment would help me settle what to do in the mundane sense.

With this distinction in mind, let us return to Fatima. We have already seen that it could not be a platitude that deontic evaluation helps her to correctly settle what to do; whether or not her deontic evaluation tracks the objective deontic properties of ϕ-ing is entirely open. Next, does it help her to *mundanely* settle what to do? The answer, I think, must be no. Other things equal, her deontic evaluation does not change her subjective deontic landscape at all. According to her evaluation, ϕ-ing is wrong (not-to-be- done) *and* right (not not-to-be-done). This evaluation is entirely unhelpful. Other things equal, Fatima is not any closer to figuring out whether or not to ϕ, at least not (solely) in virtue of her deontic evaluation. Therefore, we may conclude, Fatima's deontic evaluation does not help her settle whether or not to ϕ. And this means, crucially, that she has *not* been actually guided by it, at least not in the mundane sense of "actual guidance."[15]

But Sharadin and van Someren Greve argue that Fatima's deontic evaluation *does* provide her with actual guidance. This must mean that their understanding of "help settle" differs from our mundane sense. They must have an even less ambitious, broader sense in mind. What sense is that? As we saw in the above quote, it is a sense according to which Fatima "will have been actually guided"

[15] An anonymous referee points out that the thought "ϕ is right and wrong" might help an agent mundanely settle what to do in a special case—namely, when she previously thought that ϕ was *only* right. In this case, she would come to think that the action is also wrong, which would change her subjective deontic landscape and, thereby, her motivation toward ϕ-ing. Two quick replies. First, this is why I use the "other things equal" qualification in the main text. The "other things equal" is meant to exclude scenarios like the one envisioned by the anonymous referee. I agree with the referee's objection insofar as in the scenario(s) he envisions, the deontic evaluation helps the agent to mundanely settle what to do; but it would not do so "on its own." It would only help the agent to mundanely settle what to do because she *previously* had evaluated the situation differently. The deontic evaluation alone, without the previous evaluation, remains entirely unhelpful. Second, and more importantly, call to mind the relevant disagreement between the authors and myself: they purport that Fatima's deontic evaluation *always* provides actual guidance; I claim that it does not. All I need to defend my claim is my mundane sense of actual guidance and *one* scenario in which Fatima is not guided in the mundane sense by her thought that ϕ is right and wrong. And there clearly are scenarios of this kind. So, as far as I can see, the anonymous referee's imagined case, in which this deontic evaluation helps the agent mundanely settle what to do *in the context of* another, previously made deontic evaluation, does not affect my argument against Sharadin and van Someren Greve.

by her deontic evaluation because it steers her toward ϕ-ing and it steers her away from it.

I find this claim highly unintuitive; the respective sense of "help settle" would have to be spelled out.[16] But my main point does not rely on an intuition. What is important is that we *can* (naturally and coherently) understand "actual guidance" in the above, mundane sense. And this mundane sense makes Smith's argument sound. Given the mundane sense, Fatima is not actually guided by her deontic deliberation. And thus, contrary to the authors' claims, we do not need Capable$_{\text{CG}}$ to render Smith's argument sound.

Remember: the authors claim that a deontic evaluation according to which an action is both right and wrong still provides *actual* guidance. But, as we have now seen, this move presupposes two things: (1) an extremely broad sense of "actual guidance" and (2) the claim that this is *the only sense* of "actual guidance" available to Smith. But this extremely broad sense of "actual guidance" does not seem to be the only available sense. We formulated a different, narrower sense of "actual guidance" that is natural, coherent, and *prima facie* plausible.

Moreover, our mundane sense of "actual guidance" can be defended with the platitude argument. While it is clearly not platitudinous that deontic evaluation is capable of *correctly* guiding action, it seems quite obvious that deontic evaluation is capable of *mundanely* guiding action, that is, capable of helping to settle what to do by changing the agents' subjective deontic landscape and, thereby, her motivational setup.

1.2. Actual Guidance Does not Render Smith's Argument Uninteresting

Smith's argument concludes that the concepts RIGHT and WRONG never apply to the same action. According to Sharadin and van Someren Greve, this is a "substantive, first-order normative result" (207). But, they continue, in order to reach this result, Smith needs Capable$_{\text{CG}}$. Capable$_{\text{AG}}$, they think, is not enough.[17] But again, I am not convinced.

Let us start by reflecting on the meaning of "apply" in the conclusion of Smith's argument. "Apply," as used here, cannot be a success term. A deontic evaluation is the application of deontic concepts to actions. And such applications

[16] In an email conversation, one of the authors suggests that one actually guides a remote-controlled car by randomly waggling the controls. This strikes me as grist for my mill. It is at least not self-evident, I think, that "providing" a remote-controlled car with completely random directions can naturally be called "guidance." Instead, it seems more natural to say that the car is "shoved around"—and, intuitively, something similar seems to hold in the case of Fatima's deontic evaluation.

[17] "It it not clear that there is *any* interesting instance of [the schematic argument] that is true when 'guidance' is interpreted as 'actual guidance'" (214).

can be incorrect. So, according to Smith's argument, it is a mistake *to think* that both concepts apply to one and the same action. This means that her conclusion—that RIGHT and WRONG never apply to the same action—must be understood as a claim about *concepts* and not about *properties*. The argument tells us that it is *incoherent to think* that an action is both right and wrong.

Now, as we saw, Sharadin and van Someren Greve claim that Smith's argument is interesting and substantive. But what exactly, according to them, renders a conclusion interesting and substantive? Unfortunately, this does not become entirely clear. At times, they seem to suggest that the schematic argument delivers substantive (and interesting) results only if it tells us something about deontic properties, or an "independent realm of facts" (218, 227).[18] At other times, they sound as if the argument is substantive because it has "first-order normative" implications about the correctness of deontic principles (207). Then again, they also describe the schematic argument as providing a "substantive conclusion about the nature of deontic concepts" (206), a claim that does not involve anything about deontic properties. What to make of this?

Three remarks are in order. First, it seems clear to me that the schematic argument, since it is an argument about the nature of deontic concepts, *could not*, by itself, deliver insights about an "independent realm of facts" or deontic properties. The reason is simple: concepts are not properties or facts. Therefore, expecting that the schematic argument may yield conclusions about deontic properties or facts rests on a confusion between concepts on the one hand and properties (or facts) on the other.[19]

Second, note that arguments need *not* imply anything about deontic properties or facts in order to have first-order, normative implications. Suppose somebody proposes the first-order moral principle that an action is right just when it increases the number of squared circles in the universe. It is a legitimate criticism of this principle that, given the *meaning* of "square" and "circle," the principle is incorrect.[20]

[18] This, I take it, is what motivates their discussion of the platitude argument for realists in section 2.2.

[19] It is a main point of the authors' paper that "we cannot infer from the fact that a kind of evaluation (and the concepts it involves) provides [actual guidance] to the fact that it provides [correct guidance]" (215). I fully agree. And the reason is, as we just said, that concepts are not properties or facts.

[20] In a footnote, the authors note that Smith's argument is "explicitly directed" at an act-utilitarian principle proposed by Tännsjö, "Moral Conflict and Moral Realism" (Sharadin and van Someren Greve, "Is Deontic Evaluation Capable of What It Is For?" 207n14. My point here is simply that the argument does not need to appeal to properties or facts in order to be "directed" at a first-order principle.

Third, this means that, if Smith's argument is "interesting and substantive," it must be because it tells us something about the nature of deontic concepts and *not* about properties or facts. But then it becomes unclear why Sharadin and van Someren Greve believe that we need Capable$_{CG}$ in order for Smith's argument to generate an interesting and substantive result. After all, the notion of "correct guidance" is only introduced to distinguish deontic evaluations that point us toward the *objective deontic properties* from those that do not. But, as we just said, the conclusion of Smith's argument—as the authors themselves formulate it (207)—says nothing about objective deontic properties. And, being an argument about *concepts*, how could it?

So, the crucial question is whether or not Smith's argument can deliver "interesting and substantive" results without implying anything about objective deontic properties or an independent realm of deontic facts. I think it can, for the following two considerations. First, as the authors themselves stress, Smith's argument has the already mentioned first-order normative implication; it tells us something about the conditions under which deontic principles are incoherent. This, I think, qualifies as interesting. Second, Smith's argument tells us something about the relation between the concepts RIGHT and WRONG: they are mutually exclusionary. This means, like we just said, that it is incoherent to think they apply to the same action, which, interestingly, implies that the following two claims are inconsistent:

1. Our concepts RIGHT and WRONG refer to rightness and wrongness.
2. There are moral dilemmas; it is possible that an action instantiates rightness and wrongness.

These claims are inconsistent because RIGHT and WRONG are exclusionary concepts and, thus, could not refer to non-exclusionary properties. This result, I think, qualifies as "interesting" and, arguably, as "substantive." It implies that moral philosophers who use *our* concepts RIGHT and WRONG and who believe in moral dilemmas must change something about their views. This, I think, qualifies as an interesting and substantive result. Thus, we do not need Capable$_{CG}$ to make Smith's argument interesting and substantive. Capable$_{AG}$ (perhaps in our mundane sense) is sufficient. With actual guidance, Smith's argument bears the result that it is conceptually incoherent to believe that an action is both right and wrong—a result that bears interesting implications, two of which we just mentioned.

This concludes my discussion of Sharadin and van Someren Greve's treatment of the platitude argument. As we have seen, they claim (1) that we must understand *Capable* in terms of correct guidance or else Smith's argument is neither sound nor interesting. And, they continue, (2) *this* sense of *Capable* is not a platitude. Therefore, (3) the platitude argument fails. We have argued that 1 is false. There is a natural and coherent ("mundane") understanding of "actual guidance" that renders Smith's argument both sound and interesting. Moreover, as already suggested at the end of the last section, it does not seem far-fetched to defend the idea that deontic evaluation is platitudinously capable of guiding action in the mundane sense. It seems quite obvious that, at least in some contexts, coming to believe that *ϕ* is wrong changes the agent's subjective deontic landscape and her motivational setup.[21]

2. The Pragmatist Argument

Some authors claim that deontic concepts are *essentially* (by definition) capable of guiding action.[22] Sharadin and van Someren Greve discuss this proposal under the label "pragmatism." While they agree that it is a promising defense of Capable and declare their sympathies with it, they stress, again, that it would render Smith's argument uninteresting und non-substantive (225–26). The pragmatist strategy, they think, prevents instances of the schematic argument from "representing independent arguments for novel, surprising conclusions about the nature of deontic concepts," and such arguments would simply become "long-winded ways of reiterating one's view about the functional role of deontic concepts" (226).

[21] In response, the authors might suggest that, platitudinously, even astrological concepts like TAUREAN can help agents settle on what to do in the mundane sense, which hardly generates philosophically interesting results (cf. 214–15). Two short comments. First, I never claimed that we get interesting philosophical results from the fact that, platitudinously, deontic concepts can guide action in the mundane sense. The interesting results are generated from this platitude *in the context of Smith's argument*; and, in particular, from the idea that it is the *function* of deontic concepts to guide action. The fact that even astrological concepts meet the "incredibly low bar" (215) of being capable of actually guiding action changes nothing about that. Second, the authors discuss astrological concepts to drive home the point that *actual* guidance and *correct* guidance are worlds apart (cf. 214–5). This is, of course, true (since concepts are not properties or facts). But, again, this changes nothing about the interesting results generated by Smith's argument.

[22] For a discussion of the pragmatist argument, see Sharadin and van Someren Greve, "Is Deontic Evaluation Capable of What It Is For?" sec. 3. One "pragmatist" author they do not mention is Simon Kirchin, who characterizes all normative concepts by saying that their "logic … dictates that direction follows" ("Evaluation, Normativity, and Grounding," 183). A further note: I agree with Sharadin and van Someren Greve's suggestion (225) that Smith is most charitably understood as a pragmatist. If true, pragmatism would explain why Capable is a platitude, thereby rendering an independent discussion of the platitude argument obsolete.

Again, and for the reasons already mentioned in the previous section, I find this overly pessimistic. "Mere" implications of one's view about the functional role of deontic concepts can bear philosophically interesting fruits. For example, Smith's argument implies that certain deontic principles are incoherent. So, it seems, even by the authors' own standard of what renders an argument interesting and substantive (207), Smith's argument—with a pragmatist defense of Capable—is interesting. Therefore, Sharadin and van Someren Greve's criticism of the pragmatist argument seems unwarranted.

Let me close with one further remark about the authors' final assessment of pragmatism:

> But adopting [pragmatism] requires giving up a certain degree of ambition when it comes to limning the nature of the deontic by way of our practices. Or rather, it requires giving up the thought that, in doing so, one is limning the nature of some independent realm of facts that can be characterized independently from the way in which it is embedded in the lives of creatures like us. (227)

I believe that this is a mischaracterization of the situation we find ourselves in at end of the authors' paper. As they themselves rightly pointed out, the *actual* application of a concept to an action and its *correct* application can be worlds apart. This gap could not, in principle, be closed by reflections on the nature of deontic concepts alone. So, *any* attempt to close the gap in this way is bound to fail. Consequently, anyone who shares the above "ambition" confuses the subjective deontic landscape—the deontic properties as they present themselves in thought—with the objective one. And, therefore, giving up the endeavor to close the gap between the subjective and the objective deontic landscape via the schematic argument is not giving up an ambition—it is clearing up a confusion.

In the end, Sharadin and van Someren Greve's article leaves me somewhat perplexed: Do proponents of the schematic argument typically believe to limn an independent deontic realm? I suspect that many authors who use the schematic argument are pragmatists (about Capable) and do not aim to limn an independent deontic realm—in which case a part of the critical points Sharadin and van Someren Greve raise would not apply.[23] In any case, we learned this much: if you are a realist about deontic properties, you cannot use the schematic argument to gain insights about an independent deontic realm. My criticism notwithstanding, the authors' considerations nicely drive this point

[23] I thus think that the importance of the authors' considerations (partly) depends on the number of realists who actually proceed in this way.

home. But their assessment of the pragmatist strategy still seems overly pessimistic. As long as we do not expect the schematic argument to tell us something about an independent deontic realm (which, again, most pragmatists probably would not expect), the argument is capable of providing interesting philosophical results. Given their sympathies with the pragmatist strategy, Sharadin and van Someren Greve might welcome this result.

3. Conclusion

In this note, I have argued that Sharadin and van Someren Greve's attempt to pull into doubt a seemingly self-evident idea, Capable, is unsuccessful. According to the authors, two possible defenses of Capable—the platitude and the pragmatist argument—fail because they cannot defend Capable in the sense needed ("correct guidance") to achieve interesting and substantive results. Here, I have argued that we do not need to understand Capable in this sense to achieve results of this kind. In particular, and despite the authors' claim to the contrary, I showed that Smith's instance of the schematic argument is both sound and interesting if we understand Capable in a natural, intuitively plausible, mundane sense of "actual guidance" that falls well short of "correct guidance." And, in defense of the pragmatist strategy, I have argued that, as long as we do not expect the schematic argument to deliver results it would be (metaphysically) confused to expect, we may remain optimistic about the prospects of learning something philosophically interesting from it.[24]

University of Konstanz
stefan.fischer@uni-konstanz.de

References

Bykvist, Krister. "Violations of Normative Invariance: Some Thoughts on Shifty Oughts." *Theoria* 73, no. 2 (October 2008): 98–120.

Carlson, Erik. "Deliberation, Foreknowledge, and Morality as a Guide to Action." *Erkenntnis* 57, no. 1 (July 2002): 71–89.

Copp, David. "'Ought' Implies 'Can' and the Derivation of the Principle of Alternate Possibilities." *Analysis* 68, no. 1 (January 2008): 67–75.

Kirchin, Simon. "Evaluation, Normativity, and Grounding." *Aristotelian Society*

[24] Thanks to the anonymous referees from *JESP* for their constructive feedback and to Nate Sharadin and Rob van Someren Greve for a short but helpful email discussion.

Supplementary Volume 87, no. 1 (July 2013): 179–98.

Sharadin, Nathaniel, and Rob van Someren Greve. "Is Deontic Evaluation Capable of Doing What It Is For?" *Journal of Ethics and Social Philosophy* 19, no. 3 (March 2021): 203–29.

Smith, Holly M. "Moral Realism, Moral Conflict, and Compound Acts." *Journal of Philosophy* 83, no. 6 (June 1986): 341–45.

Tännsjö, Torbjörn. "Moral Conflict and Moral Realism." *Journal of Philosophy* 82, no. 3 (March 1985): 113–17.

Wiland, Eric. "How Indirect Can Indirect Utilitarianism Be?" *Philosophy and Phenomenological Research* 74, no. 2 (March 2007): 275–301.