JOURNAL *of* ETHICS
& SOCIAL PHILOSOPHY

VOLUME XXVII · NUMBER 3
May 2024

ARTICLES

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 Division of Arts and Humanities at New York University Abu Dhabi.

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 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
vol. 27, no. 3 · may 2024

https://doi.org/10.26556/jesp.v27i3.2005

A THEORY OF COLLECTIVE VIRTUE

Matthew Baddorf and Noah McKay

We say things like "Enron's greed led to catastrophe for many investors" and "the Eighty-Second Airborne was very brave." This language suggests that we believe that groups are capable of virtues and vices. It is hard to know what to make of this idea, however. Humans have virtues and vices due in part to our mental capacities, but attributing mental capacities to groups can sound absurd, like invoking vaguely Hegelian group spirits who work over and above their human members.[1] Pressure to avoid such results can lead to summative views of virtues and vices: on such views, a group's virtue or vice is just a result of "summing up" the same trait in its members. (So perhaps the Eighty-Second Airborne was brave just because most of its members were.) Unfortunately, this safely reductive view suffers from counterexamples, such as Lahroodi's example of a church committee that is closed-minded due to social pressures despite being made up of open-minded individuals.[2]

If we want to understand how groups can have important traits such as greed, bravery, and open-mindedness, we need a credible view that avoids the implausibility of Hegelian group minds and the counterexamples to summativism. Or so we think, anyway; and even if some readers think those Scylla and Charybdis safer than we do, we hope they will agree that there is room to attempt a middle course. Here, we set out a view that does just that. "Imitationism" is a kind of nonreductive theory that explains how a virtue can be genuinely collective without requiring collective minds.[3] We articulate and defend

[1] We should note that we are not actually accusing Hegel of holding this view; this is a caricature of Hegel, not an exegesis of his social philosophy. Also, not everyone agrees that collective minds are implausible. For a position friendly to group minds, see Theiner, "A Beginner's Guide to Group Minds." And for arguments that some popular positions in philosophy of mind imply the existence of groups minds, see Theiner and O'Connor, "Emergence of Group Cognition"; and Schwitzgebel, "If Materialism Is True, the United States Is Probably Conscious."

[2] Lahroodi, "Collective Epistemic Virtues," 287.

[3] Theories of collective virtue are theories about normative properties of collectives. These properties have been of interest to analytic philosophers since the Second World War. (Unsurprisingly, interest sprung up shortly after the discovery of particularly shocking widespread human rights abuses in group contexts; My Lai prompted an important early

our theory in section 1, explain how it accounts for some examples of collective virtue in section 2, and address two objections in sections 3 and 4. But first, a few comments about the rationale for and scope of our project are in order.

A key feature of our view is that it attempts to do justice to the intentional nature of collective virtue without making commitments to collective minds or collective phenomenal consciousness that many find implausible. The idea, in other words, is that we can not only deny collectives minds of their own, but we can preserve an insight at the heart of opposition to collective minds—the importance of intentional content.[4] And we can do this without having to deny the possibility of nonreductive collective virtue.

We will not consider all sorts of groups. The sort of groups we are going to discuss are groups we call "collectives."[5] Collectives are organized and structured groups that can survive the departure of at least some of their members.[6]

edited volume on the subject (discussed in French, *Individual and Collective Responsibility*, vii). Most work has focused on collective moral responsibility (and the free will required for it). For recent examples, see Copp, "On the Agency of Certain Collective Entities"; Haji, "On the Ultimate Responsibility of Collectives"; Braham and van Hees, "Responsibility Voids"; Hess, "Free Will of Corporations"; Dempsey, "Corporations and Non-Agential Moral Responsibility"; and Baddorf, "Phenomenal Consciousness, Collective Mentality, and Collective Moral Responsibility."

Attention to collective virtue lagged somewhat but has increased in the last two decades. Much work on collective virtue has been oriented around attempts to establish and account for nonreductive collective virtue. In our judgment, however, we do not yet have a fully satisfactory theory. For example, Lahroodi, "Collective Epistemic Virtues," considers and Fricker, "Can There Be Institutional Virtues?" develops accounts drawn from Gilbert's idea of joint commitments ("On Social Facts"), but (as we will argue in section 2) these cannot account for the wide variety of collective virtue cases. Beggs, "The Idea of Group Moral Virtue," gives a theory that (as we understand it) is consistent with our own but not as developed. Jones, "Numerous Ways to Be an Open-Minded Organization," gives helpful examples of the variety of collective virtues but does not attempt to provide a unifying theory explaining them. For more examples of contemporary work on collective virtues, see Anderson, "Epistemic Justice"; Ziv, "Institutional Virtue"; Gowri, "On Corporate Virtue"; Cordell, "Group Virtues"; and Diamantis, "The Law's Missing Account."

[4] For an argument readily adaptable for use against collective minds built around the importance of phenomenal consciousness, see Horgan and Kriegel, "Phenomenal Intentionality Meets the Extended Mind." One of the authors discusses the issue of collective minds further in Baddorf, "Phenomenal Consciousness, Collective Mentality, and Collective Moral Responsibility."

[5] Sometimes these are called "corporate agents" (e.g., in the title of Björnsson and Hess's "Corporate Crocodile Tears? On the Reactive Attitudes of Corporate Agents"). We will stick with the word "collective" and refer to virtues held by collectives as "collective virtues."

[6] The metaphysics of collectives is a very difficult topic, and we do not intend to make any contribution to it here. For readers who would like some idea of the metaphysical framework we are assuming, though, we can say this: we suspect that collectives are artifacts

They are typically capable of group goals and actions, at least in some sense; and these are not necessarily shared by all their members.[7] The two of us going to the store together are not a collective; British Petroleum and the French government are. Collectives need not all be legally recognized or subject to a written code of rules, as these groups are. An informal chess club, organized by volunteers who rely only on informal social norms to govern their behavior, could well be a collective also.

There may not be any bright, clear line between the groups that are collectives and those that are not. Some cases may be hard to classify. The important thing is that our claims are only intended to apply to groups that are clearly collectives; whether they apply to other groups is a question we will not address.

We are interested in both virtues and vices, but here we will often just speak about virtues. We suspect most such statements will apply to vices as well.

With all that said, we can turn to motivating our view. Some readers may wonder why they should care about what is basically a classificatory debate. Almost everyone agrees that collectives can have dispositions that are similar to human virtues but not exactly like them: there is something to the claim that the Red Cross is compassionate, even if it is not compassionate in just the way a person might be. Why not say, as Cordell suggests, that what appear to be genuine collective virtues are really not virtues but just highly desirable structural features?[8]

We are somewhat sympathetic to this worry. But there are practical advantages to recognizing the existence of collective virtues and vices. First, and likely most importantly, recognizing collective virtue allows for a sort of broadly moral criticism of collectives that recognizing merely beneficial structural features does not. "Moral blame" is a complex and disparate set of phenomena, and we think that there is at least one sort of moral blame for which collectives are not apt targets.[9] But there are some practices that have gone by the title "moral blame" for which collectives can be apt targets.

that have humans as proper parts (and which are capable of persisting through changes to their human parts). See Baker, *The Metaphysics of Everyday Life*, for an influential discussion about the reality of artifacts, and Uzquiano, "Supreme Court and Supreme Court Justices," for an adaptation of Baker's account to groups. For some good recent discussion of the metaphysics of groups in general, see Ritchie, "What Are Groups?" and Epstein, "Ontological Individualism Reconsidered."

[7] There has been a good deal of discussion recently about the nature of collective decision-making. For an authoritative summary of the ways that collective decision-making is importantly irreducible to that of individuals, see List and Pettit, *Group Agency*.

[8] Cordell "Group Virtues."

[9] One of the authors defends this claim in Baddorf, "Phenomenal Consciousness, Collective Mentality, and Collective Moral Responsibility."

Some of these practices are associated with attitudes such as disdain and admiration.[10] Paradigmatic examples of things we disdain and admire are human agents, but we can sensibly adopt these attitudes toward anything capable of a bad or good character. It is plausible that we can sensibly adopt these attitudes toward collectives; one can disdain a charity for its tendency for mission creep or admire a government for its justice. Collective virtues help account for and legitimize such attitudes. But mere structural or individual features—considered in themselves rather than as parts of the virtues they help constitute—do not.

Relatedly, consider the motivational effects of a belief that one's organization has a virtue or vice versus the belief that it merely has a feature with positive or negative effects. While the latter can certainly motivate one to preserve a collective's good features or reform its bad ones, we doubt that mere features have the motivational oomph most of us would get from learning that we are part of a virtuous or vicious organization. Collectives can have all sorts of positive and negative effects due to structural and individual features, many of which we do not have any strong obligation to encourage or discourage. But learning that your collective has a virtue involves learning more than this: it involves learning that you are participating in a morally praiseworthy endeavor. Recent empirical research suggests that this has a powerful effect on employees' willingness to identify with their employers.[11] Similarly, there is a natural horror one can feel when one discovers that one is participating in a morally diabolical group that is not felt simply by participating in a system with some equally undesirable consequences.

Finally, thinking in terms of collective virtues can also be practically advantageous for epistemic reasons. More particularly, it can help us conceptualize what it is that we are trying to get our collectives to do and be. Consider a scientific research group whose primary output is journal articles and that is thus analogous to an individual researcher. Thinking about their group in terms of the collective epistemic virtues that it does or does not exemplify can help the members of the team understand what it is that they are trying to make their group like. Making their group more virtuous might require moving beyond merely thinking about collective virtue, but that does not mean that it is not helpful to start with collective virtue. Skeptics might reply that one could, instead of starting with thoughts about what collective virtues the group should instantiate, start by thinking about what desirable structural features the group should have. But structural features combine with individual features in order to have their effects, and there are many different combinations of the

[10] Here we draw from Shoemaker, *Responsibility from the Margins*, particularly ch. 1.

[11] Chun, "Organizational Virtue and Performance."

two sorts of features that might result in a given effect. Thinking in terms of collective virtues would allow the group's members to start in the place that makes the most sense, with the sort of characteristic traits they want the group to have, rather than from the building blocks of those traits. Thinking in terms of collective virtues, then, can help us conceptualize what sort of agency-imitating collectives we should be striving for.

So much for the rationale. Now for the theory.

1. Imitationism: A Theory of Collective Virtue

Here is a summary of our view:

> *Collective Virtue* (cv): Collectives can possess virtues and vices; they can do so because they can possess reasonably broad and stable dispositions that functionally and intentionally imitate individual virtues and vices.

The gist of cv is just this: collectives have virtues because they have dispositions that meet certain conditions. Just what those conditions are needs explication. The basic idea, though, is that these collective dispositions resemble individual virtues closely enough to be virtues themselves.

The first notion in cv that needs some explanation is that of "reasonably broad and stable dispositions." By "stable dispositions," we mean dispositions that tend to endure over time; if a collective has one at a given time, then the collective will tend to have it in the future as well. By "broad," we mean that these dispositions trigger in an appropriately wide range of circumstances; a generous collective, like a generous individual, will usually not be arbitrarily generous in one situation but fail to be so in a similar one. Both stability and broadness admit of degrees; the claim that these dispositions are "reasonably" stable and broad is meant to indicate that collectives may not need a very high degree of either for their dispositions to count as virtues. Just as individuals need not always behave in accordance with a virtue in order to have it, so it is with collectives. Where the line ought to be drawn between virtue possession and lack is highly dependent on the virtue and is, of course, almost always difficult to discern precisely.[12]

[12] The extent to which individuals have dispositions that meet these conditions has been the subject of a great deal of dispute in (individual) virtue theory over the past fifteen years. See Phillips, "Towards an Empirically Adequate Virtue Ethics," for discussion. We will assume that individuals have sufficiently broad and stable dispositions that they can have virtues. It is worth noting, though, that almost all parties to the debate over individual virtues think that individuals' dispositions are often less broad or stable than we tend to pretheoretically think. This gives us reason to think that collectives' dispositions may not need to be as broad or stable as we might pretheoretically think in order to be on as good a footing to qualify as virtues as those of individuals.

The second notion that bears comment in CV is that of functional imitation. What we mean here is that collective dispositions can play either the same or very similar functional roles as those played by individual virtues. For example, collective bravery might enable an organization to correctly determine what should be done in cases of danger and to implement its decisions; this role is identical to or similar to the role that bravery plays in the case of the brave individual.

The third notion to discuss—and the one that we suspect will provoke the most controversy—is that of intentional imitation. The idea is similar to that of functional imitation: collective dispositions can involve the same or very similar intentional content as those involved in individual virtues. Let us first say a bit more about what this means and why it matters, and then explain why we should think it is true.

What would it mean for a collective disposition to involve intentional content? To answer this question, we should first figure out what it is for an individual virtue to involve intentional content. The phrase "intentional content" is not often discussed with regard to virtues, but there is a straightforward sense in which they can involve intentional content: having a virtue can sometimes involve possessing intentional states. For example, courage involves a disposition to certain attitudes toward danger. These attitudes are intentional: they are directed toward danger. Some collective dispositions can involve similar states that are directed toward danger in the same way (or similar ways) as the individual attitudes involved in courage.

These collective states are not necessarily the same states as those in individuals. For example, individual courage likely involves certain beliefs and desires. We doubt that collectives have beliefs and desires, strictly speaking. Why not? Briefly: because collectives lack minds, and mental states, beliefs, and desires can only be had by creatures with minds. (If we are wrong about this, of course, the similarity with individual states is only strengthened.) But even if they cannot have beliefs and desires, collectives can have states that have the same intentional contents as beliefs and desires. Intentional content is a more plausible feature of collectives than mental states since there are many things with intentional content but no mental states: spoken and written communication and much art, for instance. (See section 4 for more discussion of this sort of intentional content.)

For example, the Biden administration has taken an official stance on immigration, and this official stance might (and probably does) have the same intentional content as a set of beliefs that could be held by an individual. When we say that collective dispositions can intentionally imitate individual virtues, we mean that they can involve intentional states with the same (or very similar) intentional content. We can call this the Content Imitation view.

Why is this important to Imitationism? In short, because some virtues are essentially intentional (or so we think). Courage is not merely a matter of behaving fearlessly in the presence of danger—it involves a certain attitude *toward* danger. Similarly, compassion is not merely a matter of behaving a certain way, but also of being concerned *about* the well-being of others. So, in order to be courageous or compassionate, collectives must be capable of imitating these attitudes and concerns. It would be exceedingly odd to admire a collective for its courage and compassion while denying that it had any concern for the vulnerable or stalwartness toward danger.

Why think that the Content Imitation view is true? Briefly, collectives can be fruitfully understood from the intentional stance. In other words, attributing intentional contents to states of collectives allows us to make sense of their behavior in the same way that attributing intentional contents to states of individuals does. That is why we often explain and predict the behavior of collectives in intentional terms: the sentence "the Roman Catholic Church opposes birth control" works as an explanation of some behaviors of the Roman Catholic Church because the Roman Catholic Church is related to birth control in something very like the way that an individual with a con-attitude toward birth control is.

Some philosophers have suggested that intentionality *consists in* intelligibility from the intentional stance.[13] If that is right, then collectives can certainly have intentional states. However, we note that the Content Imitation view does not require anything this strong. Even philosophers who deny that intelligibility from the intentional stance is sufficient for intentionality normally think that it counts as good evidence for intentionality. So, while we cannot be certain that collective states are sometimes intentional, we are reasonably confident that they are. We are not alone in this respect; many philosophers have defended detailed accounts of collective intentionality in recent years.[14] We do not endorse any one of these accounts in particular, but we think they render the Content Imitation view plausible. (We realize this brief explanation will not satisfy everyone—for reasons of scope, we cannot expand on it much. But in section 4, we try to anticipate an objection to this part of the account without moving too far afield.)

[13] Dennett, in *The Intentional Stance,* holds that this is true for intentionality of all kinds. Kriegel has argued that it applies to derivative intentional content, even if nonderivative content is essentially linked to consciousness (*Sources of Intentionality,* ch. 4). See Tollefsen, "Organizations as True Believers," for an application of Dennett's view to collectives, and our section 4 for more on how our view is consistent with views like Kriegel's.

[14] See Schweikard and Schmid, "Collective Intentionality," for an excellent survey, and Toumela, *Social Ontology*, for a developed and illuminating account.

To bring what has been said thus far together: we are saying that the fact that collectives can possess broad, stable traits that functionally and intentionally imitate human virtues is the reason that collectives can possess virtues and vices. The idea is that in cases of collective virtue, the traits we have described form an explanation for the collective virtue in question. We are not committed to a particular view of what sort of explanation is involved here; presumably it is not a causal one, but some other sort—probably some kind of constitution or grounding.[15]

Here is the thought behind this explanatory claim (however its details are worked out): we can determine whether something counts as a virtue or vice by comparing it with paradigmatic instances of such things. (Paradigmatic instances of a virtue might include generosity, or courage, or other examples of traits commonly attributed to individuals.) Collective virtues and vices are not paradigmatic instances, but they are sufficiently similar to paradigmatic instances to count as instances. In large part, this is because they are, like paradigmatic instances, broad and stable, and because they both play the same or similar functional roles and involve the same or similar intentional content as those of individual virtues and vices.

We are not saying that collectives can possess all virtues and vices; there may be some virtues that involve things that collectives lack. In fact, when we attribute some trait to a collective—say, open-mindedness—it may be that the word "open-mindedness" refers to a somewhat different trait than when we use the word with respect to an individual. Obviously, the two traits could not be too dissimilar, or else it would not make sense to use the same word for each, but it is nonetheless possible that the collective trait is different from the individual one. This is worth bearing in mind when we consider a virtue like courage, which we might be tempted to think that collectives cannot possess. (Maybe it involves rightly facing a phenomenally conscious fear.) Even if they cannot possess human courage, it is natural to think that they can possess some collective analogue of it, and that is all our theory needs. Indeed, CV could be true even if collective virtues are not close analogues of any individual virtue. So long as the collective traits count as virtues due to their "formal" similarity to paradigmatic cases of individual virtue as described by CV, Imitationism allows for the possibility of "alien" collective virtues that are little like the virtues of human beings.

One last point about CV: one might wonder why we are not simply functionalists about virtue, claiming that possessing a virtue is a matter of instantiating

[15] Note that while we are providing explanations, we are not providing an analysis of the concept of collective virtue. For some reasons why we do not think this is a helpful endeavor, see Huemer, "Failure of Analysis and the Nature of Concepts."

certain functional roles.[16] If we were, then collective virtue would be comparatively straightforward, perhaps even easy to establish, and certainly the concept of a functional role has come up often enough thus far that it is plausible to think that functional roles are important for understanding virtue. In part, we do not endorse this view because it is not at all clear to us that the virtues possessed by individual human beings are actually simply a matter of functional roles. Phenomenal consciousness may well play an ineliminable part of certain human virtues. (In addition to the example of courage above, consider virtues that involve empathy, such as compassion.) Or perhaps not: but we think neutrality about this issue is an advantage of our view. If Imitationism is correct, then the existence of collective virtue is not hostage to the fortunes of functionalism.

2. Examples of Collective Virtue

Soon we will consider how our view can deal with some objections, but first, we want to consider some further examples of collective virtue and vice to show how our view can account for them. Let us start with a simple case of collective virtue, one similar to a case suggested by Donald Beggs.[17] Suppose a quilting group consisting of white, middle-class Americans regularly sells their quilts and donates the proceeds to charities. The members of the group each think that the group should not donate funds to charities that simply benefit people like themselves (e.g., charities that focus on finding cures for diseases that threaten the affluent); instead, they want to donate to charities that help the most disadvantaged members of society. So, each member is careful to ensure that the group's chosen charities fit with this goal. In particular, they exert pressure on the committee that selects charities to choose appropriately. In this case, the group has a virtue: the group cares for those who are most disadvantaged, despite the fact that they are often socially, ethnically, and economically different from the group's members. Beggs calls this virtue "radical tolerance," but we could just as aptly name it "compassion."[18]

[16] We would like to thank an anonymous reviewer for pressing this question.

[17] This case is a variation on Beggs's example of a quilting group ("The Idea of Group Moral Virtue," 467–68). In his original case, the members do not each exemplify the care and concern mentioned below.

[18] It is worth remembering here that, on our view, the trait of collective compassion might not be identical to the individual trait that gives it its name. Individual compassion might require, say, a phenomenally felt emotion, and probably collectives cannot have those. See the previous section for more discussion (and thanks to an anonymous reviewer for helping us see that a reminder is in order here).

Now, the individual members of this group have the same virtue as the group. So, we do not give this example as an example of a group whose virtue is in some strong way irreducible to those of its members. It does show a simple and straightforward case of collective virtue, however, and one that is aptly explained by CV. For in this case, the collective has traits that meet CV's requirements. It is disposed to perform actions that are functionally and intentionally similar to an individual's virtuous choice: the group selects charities *because* donating to those charities is likely to improve the lot of those who are worst off. And the vigilance of the members ensures that these dispositions are reasonably broad and stable; in a variety of different sorts of fundraising cases, the quilting group will make these choices, and it is likely to retain this disposition for some time. We think it is plausible that these sorts of considerations explain the fact that the quilting group has the virtue of compassion. It is by way of these traits that it does so. In this case, the quilting group has the appropriate dispositions because each of the individual members has the same virtue. But the virtues of individuals do not automatically "rise" to the level of the group; each individual quilter could be compassionate and simply express it outside the group. So, CV is playing some explanatory role here, despite the fact that the example is not an obvious counterexample to summativism.

Let us turn to a less apparently reductive case. This next example involves a collective vice—namely, collective carelessness. It is based on an example given by Kendy Hess, who gives it as part of an argument for irreducibly collective moral responsibility, but it is equally interesting as an example of a collective with a vice not shared by its members. In Hess's example, ACME corporation has employees who are committed to protecting the environment.[19] However, after ACME decides to produce steel additives, the distributed nature of its decision-making about the details causes a problem. In Hess's account:

> Member *A* requests proposals from Departments *α*, *β*, and *γ*,
> Member *B* picks the one from *α* and modifies it slightly to reduce costs,
> Member *C* modifies the proposal to improve materials handling,
> *D* modifies the proposal to improve efficiency,
> *E* modifies it to improve health and safety compliance (less worker exposure),
> *F* modifies it to use different (nationally available) chemicals, and
> *G* modifies it to reduce costs again.
>
> In the end, as a result of these piecemeal modifications and others during implementation—each innocuous and rational enough within

[19] Hess, "Free Will of Corporations (and other Collectives)," 247–48.

> its own limited sphere—the new production line results in a continuing discharge that pollutes a local river.[20]

We can suppose that the pollution involved is of a sort and level that is inconsistent with appropriate regard for the environment. So, it seems that ACME corporation is environmentally careless. Although the individuals involved may not have been able to realize it (given their limited individual knowledge and time), the fact that their actions together resulted in the discharge shows that ACME's decisions do not take environmental impacts into account. It is, therefore, plausible that ACME has a vice of disregard for the environment insofar as its method of decision-making does not seem to provide for environmental effects, at least when they are not obvious.

One might think that ACME lacks a disposition stable enough to count as a vice since, for all we know, given Hess's description, a slight change in circumstances could have resulted in a very different outcome.[21] Maybe if *G*, for instance, had not made that last modification, no pollution would have occurred. But note two things here: first, we could foreclose this possibility by assuming that the environment is extremely sensitive to ACME's activities and that the few ways of carrying out those activities in an environmentally responsible way are less efficient, affordable, and safe for personnel than the alternatives. On these assumptions, ACME is likely to cause some environmental damage in any given case. Second, and more importantly, vices of carelessness do not need to have a high probability of resulting in bad outcomes to be broad and stable enough to count as vices. Most careless drivers do not cause crashes on any given trip, and when they do cause a crash, it is often the case that the crash might easily have been avoided if they had made some apparently unimportant change (such as driving down a different road). Similarly, pollution need not be a high likelihood event for ACME to have a vice due to its lack of care to ensure such pollution does not occur.

To sum up: in this case, the corporation has dispositions to make decisions in certain ways, and these dispositions are functionally and intentionally similar to an individual who behaves with careless disregard for the environmental consequences of their actions. CV thus captures ACME's vice nicely.

Unlike the previous example, this is a case where the individuals involved do not share the collective trait: all the employees of ACME (and all the owners as well) might be environmentally virtuous yet not realize that their collective efforts make ACME vicious. We could also imagine a reversed case, in which

[20] Hess, "Free Will of Corporations (and other Collectives)," 248.

[21] We would like to thank an audience at the University of Rochester for making this objection.

most (perhaps all) employees were not environmentally virtuous, but a carefully followed company policy of checking for and correcting environmental problems resulted in a virtuous company. Either way, this sort of case is not susceptible to a reduction to the virtues and vices of individuals.[22] The reversed case is also an example of collective virtue that cannot be explained by theories that rely on the notion of a joint commitment to collective virtue, such as those of Fricker and Lahroodi, since in the reversed case, the employees of ACME never jointly commit themselves to caring for the environment.[23] These sorts of cases, by contrast, are nicely explained by CV.

The last case we want to look at now illustrates how Imitationism can account for virtues in "invisible hand" cases: cases where a collective outcome is the result of conflicting individual behavior. Miranda Fricker considers a case like the following one.[24] A jury is made up of biased individuals, but because their biases cancel each other out, the jury reaches a fair verdict. Fricker acknowledges that there is a sense in which the jury is fair-minded, considering all the evidence of the case and weighing it appropriately before coming to the most reasonable conclusion. Yet she thinks the jury unvirtuous because the members are all biased—the jury is fair-minded only because their biases cancel each other out. (Presumably, when juror *A* ignores some key bit of evidence because it does not conform to her bias, juror *B* points out the importance of the evidence, and the group ends up weighing it appropriately.) In this case, Fricker thinks there is no collective virtue because the resulting correct verdict was not the result of any good motivation or skill—it is merely the chance result of the individuals who happened to be selected.

We think that Fricker may be right to think that collectives whose good actions are the result of happenstance are not thereby virtuous. And CV can account for this since these good results are not the product of broad and stable dispositions. However, it is also possible that the jury's fair-minded result is not mere happenstance. (Fricker does not discuss this possibility.) Suppose the jury was formed by a legal system that reliably forms juries with individuals of sufficiently varied backgrounds and biases that juror's biases are likely to cancel each other out. This might or might not be the product of intentional design of the legal system. (We have no position on the extent to which any actual

[22] We think that at least some cases of collective virtue and vice are nonreductive in a stronger way: they do not supervene on (or reduce to) any set of intrinsic properties of individual agents. But defending that claim would take us too far afield here.

[23] Fricker, "Can There Be Institutional Virtues?"; and Lahroodi, "Collective Epistemic Virtues."

[24] See Fricker, "Can There Be Institutional Virtues?" 239. Fricker's original case concerned a debating society, but the point is the same.

legal system manages to achieve this, but it is our impression that the United States' adversarial jury selection system may be designed to produce this sort of effect.) In this case, it is not a coincidence that the jury behaves fair-mindedly. The jurists have been placed together in such a way that their behavior will lead to an outcome far better than the outcome that any individual jurist would have produced. This seems to us to be a case of a collective virtue. Our theory would handle this case by saying that the jury's reliable behavior is a collective functional and intentional analogue of an individual's disposition to be fair-minded; thus, the jury's finding would be interpreted as a result of a collective virtue.[25]

This case illustrates an advantage of Imitationism. We want to be able to differentiate between cases where a process simply happens to result in a good outcome and cases where there is a genuine collective virtue. Existing joint commitment accounts of collective virtue do this by requiring a joint commitment among individuals to the virtue, but this leaves out cases of collective virtue such as this one. CV, by specifying that collective virtues must be reasonably broad and stable, can do the same work and account for these cases.

In sum: we have good reason to believe that collectives can have virtues through broad and stable functional and intentional imitation of characteristics of paradigmatic individual virtues. This view allows collective virtues to be nonreductively held in a reasonably strong sense, and accounts nicely for a range of different sorts of collective virtue.

3. An Objection: Truly Agential Virtues?

One major objection is likely to linger in the minds of some readers: Are the collective virtues we have described really virtues in the sense that most ethicists and epistemologists use the term? Typically, the sorts of virtues that get philosophical attention are what one might call robustly agential: they have to do with advanced capacities for decision-making—capacities that require advanced capacities for reflection (e.g., on the value of the virtue in question) that Imitationism does not capture. This sort of objection has been developed by Sean Cordell into a dilemma: either a theory of collective virtue claims

[25] Would it be a collective virtue of the jury or of some other collective, such as the legal system? We are not sure. One might think that the jury could not have the sort of broad disposition necessary. For juries might have all their members essentially (replacing a single member might result in a new jury), and if the trial were different, then different jurists might have been selected—and if that is true, perhaps the jury does not exist in enough different possible circumstances for it to be capable of a broad disposition to be fair-minded. But if so, we think that it would be reasonable to think that the collective virtue producing the fair-minded outcome in this case could be a virtue on the part of the legal system as a whole: a virtue of selecting good juries.

that advanced psychological capacities are required for collective virtue, or it does not.[26] If it does, then it is hard to see how collectives can possibly meet the requirement (at least if one is trying hard to avoid Hegelianism or group minds, as we have). If it does not, then it is not clear that collective virtues are, after all, similar enough to the sort of moral and epistemic individual virtues that matter to be worthy of the virtue label. Cordell concludes that we should replace the idea of collective virtue with a somewhat more reductive account of desirable and undesirable structural features of collectives.

Both horns, we think, can be resisted. Consider Cordell's second horn, on which collectivists like us claim that collectives can have virtues *without* having advanced psychological capacities. Suppose that we humans need these sorts of psychological capacities in order to have full agential virtues. Even so, it could be that collective virtues can still be similar enough to our own to count as virtues without such collective capacities. For the collective might still be capable of virtues that are partially agential. Imagine a spectrum: on one side, there are beneficial qualities without any agential features. (Knives, for example, can have beneficial qualities such as sharpness, though they lack agential virtues.) On the other side of the spectrum are fully agential human virtues, virtues that involve the capacity for reflection on value. Even if they are not at the same end of the spectrum with fully human virtues, collectives might have virtues that are considerably closer to that end than to merely beneficial qualities. Collectives (unlike knives) could still be in states that are functionally and intentionally very similar to those of individuals with beliefs and desires. This imitation can result in robustly nonreductive cases of collective virtue and vice in ways that mimic those of individual agents. It is plausible that this is enough to make collective traits agential in ways that make them count as genuine virtues, in part because they are apt for the same kinds of appraisal and emotional response as virtues (as when we are, for instance, ashamed of associating with a vicious collective). In other words, they could be agential enough even if they are not as agential as us.

Still, we hold out hope that Cordell's challenge can be met by addressing the first horn of his dilemma.

On the first horn of Cordell's dilemma, virtues require that their possessor have certain psychological capacities: capacities for reflection, or at least the ability to act upon the value of those virtues. We are not sure whether such capacities are required for virtue in individual human agents, but suppose that they are. On our view, collectives might well be capable of such capacities—or

[26] See Cordell, "Group Virtues," 53–56. Cordell's original target was Donald Beggs's account (see Beggs, "Idea of Group Moral Virtue," 464–66). We are adapting and generalizing Cordell's dilemma somewhat.

of capacities that are functionally and intentionally similar enough to their individual counterparts to suffice for genuinely agential virtue. Key to any such account would be to work out functional and intentional equivalents of the required capacities that are equivalent in that they fulfill the reason(s) such capacities are required. What these reasons are depends on what the right account of individual virtue says about why these psychological capacities are required in individual cases.[27]

We will not pursue that project here, but we will note one promising avenue. It seems plausible that some individuals could reflect on the value of a collective virtue in such a way that their reflection instantiated collective reflection. (The collective would be functionally and intentionally imitating such reflection by way of the individual reflection.)[28] This would not make the collective redundant; the virtue would be the collective's since (by hypothesis) other elements of the collective (individuals as well as nonhuman things like computer systems and bylaws) will be needed to instantiate other aspects of the collective virtue. If all this is correct, then Imitationism need not make the concession that collective virtues are not as fully agential as individual human virtues. Collectives can, in their own way, have virtues as robustly agential as those of individuals.

4. Another Objection: Genuine Intentional States?

In section 1, we claimed that collectives are able to imitate human virtues in part by having states with the same intentional contents as states that partly constitute human virtues. We think this is reasonable since collectives are sometimes best understood from the intentional stance, and this is good evidence that some of their states are intentional. However, some philosophers might worry that the Content Imitation view is incongruent with a growing research project in the philosophy of mind, which Kriegel dubs the "phenomenal intentionality research program," or "PIRP."[29] PIRP theorists hold that there is some kind of tight connection between intentional states and phenomenal consciousness.

[27] For promising examples of how this sort of argument can go in a discussion of the psychological requirements for collective moral responsibility, see Björnsson and Hess, "Corporate Crocodile Tears?"; and Collins, "I, Volkswagen."

[28] An anonymous reviewer has suggested that this proposal might allow a collective to have a mind after all. There is a lot that could be said here, but in brief, while this might be true, we suspect that even here there would not be a collective mind, but simply a collective making use of an individual mind. To put it another way, the collective would have no mind of its own, but would merely appropriate the reflection of a human mind. More inspiration for how this might work could be gotten from Collins, "I, Volkswagen."

[29] Kriegel, *Sources of Intentionality*.

Since most people, including us, think collectives cannot be phenomenally conscious, it is difficult to see how they could be in states with intentional content if some version of PIRP is true.

The nature of intentionality is among the most vexed topics in all of philosophy, and we will not try to add to the debate about it here. We will simply offer a few reasons to think that, even if one of the aforementioned theories of intentionality is true, this does not *prima facie* pose a serious problem for Imitationism. In fact, we think our view is consonant with PIRP in that we, like PIRP theorists, are keen to avoid the overbroad ascriptions of mental states that can come from excessive reliance on functional analyses.

Even those sympathetic to PIRP typically allow that some nonconscious states can have intentional content in a derivative way.[30] For example, most would allow that a written token of the sentence "All men are mortal" has the same intentional content as the thought that all men are mortal, though only derivatively. Similarly, many allow that subconscious mental states have intentional content derivatively, and some of these subconscious states are partly constitutive of virtues and vices. A subconscious bias against members of a certain racial group, for instance, might have that racial group as its intentional object, and this might be part of what makes the bias genuinely vicious.

So, there is reason to think that some nonconscious states can have derivative intentional content, even if phenomenal consciousness is necessary for underived content, and some such states can apparently partly ground or constitute virtues and vices. To those readers who are allies of PIRP, we submit that the intentional states with which the Content Imitation view is concerned have their content derivatively.[31] As we said before, we do not have a theory of collective or derivative intentionality. But we hope that the right theories will accommodate collective intentional states like the ones Imitationism requires.

[30] Not all PIRP theorists believe this. For strong eliminativist theories of nonconscious intentional content, see Strawson, "Intentionality and Experience" and "Real Intentionality"; and Georgalis, *Primacy of the Subjective*. We think it strains credulity to insist that there are no subconscious intentional states or that sentence tokens have no intentional content.

[31] For various accounts of derived intentional content among PIRP theorists, see Searle, *Rediscovery of Mind*, ch. 7; Loar, "Reference from the First-Person Perspective"; Horgan and Tienson, "Intentionality of Phenomenology and the Phenomenology of Intentionality"; Horgan and Graham, "Phenomenal Intentionality and Content Determinacy"; and Kriegel, *The Sources of Intentionality*, ch. 4. We think a satisfactory account will allow for collective intentionality. For example, Kriegel argues that a nonconscious state has derivative intentional content if an ideal observer approaching the world from the intentional stance would ascribe intentional content to that state. On a view like this, the Content Imitation thesis is unproblematic: collectives have derivative intentional states because an ideal observer would interpret some of their states as intentional.

Imitationism can, then, deliver robustly agential virtues even given theories of intentional content that accord great importance to phenomenal consciousness, all while eschewing Hegelian group minds.[32]

Walters State Community College
matthew.baddorf@ws.edu

Purdue University
mckay24@purdue.edu

References

Anderson, Elizabeth. "Epistemic Justice as a Virtue of Social Institutions." *Social Epistemology* 26, no. 2 (2012): 163–73.

Baddorf, Matthew. "Phenomenal Consciousness, Collective Mentality, and Collective Moral Responsibility." *Philosophical Studies* 174, no. 11 (November 2017): 2769–86.

Baker, Lynne Rudder. *The Metaphysics of Everyday Life: An Essay in Practical Realism*. Cambridge: Cambridge University Press, 2007.

Beggs, Donald. "The Idea of Group Moral Virtue." *Journal of Social Philosophy* 34, no. 3 (September 2003): 457–74.

Björnsson, Gunnar, and Kendy Hess. "Corporate Crocodile Tears? On the Reactive Attitudes of Corporate Agents." *Philosophy and Phenomenological Research* 94, no. 2 (March 2017): 273–98.

Braham, Matthew, and Martin van Hees. "Responsibility Voids." *Philosophical Quarterly* 61, no. 242 (January 2011): 6–15.

Chun, Rosa. "Organizational Virtue and Performance: An Empirical Study of Customers and Employees." *Journal of Business Ethics* 146, no. 4 (December 2017): 869–81.

Collins, Stephanie. "I, Volkswagen." *Philosophical Quarterly* 72, no. 2 (April 2022): 283–304.

Copp, David. "On the Agency of Certain Collective Entities: An Argument from 'Normative Autonomy.'" *Midwest Studies in Philosophy* 30, no. 1

[32] We would like to thank Abby Brengle, Sean Cordell, Bill FitzPatrick, R. E. Fulton, Salim Furth, Kate Ritchie, Michael Robinson, Bill Rowley, Brett Sherman, and anonymous reviewers for the *Journal of Ethics and Social Philosophy* for their feedback and discussion. Similar thanks for constructive feedback are due to audiences at the 2016 Pacific APA, the University of Rochester, the 2017 Philosopher's Cocoon Conference, and the 2018 International Social Ontology Society.

(September 2006): 194–221.

Cordell, Sean. "Group Virtues: No Great Leap Forward with Collectivism." *Res Publica* 23, no. 1 (February 2017): 43–59.

Dempsey, James. "Corporations and Non-Agential Moral Responsibility." *Journal of Applied Philosophy* 30, no. 4 (November 2013): 334–50.

Dennet, Daniel C. *The Intentional Stance*. Cambridge, MA: MIT Press, 1989.

Diamantis, Mihailis E. "The Law's Missing Account of Corporate Character." *Georgetown Journal of Law and Public Policy* 17 (2019): 865–96.

Epstein, Brian. "Ontological Individualism Reconsidered." *Synthese* 166, no. 1 (January 2009): 187–213.

French, Peter A., ed. *Individual and Collective Responsibility*. Rochester, VT: Schenkman Books, 1998.

Fricker, Miranda. "Can There Be Institutional Virtues?" In *Oxford Studies in Epistemology*, vol. 3, edited by Tamar Gendler and John Hawthorne, 235–52. Oxford: Oxford University Press, 2010.

Gowri, Aditi. "On Corporate Virtue." *Journal of Business Ethics* 70, no. 4 (February 2007): 391–400.

Haji, Ish. "On the Ultimate Responsibility of Collectives." *Midwest Studies in Philosophy* 30, no. 1 (September 2006): 292–308.

Hess, Kendy M. "The Free Will of Corporations (and other Collectives)." *Philosophical Studies* 168, no. 1 (March 2014): 241–60.

Horgan, Terence, and George Graham. "Phenomenal Intentionality and Content Determinacy." In *Prospects for Meaning,* edited by Richard Schantz, 321–44. Amsterdam: de Gruyter, 2012.

Horgan, Terence, and Uriah Kriegel. "Phenomenal Intentionality Meets the Extended Mind." *Monist* 91, no. 2 (April 2008): 347–73.

Horgan, Terence, and John Tienson. "The Intentionality of Phenomenology and the Phenomenology of Intentionality." In *Philosophy of Mind: Classical and Contemporary Readings,* edited by David J. Chalmers, 520–33. New York: Oxford University Press, 2002.

Huemer, Michael. "The Failure of Analysis and the Nature of Concepts." In *The Palgrave Handbook of Philosophical Methods,* edited by Chris Daly, 51–76. New York: Palgrave MacMillan, 2015.

Jones, Todd. "Numerous Ways to Be an Open-Minded Organization: A Reply to Lahroodi." *Social Epistemology* 21, no. 4 (2007): 439–48.

Georgalis, Nicholas. *The Primacy of the Subjective: Foundations for a Unified Theory of Mind and Language*. Cambridge, MA: MIT Press, 2006.

Gilbert, Margaret. *On Social Facts*. Princeton: Princeton University Press, 1989.

Kriegel, Uriah. *The Sources of Intentionality*. New York: Oxford University Press, 2011.

Lahroodi, Reza. "Collective Epistemic Virtues." *Social Epistemology* 21, no. 3 (2007): 281–97.

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

Loar, Brian. "Reference from the First-Person Perspective." *Philosophical Issues* 6 (1995): 53–72.

Phillips, Kathryn. "Towards an Empirically Adequate Virtue Ethics." PhD diss. University of Rochester, 2015.

Ritchie, Katherine. "What Are Groups?" *Philosophical Studies* 166, no. 2 (November 2013): 257–72.

Schweikard, David P., and Hans Bernhard Schmid. "Collective Intentionality." *Stanford Encyclopedia of Philosophy* (Fall 2021). https://plato.stanford.edu/archives/fall2021/entries/collective-intentionality.

Schwitzgebel, Eric. "If Materialism Is True, the United States Is Probably Conscious." *Philosophical Studies* 172, no. 7 (July 2015): 1697–721.

Searle, John R. *The Rediscovery of Mind*. Cambridge, MA: MIT Press, 1992.

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

Strawson, Galen. "Intentionality and Experience: Terminological Preliminaries." In *Phenomenology and Philosophy of Mind*, edited by David Woodruff Smith and Amie L. Thomasson, 41–66. Oxford: Oxford University Press, 2005.

———. "Real Intentionality 3: Why Intentionality Entails Consciousness." In *Real Materialism: and Other Essays*, edited by Galen Strawson, 281–306. Oxford: Oxford University Press, 2008.

Theiner, Georg. "A Beginner's Guide to Group Minds." In *New Waves in Philosophy of Mind*, edited by Mark Sprevak and Jesper Kallestrup, 301–22. New York: Palgrave MacMillan, 2014.

Theiner, Georg, and Timothy O'Connor. "The Emergence of Group Cognition" In *Emergence in Science and Philosophy*, edited by Antonella Corradini and Timothy O'Connor, 78–117. Abingdon-on-Thames: Routledge, 2010.

Tollefsen, Deborah "Organizations as True Believers." *Journal of Social Philosophy* 33, no. 3 (Fall 2002): 395–410.

Toumela, Raimo. *Social Ontology: Collective Intentionality and Group Agents*. New York: Oxford University Press, 2013.

Uzquiano, Gabriel. "The Supreme Court and the Supreme Court Justices: A Metaphysical Puzzle." *Noûs* 38, no. 1 (March 2004): 135–53.

Ziv, Anita Konzelmann. "Institutional Virtue: How Consensus Matters." *Philosophical Studies* 161, no. 1 (October 2012): 87–96.

Journal of Ethics and Social Philosophy
vol. 27, no. 3 · may 2024

https://doi.org/10.26556/jesp.v27i3.2641

THE VALUE OF UPTAKE

Anni Räty

Much of the recent philosophical literature on consent focuses on a debate between two kinds of views about what consent is. So-called mental views of consent claim that consent consists in a mental state or an attitude of some kind.[1] According to these views, having the right kind of mental state or attitude is both necessary and sufficient for morally transformative, successful consent.[2] "Behavioral" views of consent deny that a mental state is sufficient—something more than that is necessary for consent to work its "moral magic." According to one popular view, the additional element is communication—for example, a verbal yes, a nod, or an inviting gesture.[3]

Arguments for both types of views often appeal to ideas about what the *function* of consent is. Proponents of mental views tend to emphasize how consent functions to extend the consent giver's individual autonomy or control over her normative boundaries with others. Proponents of behavioral views sometimes emphasize how consent serves as a tool that lets us coordinate our actions with other people. These different ideas about what consent *does for us*—what its function is—motivate different views of what consent is and what it takes to give morally transformative consent.

My argument here will follow a similar strategy. I will argue that consent has an often-overlooked *relationship-shaping* function: acts of consent can shape our relationships with others directly when we gain permissions that are constitutive of a new kind of relationship. Indirectly, acts of consent can create trust, intimacy, and other preconditions of personal relationships. I will then argue that this function grounds an argument for a claim about what it takes to give

[1] Alexander, "The Moral Magic of Consent (II)"; Alexander, Hurd, and Westen, "Consent Does Not Require Communication"; Ferzan, "Consent, Culpability, and the Law of Rape"; and Hurd, "The Moral Magic of Consent."

[2] When *A*'s consent to *B*'s ϕ-ing is successful, or morally transformative, it releases *B* from an obligation not to ϕ. An attempt to consent can be undermined by factors such as coercion, deception, or incapacitation (e.g., due to intoxication)—in what follows, and for all cases discussed below, I will assume that none of these undermining conditions are present (consent is given voluntarily, with sufficient information, and so on).

[3] Dougherty, "Yes Means Yes."

morally transformative consent: when consent serves its relationship-shaping function, an act of consent needs to be cosigned by both parties. More precisely: when *A*'s consent to *B*'s ϕ-ing plays a relationship-shaping function, *A*'s consent needs to be *accepted* by *B* in order for it to be morally transformative (where ϕ is an action).

This argument has an upshot for the debate between mental and behavioral views of consent. Mental views of consent deny that consent requires acceptance by its recipient. Some (but not all) behavioral views also deny that consent requires acceptance. So there are views in both camps that are committed to saying that in all cases, a consent giver can *unilaterally* change her moral boundaries with others. If my argument here is correct, this is a mistake. Consent cannot be unilateral if it alters the parties' relationship.

I will start by discussing the backdrop of the debate between mental and behavioral views of consent in some more detail and explaining the distinction between unilateral and bilateral conceptions of consent. I will then explain consent's relationship-shaping function and show how it creates a need for acceptance, uptake, or cooperation of some kind on the consent recipient's part. We will want to know next what precisely is required—what *is* uptake? The answer to this question depends in part on our background view of consent and our motivations for it.

1. Two Distinctions in the Ontology of Consent

Sometimes when we talk about consent, it can be unclear whether we are talking about a speech act, a legal concept, or something else. When I talk about consent here, I am talking about a normative power.[4] More precisely, I am talking about the normative power that you exercise when you permit someone else's doing something by releasing that person from an obligation to not do that thing.[5]

[4] An anonymous reviewer has suggested that it may be a mistake for theorists of consent to assume that consent is a unified phenomenon at all. I doubt that the fact that "consent" means different things in different domains (philosophy of language, legal discourse, everyday parlance, etc.) gives us good reason to think that the *normative power* of consent is not a unified phenomenon. I believe it is: its essential feature is that it releases others from obligations. The reviewer may have in mind the idea that the normative power of consent looks different in different contexts, and I wholeheartedly agree: what it takes to release someone from an obligation they owe to us can depend heavily on things such as the parties' relationship, the risks of the interaction, and the stringency of the obligation in question.

[5] More precisely, you do this by releasing that person from an obligation *owed to you, the consent giver*, to not do that thing. Consent operates on what are sometimes called "directed"

This notion of consent may not cover everything that gets called "consent" in everyday parlance or in specialized domains such as the legal realm. For example, in everyday conversations about sexual consent, the word "consensual" is sometimes just meant to mean "morally permissible." This does not track the normative power of consent, because consent alone is not enough to guarantee that any encounter—sexual or otherwise—is morally permissible all things considered.[6] What matters is that what I talk about when I talk about consent here does track what is at issue in the debate between mental and behavioral views—let us turn to these now.[7]

The primary point of disagreement between mental and behavioral views is whether morally transformative consent requires an expression of consent in the consent giver's outward behavior. According to mental views, it does not; the right mental state—provided that the agent is not coerced or deceived or in a state where she is incompetent to consent—is necessary and sufficient for morally transformative consent.[8] Different mental views of consent differ on the details of which mental state they take to be relevant to consent. For instance, according to Heidi Hurd, to consent to someone's ϕ-ing is to intend that person's ϕ-ing.[9] Larry Alexander identifies consent with the "subjective

or "bipolar" obligations (Thompson, "What Is It to Wrong Someone?"; see also Darwall, "Bipolar Obligation").

In some exceptional circumstances, a third party *A* can release *B* from their obligation to *C*. For example, next of kin can be authorized to consent to a medical procedure on behalf of a comatose patient. I will set scenarios such as this aside here, and focus on the more common case where the relevant obligation is owed to the consent giver herself.

It is common to assume that bipolar obligations correspond to rights (the thought has its origins in Wesley Newcomb Hohfeld's influential analysis of legal rights in *Fundamental Legal Conceptions as Applied in Judicial Reasoning and Other Legal Essays*. If they do, then consenting is the very same thing as waiving a right. In light of recent challenges to the idea that all bipolar obligations correspond to rights, I will refrain from saying that consent is the power to waive one's rights against others. (See, e.g., Cornell, "Wrongs, Rights, and Third Parties"; and Martin, "Personal Bonds.")

[6] Suppose *A* consents to have sex with *B* and vice versa; their encounter may still be morally impermissible because it is harmful, or alienating, or infringes the rights of a third party (suppose *A* has promised *C* to never have sex with *B*). The same applies to interactions not involving sex.

[7] See, e.g., Wertheimer, *Consent to Sexual Relations*; Hurd, "The Moral Magic of Consent"; Bolinger, "Moral Risk and Communicating Consent"; and Dougherty, *The Scope of Consent*.

[8] Communicating that one has the relevant mental state can be instrumentally useful: it gives others good evidence that one has in fact consented. There may even be good reason to require that we secure good evidence of others' consent before acting in ways that risk being wrong if consent is not given. But this is different from saying that the behavior or evidence of it constitutes part of the consent giver's consent.

[9] Hurd, "The Moral Magic of Consent," 125–34.

mental state" of choosing to forgo a moral objection to another's action.[10] Kimberly Ferzan argues that consent is "willed acquiescence."[11]

There is a popular argument motivating mental views that appeals to the connection between consent and *autonomy*. By giving consent, a person can voluntarily choose to permit something that would otherwise wrong them. By revoking consent that was previously given, they can choose to impose an obligation on another person. The power to consent makes the consent giver, to use H. L. A. Hart's memorable phrase, a "small-scale sovereign" over the obligations that others owe to them.[12] Hurd, for example, appeals to this function to argue that consent must consist in a mental state:

> If autonomy resides in the ability to will the alteration of moral rights and duties, and if consent is normatively significant precisely because it constitutes an expression of autonomy, then it must be the case that to consent is to exercise the will. That is, it must be the case that consent constitutes a subjective mental state.[13]

We might wonder whether autonomy in fact resides in the alteration of one's rights and others' duties and whether consent is significant just because it plays this function. But even if we grant both of these points, there is a gap in the argument: outward behavior such as communication can also constitute an expression of the consent giver's autonomy. The fact that consent functions as an expression of autonomy does not tell decisively in favor of a mental view of consent.[14] Ferzan bridges this gap by claiming that "autonomy is best respected by recognizing that the consenter has it within his or her power to allow the boundary crossing simply by so choosing."[15] The thought here is that if consent is important because it expresses the consent giver's ability to will the alteration of her rights and others' duties, then it best expresses that ability if it is maximally within the consent giver's control. Our mental operations are more fully within our control than our outward behavior. So consent consists in a mental state—or so the argument goes.[16]

[10] Alexander, "The Moral Magic of Consent (II)," 165–66.

[11] Ferzan, "Consent, Culpability, and the Law of Rape," 402–7.

[12] Hart, *Essays on Bentham*.

[13] Hurd, "The Moral Magic of Consent," 124–25.

[14] Cf. Dougherty, *The Scope of Consent*, 25–26.

[15] Ferzan, "Consent, Culpability, and the Law of Rape," 405.

[16] Due to space constraints, I am not going to evaluate how well this argument supports mental views over alternatives. See Dougherty's *The Scope of Consent*, 30–34, for a more thorough assessment of this motivation for mental views. I will come back to the connection between consent and control in section 3.

Suppose now that *A* wants to consent to *B*'s doing something—say, entering *A*'s apartment. If consent consists in a mental state, then *B* has no part to play in this process. Consider, for example, Hurd's view: *A*'s consent consists in *A*'s intending that *B* enter *A*'s apartment. *A* can form this intention without any cooperation from *B*, without *B*'s knowledge, and with no regard to whether *B* wants to have permission from *A* to enter *A*'s apartment. This is a one-sided, *unilateral* conception of consent. So are other versions of the mental view: consent is given in the privacy of the consent giver's mind, and no one else needs to enter the picture.[17]

What about behavioral views? Behavioral views of consent reject the idea that a mental state is *sufficient* for morally transformative consent. As with mental views, many of the motivating arguments for behavioral views appeal to the functions of consent. For instance, consent plausibly serves to coordinate complex behavior between people and enables us to undertake joint projects with others. Consent that is publicly observable by its recipient and by third parties seems best suited for this purpose.[18]

Behavioral views may disagree over whether a mental state is *necessary*. Alan Wertheimer distinguishes between "hybrid" views, which consider both a mental state and an expression of consent in outward behavior necessary for consent, and "performative" views, which consider an expression of consent necessary and sufficient.[19] As I am using the term here, both kinds of views count as behavioral views of consent.

Whether a given behavioral view is *unilateral* or not depends on what kind of behavior is necessary for morally transformative consent. Consider, for example, the following view:

> *Successful Communication View*: *X* gives consent to *Y* if and only if *X* successfully communicates to *Y* that *X* is giving permission to *Y*.[20]

[17] An anonymous reviewer has suggested that there could be a view according to which morally transformative consent consists in (1) a mental state of some kind in the consent giver and (2) a justified belief in the recipient that the consent giver has the right kind of mental state. They suggest that a view such as this would be an example of a mental view that incorporates an uptake requirement. I have characterized mental views as views that are committed to the claim that a mental state is both necessary and sufficient for morally transformative consent. So strictly speaking, what we have here is neither a mental nor a behavioral view of consent. More importantly, I argue below that uptake is not a matter of the recipient knowing that a consent giver has done their part of the permission giving (section 2.3). The considerations I offer also rule out the proposal that uptake consists in a justified belief.

[18] Cf., e.g., Bolinger, "Moral Risk and Communicating Consent," 181.

[19] Wertheimer, *Consent to Sexual Relations*, 144–62.

[20] Cf. Dougherty, *Scope of Consent*.

Successfully communicating anything to an audience takes some work from the latter. If I tell my friend about my day but they pay no attention to what I am saying, then our communication is not successful—it falls apart. The successful communication view is what I will call a *bilateral* view of consent: consent is not given in the privacy of the giver's mind, and the sort of behavior that is necessary cannot be performed fully privately either. Compare this view to the following:

> *Pure Behavioral View*: *X* gives consent to *Y* if and only if *X* deliberately engages in behavior *B* that indicates that *X* is releasing *Y* from a duty.[21]

Clarifying which behaviors "indicate that *X* is releasing *Y* from a duty" would tell us whether the pure behavioral view is a hybrid view or a performative view of consent.[22] But whatever those behaviors are, the pure behavioral view does not require the consent giver's release-indicating behavior to be observed by the consent recipient (or by anyone else for that matter), nor does it require anyone else to take part in the behavior (by, e.g., being a receptive audience). The pure behavioral view is unilateral, just like mental views of consent.

In asking what consent is, philosophers have tended to focus on whether consent needs to be communicated. The literature therefore tends to focus on the distinction between mental and behavioral views. The distinction between unilateral and bilateral views has not been previously appreciated in the literature, and it carves the space of existing views of consent in a novel way.

I want to argue next that we should favor a bilateral view of consent. I will not say which one—what I say here leaves open the question of which substantive view of consent is correct. I will argue in favor of bilateral views by arguing that consent sometimes needs to be taken up or accepted by the consent recipient.

2. An Argument for Uptake

I will proceed in the following order. First, I will argue that consent has a function that many authors overlook. It has what I earlier called a relationship-shaping function—more on this in a moment. I will then show how this function

[21] Dougherty, *The Scope of Consent*, 120. This claim is half of Dougherty's "expression of will" view of consent: "*X* gives consent to *Y* if and only if either *X* gives consent to *Y* via a directive or *X* gives consent to *Y* via expressing permission" (*The Scope of Consent*, 124).

[22] Verbal communication is one type of behavior that can indicate release from a duty, as is signing a waiver, putting out a public notice, a nod, and so forth. If deliberately engaging in these entails that the agent does so with a particular mental state, the view is a hybrid behavioral view of consent. If not, it is a performative view.

supports the claim that consent needs to be accepted by its recipient in order to be morally transformative, at the very least in cases where consent plays its relationship-shaping function. I will then ask what acceptance or uptake *is*—what does the recipient need to do in order to gain the permission that is offered to her?

2.1. *How Consent Shapes Relationships*

The kinds of relationships that I am primarily interested in here includes relationships such as friendship, romantic partnership, relationships between family members, relatives, and colleagues. Relationships such as these are of interest to moral philosophers because they affect what we have reason to do and what we owe to one another. Friends typically have reason to help one another with their projects, family members owe one another duties of care and support, and monogamous romantic partners owe it to one another not to have other romantic relationships.[23]

In many cases, these reasons and obligations are not just an incidental feature of the relationship. They are an essential part of what it is to have that particular kind of relationship with someone. Most obviously, part of what it is to be in a monogamous relationship with another person is to owe it to that person not to have other romantic relationships. Likewise, you and I are friends in part because we have special reason in times of need to lend one another a hand, advice, or a shoulder to cry on. Family members who neither care for nor support one another may still be relatives, but their relationship is closer to one between strangers or acquaintances.

On the flip side of our relationship-based obligations are *permissions* that we have in virtue and as part of our special relationships. For example, casual touch, such as placing a hand on another person's shoulder, is typically permitted between friends and close acquaintances but not between strangers. A parent may be permitted to enter a child's bedroom to clean it up, but if a house guest were to do this, it would be an infringement of the child's privacy. People who are dating often give one another keys to their respective apartments, along with permission to enter when they please. And so on. The range of permissions

[23] Not all friendship, families, and partnerships are alike, of course. Which permissions and which obligations I have toward a particular friend, for example, is a complicated function of things such as our own understanding of our friendship, the prevalent understanding of friendship in our culture(s), past interactions between us, explicit agreements, personal preferences, and much, much more. I am going to rely here on what I believe to be commonly accepted ideas about friendship, family, partnership, and so on. But I acknowledge that these ideas are culturally specific, and that personal relationships and their attendant obligations are very malleable.

between two people tells us a great deal about their boundaries and their relationship with one another.

Consider now consent. Consent is a normative power that, when it is morally transformative, gives the recipient permission to act in a way that would otherwise wrong the consent giver. Since personal relationships are characterized by both the obligations they impose on us and the permissions that they grant to us, acts of consent can affect what our relationships with others look like. To illustrate, consider this case:

> *Nonmonogamy*: Colt and Larissa are a monogamous married couple. They are both interested in having romantic relationships with other people. After carefully discussing the matter, they both decide to give the other permission to date people outside of their marriage.

Granting one's monogamous partner permission to date other people is a clear and direct alteration of the existing relationship. In this case, the alteration is welcome to both parties: Colt and Larissa are both enthusiastic about their new nonmonogamous relationship. But we can easily imagine a case where this is not so—I will discuss a case like that in a moment. First, consider another case where consent alters a relationship but in a way that is less obvious:

> *Apartment Key*: Fernanda and Robbie have been dating for a few months. Robbie offers Fernanda a key to his apartment and says: "You can have this, and feel free to come and go as you please."[24]

As I mentioned earlier, people who are dating often give each other this particular permission. In modern Western dating culture, the act serves as a way of signaling a certain level of commitment to the relationship. The change that this permission might cause in Robbie and Fernanda's relationship is not as clear-cut as the change in Nonmonogamy. But accepting (or rejecting!) the permission clearly does make a difference to their relationship and takes it a step further.

[24] This is a case of *unsolicited* consent (cf. Pallikkathayil, "Consent to Sexual Interactions"). Some authors have recently argued for notions of consent that rule out the possibility of unsolicited consent: Jonathan Ichikawa argues that attributions of consent (and nonconsent) are linguistically inappropriate unless the consent giver is responding to a request, order, or command, or otherwise acting at "someone else's behest" (Ichikawa, "Presupposition and Consent," 1). For a similar notion, see also Rebecca Kukla, "That's What She Said." As I understand their arguments, Kukla and Ichikawa are primarily interested in a very specific speech act, which they contend can only be performed in response to someone else's request. My focus here is on the power that we have to release others from the obligations they owe to us, and it seems clear to me that this power is operative in cases of unsolicited permission giving, just as it is in cases where one person requests a permission from another.

There is some room for negotiation in cases such as this. Fernanda could respond to Robbie's proposal by saying: "I am okay with having the key to your apartment, but I do not want that to change anything between us. I want to be very clear that taking this key does not mean that I want a serious relationship with you." This can go some way toward preventing the unwanted changes in their relationship, though "overriding" the conventional or cultural meaning of an act such as this is a delicate—and often difficult—thing to do.

Consider one more case outside of the romantic realm:

> *Friends*: Phoebe and Monica are colleagues. Their past interactions have been strictly professional, but they have a good rapport. Phoebe is going through difficulties in her personal life. She approaches Monica and asks: "I know we do not really know each other like that, but is it okay if I ask you for advice about some personal stuff?"

Permission to share personal information and to ask personal questions is characteristic of friendship, which is a relationship that Phoebe and Monica do not yet have. If Monica allows Phoebe to share her worries with her, this changes things between them. Depending on how things unfold afterward, it may be the beginning of a path toward friendship.

These cases illustrate that consent has a relationship-shaping function: acts of consent can shape and alter our existing personal relationships. Unlike the other functions of consent mentioned so far (expressing autonomy, enabling cooperation), the role of consent in shaping personal relationships has received little attention in the literature.

Some authors, however, have argued that *promising* can shape, alter, and enable personal relationships. This should be expected, since consent and promising trade in the same currency of our obligations to one another: consent releases obligations; promises generate new ones. Seana Valentine Shiffrin, for example, argues that by making a promise to someone else, you at once make yourself accountable to the other person for acting as promised and grant them a kind of discretionary authority over whether you are bound to act as promised (the promisee, and only the promisee, can release a promissory obligation at will).[25] Without the power to make a binding promise, I could tell you that I *intend* to, say, return your book, or meet you for lunch. But I could not make myself accountable to you for doing so or give you a say in the matter. Shiffrin argues that being able to do this is a precondition of healthy personal

[25] Shiffrin, "Promising, Intimate Relationships, and Conventionalism." See also Dougherty, "Yes Means Yes."

relationships in which we are not vulnerable to each other's whims and can relate to one another as moral equals.

It seems to me that consent can play a similar role in creating the preconditions of healthy personal relationships. The duties that other people owe to us typically serve to keep them at arm's length from our bodies, our property, and our sphere of private thoughts and decisions. Consent releases these duties and thereby brings others closer to us—into domains that are normally off-limits to other people. This can foster vulnerability; trust; closeness; and physical, emotional, and intellectual intimacy. These are the more indirect ways in which consent can enable personal relationships.[26]

2.2. The Need for Uptake

Which relationships we have with the people in our lives matters to us a great deal. This may seem like an obvious point, but it is crucial to the argument that I want to make next. One reason why our relationships matter to us is this: the obligations that come with special relationships can be burdensome. For instance, becoming a parent often involves restructuring your daily life, habits, and routines (especially if material support such as parental leave and free childcare is not available), and parents sometimes have to set their own plans and wishes aside to provide for their children. On the flip side, special relationships can allow us to access goods that we could not enjoy otherwise: things like the joys of childrearing, friendship, and partnership—as well as more tangible goods such as the legal privileges of marriage and guardianship. Another reason why our relationships matter to us is that our relationships can reflect our deeply held values. For example, some people forgo marriage for political reasons or because they consider the institution of marriage outdated. Meanwhile, others desire to be married precisely because of the social meaning the institution has.

Given the overwhelming importance of special relationships, it seems to me that we would lack a very important power to shape our own lives if we lacked the power to form personal relationships or to shape our existing relationships. We have a strong interest in being able to shape our own lives in accordance with our values, desires, and plans; we therefore have a strong interest in having a say in what our relationships with other people look like. To be clear, this is not to say that we ought to have complete control or a unilateral say over which personal relationships we have and with whom we have them; I might wish very much to be someone's friend or lover, but I am not entitled to anyone's friendship or

[26] By the same token, promises can also directly alter our relationships by bringing changes to the obligations that in part constitute those relationships; for example, a promise to be monogamous, or an exchange of wedding vows, is a direct alteration of the parties' relationship.

partnership.[27] But this much seems true: we have a very strong interest in having a say in which personal relationships we have and with whom we have them.[28]

Given that acts of consent can affect our personal relationships, we have an equally strong interest in having a say in whether acts of consent that affect our relationships are morally transformative. This interest creates a need for the recipient's concurrence or cooperation in creating permissions through consent—a need for something like the recipient's uptake or acceptance.

In the next section, I will say more about what this notion of uptake or acceptance might be. First, let me illustrate the need for uptake by considering Apartment Key once more.

> *Apartment Key*: Fernanda and Robbie have been dating for a few months. Robbie offers Fernanda a key to his apartment and says: "You can have this, and feel free to come and go as you please."

Fernanda's having this permission would change her relationship with Robbie in ways we have already discussed. If she welcomes that change, all is well—but suppose she does not: suppose Fernanda does not want a serious relationship with Robbie, and so she does not want the permissions that are typically associated with a relationship of that kind. It seems clear that Fernanda should have a say in whether she gains the permission that is on offer here and that Robbie should not be in a position to impose it on her unilaterally.

This case illustrates how the relationship-shaping function of consent is in tension with our interest in having a say in the shape of our personal relationships. We can ease that tension by introducing a requirement for morally transformative consent, a requirement for the recipient's uptake, acceptance, or cooperation—in the next section, we will take a closer look at how this notion should be understood. Absent a requirement such as this, consent givers would be in a position to make unilateral changes to their relationships with others; this is the cost of adopting a unilateral conception of consent and a robust reason to favor a bilateral conception.

[27] I also do not mean to say that we should always be able to disengage from existing personal relationships at will—except for relationships that are abusive, toxic, or otherwise harmful. If one party to a relationship wrongs the other or violates the norms of the relationship, the wronged party may be perfectly justified in unilaterally disengaging from the relationship. I have in mind something more like a healthy partnership where both parties are dependent on one another emotionally and materially; here, disengaging unilaterally risks harm to both parties.

[28] What about involuntary relationships? We do not get to choose our relatives, neighbors, or colleagues. But we do get to (and have an interest in being able to) shape these relationships. That is to say, we have an interest in being able to negotiate and renegotiate our boundaries with the people we have involuntary relationships with.

Is there something else besides a requirement such as this that could ease the tension?[29] What we need here is something that prevents the consent giver from making unilateral changes to their relationship with the recipient. I have argued that changes to relationships sometimes happen through acts of consent—in other words, that consent has a relationship-shaping function. The solution, therefore, has to be a requirement on morally transformative consent that prevents the consent giver from unilaterally granting permissions.

2.3. *What Is Uptake?*

Let us take a closer look at what the requirement for acceptance, cooperation, or uptake should look like—from here on, I will call it the "uptake requirement" for short. Our first choice point is between what I will call *weak* and *strong* uptake requirements. A strong requirement applies to all cases of consent; a weak one is limited to a certain class of cases. So a strong uptake requirement for consent would say the following:

> In order for *A*'s consent to *B*'s ϕ-ing to release *B* from an obligation not to ϕ, *B* must accept *A*'s attempt to consent.[30]

I do not think that my argument here supports a strong uptake requirement such as this. The interest that the requirement is meant to protect is tied to consent's relationship-shaping function; this interest is only at stake in cases where an act of consent would alter the parties' relationship. So what I have said here supports the following weak uptake requirement:

> In any case where *A*'s consent to *B*'s ϕ-ing would change the relationship between *A* and *B*, in order for *A*'s consent to release *B* from an obligation not to ϕ, *B* must accept *A*'s attempt to consent.

There may be other functions of consent and other interests of ours that support a requirement stronger than this. But note that even this weak requirement is incompatible with unilateral conceptions of consent. If consent is always given in the privacy of a person's mind or through behavior that involves no one

[29] Thanks to an anonymous reviewer for pressing me to consider alternatives. I discuss one more alternative in note 32 below

[30] In the literature on promising, it is widely accepted that there is a strong uptake requirement for promising (see, e.g., Thomson, *The Realm of Rights*; and Liberto, "Promises and the Backward Reach of Uptake"). Challenging this view, Seana Valentine Shiffrin grants that promisees have an interest in being able to avoid "the sometimes charged relation of moral debtor to the promisor," but argues that protecting this interest only requires that the promisee be in a position to reject the promise ("Promising, Intimate Relationships, and Conventionalism," 491).

but the consent giver, then in no case does it require the recipient's cooperation, acceptance, or uptake.

To understand the full scope of this weak uptake requirement, we would need to know *when* consent changes the relationship between the consent giver and the recipient. Whether a particular permission changes things for the pair will depend on a variety of factors, including which permissions the pair already have, how the particular permission would change things between the pair, their shared understanding of their existing relationship (if they have one) and of the meaning of the permission, and so on. To tell whether a relationship would be changed in any given case, we will have to rely on our understanding of details like these.[31]

Next, we will want to know what is meant by cooperation, acceptance, or uptake. What does the recipient need to do in order to complete an act of consent and gain the permission that is on offer?

Suppose *A* wants to consent to *B*'s ϕ-ing, and *B*'s gaining the permission to ϕ would constitute a change to their relationship. Consider first the following suggestion:

> *Uptake Is Knowledge of Offer*: *B* accepts *A*'s offer to permit *B*'s ϕ-ing just in case *B* recognizes that *A* is attempting to permit *B*'s ϕ-ing.

The purpose of the uptake requirement is to protect the recipient's interest in having a say in which relationships she has and with whom she has them. This rules out the proposal that uptake is knowledge of the offer. Suppose that *B* recognizes what *A* is doing but the permission is unwelcome to *B*. If uptake is mere knowledge of *A*'s offer, then *B* cannot prevent *A*'s consent from going through. So this notion of uptake is too weak to protect *B*'s interest in having a say in whether *A*'s consent goes through or not.[32]

[31] Thanks to an anonymous reviewer for raising this question. The reviewer also raises the following case, which probes the scope of the requirement: suppose a stranger on a plane offers their neighbor part of their snack. Is there a relationship shift here, and should the neighbor have a say in whether they gain the permission to eat part of the snack? I think so: as I am imagining the case, the relationship between the two strangers would change in a way that makes it okay (and not intrusive or inappropriate) to do various other things that strangers sometimes do on planes, such as engage in casual conversation about the destination and purpose of their travel. The neighbor may prefer to keep their distance, and so has an interest in not gaining the permission through the stranger's say-so. Note that this explanation relies on the cultural norms of plane travel; this is an example of the sort of information that I think we have to rely on to determine whether and how a particular act of consent changes a relationship.

[32] At the end of the previous section I raised the question of whether something other than an uptake requirement could protect the recipient's interest. An anonymous reviewer suggests the following: we could posit an additional normative power in the recipient that

Here is another suggestion:

> *Uptake Is Nonrejection*: *B* accepts *A*'s offer to permit *B*'s ϕ-ing just in case *B* recognizes that *A* is attempting to permit *B*'s ϕ-ing and *B* does not reject *A*'s offer.[33]

This would better protect *B*'s interest in having a say in whether *A*'s consent goes through. If *B* does not welcome the permission, she can reject it—provided that she has the ability and the opportunity to do so. Contrast this with the following slightly more demanding suggestion:

> *Uptake Is Communicated Willingness*: *B* accepts *A*'s offer to permit *B*'s ϕ-ing just in case *B* recognizes that *A* is attempting to permit *B*'s ϕ-ing and *B* communicates to *A* that *B* is willing to be permitted to ϕ.

This would equally protect *B*'s interest but also require that *B* communicate to *A*—verbally or otherwise—that *B* is willing to change her permissions and the relationship in the relevant way.[34] Why might this be important?

Suppose we are already committed to a behavioral view of consent. Our reasons for thinking that giving consent requires an expression in the consent giver's behavior may extend to the recipient's acceptance. For instance, suppose we believe that consent needs to be communicated because it alters third parties' reasons for action: prior to *A*'s consent, third parties may be justified in intervening (or even obligated to intervene) with *B*'s ϕ-ing. If *A* has consented to *B*'s ϕ-ing, then third parties are not justified in intervening. Unless *A*'s consent is publicly observable, third parties will not be able to reliably track their reasons for action. And unless acceptance is also publicly observable, third parties will not be able to reliably track whether *B* has the relevant permission, and so will not be able to reliably track their reasons for action.

enables them to reverse the effect of another person's consent and hand back an unwanted permission at any time. I think that the same considerations that rule out the "uptake is knowledge" proposal cause problems for this suggestion: consent would still go through without the recipient's participation and regardless of whether the recipient welcomes it. The power to reverse the effect of unilateral consent does not prevent the consent giver from giving consent and making changes to relationships unilaterally. In addition, the normative power of reversing someone else's consent would itself be a power that can alter relationships, and so would be in tension with the consent *giver's* interest in having a say in her relationships. Positing an uptake requirement on consent offers a much simpler solution to the problem at hand.

[33] Thanks to an anonymous reviewer for prompting me to consider this proposal.

[34] Sometimes acting as one has been permitted to act may be enough to communicate that the permission has been accepted.

Now, I do not intend to argue here that we should adopt a behavioral view of consent for this reason. I bring this up to illustrate that our view of what uptake is may depend on other commitments we have about consent or its functions. The interest that generates the need for an uptake requirement rules out the proposal that uptake is just recognizing that another is giving consent, but it alone does not decide between more robust notions of uptake.

3. Interlude: Consent and Control

I want to return briefly to the idea that consent functions to extend and expresses the consent giver's autonomy. I explained earlier how this function is used to motivate mental views of consent. I also explained that mental views are unilateral: if consent consists in a mental state, then no one but the consent giver needs to enter the picture.

I have argued against unilateral conceptions of consent by way of arguing for an uptake requirement for consent. You may wonder at this point whether the emerging bilateral conception of consent still retains a connection to the consent giver's autonomy, or whether an uptake requirement takes consent too far out of the consent giver's control. Tom Dougherty has raised a concern along these lines, writing:

> When we discussed the Mental View, we encountered the idea that consent enables a consent giver to exercise autonomous control over their normative boundaries. We also saw that if consent requires uptake with the consent-receiver, then the consent-giver is less able to exercise this autonomous control. Therefore, there is a tension between the ideal that the consent-receiver has control over their consent and the ideal that the consent-giver and the consent-receiver both know whether consent has been given.[35]

This seems correct, but I do not think we should be too worried about this tension. Focusing exclusively on the consent *giver*'s control over their normative boundaries obscures the fact that what those boundaries look like can matter a great deal to the consent *recipient*. In criticizing a behavioral account of consent, Alexander, Hurd, and Peter Westen—all defenders of mental views—write:

> Consent ... merely removes a moral (and sometimes legal) barrier. If it is not communicated, ... those to whom consent is given may not realize that those barriers are down and that they have permission to cross the

[35] Dougherty, *The Scope of Consent*, 60.

> consenter's moral (and legal) boundary. But so what? They have no duty to cross, only a permission to do so.[36]

Much of section 2.2 was dedicated to answering this rhetorical "So what?" Our duties matter to us, but so do our permissions. Unless we keep this in mind, it is easy to overlook the ways our autonomy, when we are the recipients of others' consent, is hampered if consent can be given unilaterally.

4. Objection: Revoking Consent (Unilaterally)

Before concluding, I want to consider an objection to uptake requirements for consent. The objection states that because consent can be revoked unilaterally, it should also be given unilaterally. Consider the following case:

> *Revocation*: Angie has moved to a new country and is making friends. In her home country, it is customary to linger after a dinner party while the host clears the dishes. In her new country of residence, clearing the dishes signals that the party is over and guests should leave. At a party at Betty's, Betty starts to clear the dishes. Angie thinks the party is still on and lingers for longer than Betty would like.[37]

By starting to clear the dishes, Betty tries to revoke her consent to Angie's presence in her house. If Angie's uptake is needed for Betty to revoke her consent, then Betty cannot do so unilaterally. But we do tend to think that consent can be revoked by the consent giver at any point, at their will, for any reason—especially in the context of sexual consent and other vulnerable or high-stakes interactions. Dougherty cites this as a reason to reject uptake requirements for consent, and writes:

> In so far as we have reason to expect that giving consent operates similarly to revoking consent, we have reason to reject the Uptake [requirement].[38]

Do we have reason to expect that giving consent operates like revoking consent? Dougherty does not provide any such reason, and proponents of uptake might take cases such as Revocation as evidence that revoking consent does *not* operate like giving consent. More importantly, I think that there is independent reason to think that revoking consent does not operate like giving consent: revocation has its own functions, and these functions are best served

[36] Alexander, Hurd, and Westen, "Consent Does Not Require Communication," 657.

[37] Cf. Dougherty, *The Scope of Consent*, 78.

[38] Dougherty, *The Scope of Consent*, 79.

if revocation can be done unilaterally. The power to revoke previously given consent is the power to reassert or reestablish a normative boundary and to thereby create distance between oneself and others. Unlike consent, which can be used to enable cooperation and intimacy, revocation serves primarily a *protective* function. We use this power to reassert our rights and our boundaries when a previously consented-to act becomes unwanted or unwelcome, or when a consent recipient's behavior becomes hostile or harmful. This function could hardly be served if revoking consent did require the recipient's cooperation. That said, this rationale for unilateral revocation does not speak against the weak uptake requirement I have argued for here—in fact, it does not threaten even a strong uptake requirement for all cases of consent.

5. Conclusion

I have argued that consent has a relationship-shaping function and that this function supports the following requirement for morally transformative consent:

> In any case where *A*'s consent to *B*'s ϕ-ing would change the relationship between *A* and *B*, in order for *A*'s consent to release *B* from an obligation not to ϕ, *B* must accept *A*'s attempt to consent.

I have discussed the worry that this may take consent too far out of a consent giver's hands and the objection that since revoking consent requires no uptake, neither does giving consent. What I have not done here is offer a complete account of what uptake is; this work will have to be done against the backdrop of a substantive view of what consent is.

The question of uptake has brought to light a distinction that does not yet exist in the philosophical literature on consent: the distinction between unilateral and bilateral views of consent. If my argument here is correct, then we ought to favor a bilateral conception of consent and reject conceptions of consent as a unilateral normative power.[39]

Harvard University
anniraety@fas.harvard.edu

[39] Thanks to Sally Haslanger, Kieran Setiya, Tamar Schapiro, Haley Schilling, Judith Jarvis Thomson, and Mallory Webber for discussion and feedback on earlier drafts of this work. Thanks also to an anonymous reviewer of *JESP* for their helpful comments.

References

Alexander, Larry. "The Moral Magic of Consent (II)." *Legal Theory* 2, no. 3 (September 1996): 165–74.

Alexander, Larry, Heidi Hurd, and Peter Westen. "Consent Does Not Require Communication: A Reply to Dougherty." *Law and Philosophy* 35, no. 6 (December 2016): 655–60.

Bolinger, Renée Jorgensen. "Moral Risk and Communicating Consent." *Philosophy and Public Affairs* 47, no. 2 (Spring 2019): 179–207.

Cornell, Nicolas. "Wrongs, Rights, and Third Parties." *Philosophy and Public Affairs* 43, no. 2 (March 2015): 109–43.

Darwall, Stephen. "Bipolar Obligation." In *Oxford Studies in Metaethics*, vol. 7, edited by Russ Shafer-Landau, 333–58. Oxford: Oxford University Press, 2012.

Dougherty, Tom. *The Scope of Consent*. Oxford: Oxford University Press, 2021.

———. "Yes Means Yes: Consent as Communication." *Philosophy and Public Affairs* 43, no. 3 (Summer 2015): 224–53.

Ferzan, Kimberly Kessler. "Consent, Culpability, and the Law of Rape." *Ohio State Journal of Criminal Law* 13 (2015): 397–439.

Hart, H. L. A. *Essays on Bentham: Jurisprudence and Political Philosophy*. Oxford: Oxford University Press, 1982.

Hohfeld, Wesley Newcomb. *Fundamental Legal Conceptions as Applied in Judicial Reasoning and Other Legal Essays,* edited by Walter Wheeler Cook. New Haven, CT: Yale University Press, 1919.

Hurd, Heidi M. "The Moral Magic of Consent" *Legal Theory* 2, no. 2 (June 1996): 121–46.

Ichikawa, Jonathan Jenkins. "Presupposition and Consent." *Feminist Philosophy Quarterly* 6, no. 4 (2020). https://doi.org/10.5206/fpq/2020.4.8302.

Kukla, Rebecca. "That's What She Said: The Language of Sexual Negotiation." *Ethics* 129, no. 1 (September 2018): 70–97.

Liberto, Hallie. "Promises and the Backward Reach of Uptake." *American Philosophical Quarterly* 55, no. 1 (January 2018): 15–26.

Martin, Adrienne M. "Personal Bonds: Directed Obligations without Rights." *Philosophy and Phenomenological Research* 102, no. 2 (January 2021): 65–86.

Pallikkathayil, Japa. "Consent to Sexual Interactions." *Politics, Philosophy and Economics* 19, no. 2 (May 2020): 107–27.

Shiffrin, Seana Valentine. "Promising, Intimate Relationships, and Conventionalism." *Philosophical Review* 117, no. 4 (October 2008): 481–524.

Thompson, Michael. "What Is It to Wrong Someone? A Puzzle about Justice." In *Reason and Value: Themes from the Moral Philosophy of Joseph Raz,* edited by R. Jay Wallace, Philip Pettit, Samuel Scheffler, and Michael Smith, 333–84. Oxford: Clarendon Press, 2004.

Thomson, Judith Jarvis. *The Realm of Rights*. Cambridge, MA: Harvard University Press, 1990.

Wertheimer, Alan. *Consent to Sexual Relations*. Cambridge: Cambridge University Press, 2003.

Journal of Ethics and Social Philosophy
vol. 27, no. 3 · may 2024

https://doi.org/10.26556/jesp.v27i3.2920

HOW TEMPTATION WORKS

John Schwenkler

I have to get this paper finished by the deadline. This means completing the next section before I have to teach at noon today. So that is what I decide to do. Then the morning unfolds, and noon rolls around—but my paper is only a few paragraphs longer. I have not followed through on my decision.

Let us ask: What can have happened between my deciding to work on my paper today, and my ending the morning with so little done, that would explain why I did not act as I said I would?

Several possibilities can be set aside as irrelevant to the topic of this paper. One is that I did not do my writing because I chose to do something else, like deal with a family emergency, that I *reasonably* found to be more important than the task I had decided on. Other possibilities are that there were occurrences outside my control, like the loss of electrical power in my office, that somehow *prevented* me from doing my work; or that at some point I simply *forgot*, either innocently or not, that I had meant to do this. A further possibility is that I did spend the whole morning working hard on my paper but came up short despite my best efforts. (Admittedly, the boundaries of this last phenomenon are vague, and it is something that we *claim* to have happened more often than it actually does.) Things like these do happen, and each has its own philosophical interest. But none of them will be my topic here.

The topic of this paper is rather the phenomenon of succumbing to the *temptation* to do something other than what one has decided to do. The argument I will make is that there is an especially devilish form of temptation, prevalent in human life, that philosophers who have written on this topic have tended to ignore or overlook. For these philosophers, to give in to temptation is always to *revise* a decision in a way that is somehow unreasonable—as when, say, recalling that there is a World Cup game that I can stream from my office, I abandon my plan to spend the morning writing. This construal of temptation fits the way it is depicted in the movies: the devil perches on my shoulder and tries to convince me to do what I know is wrong. In the present case, the devil might do this by praising the pleasures of watching soccer, while also reminding me of how far away my deadline is, how easily I can make up for missed time, and how many of the other authors are likely to be late with their submissions. In saying these things, the devil is trying to get me to *undo* my decision to work on my paper this morning, to *change my mind* about whether this is what I should do.

As many philosophers have recognized, what makes this kind of temptation both so pernicious and so philosophically interesting is the way it exploits what is often a perfectly rational process of reconsidering and revising our decisions. In the case where my work is disrupted by a family emergency, for example, it would be madness to insist that, given my plans, the emergency must take care of itself. This gives us the task of accounting for why just such a thought is so *un*reasonable in connection with the prospect of spending my morning in the office watching soccer. The challenge, in other words, is to explain the difference between reasonable *resoluteness* and unreasonable stubbornness or *inflexibility* in respect of the decisions we have made.[1]

Clearly, this is a common form of temptation, and we need to explain how we can resist it without irrationality. But I am going to argue in this paper that it is also possible to violate one's decisions, *without* ever taking those decisions back, by succumbing to a form of temptation that does not involve any inclination to change one's mind. And the case that I began with can easily be of this other sort. For even if I never take back the decision that I made to do my writing, I might still spend most of my morning doing things like formatting my bibliography, going out for coffee, staring at my bookshelf, and so on—but operating all the while under the notion that I am getting my writing done, or at least that I am going to finish it before I have to teach. When I succumb to temptation in this second way, it is not because I confront a choice between

[1] For related discussion of the so-called authority of one's decisions, see Arruda, "Sticking to it and Settling"; Bagnoli, "Hard Times"; Betzler, "Inverted Akrasia"; Bratman, *Intention, Plans, and Practical Reason*, "Temptation Revisited," "A Planning Agent's Self-Governance over Time," and "Acting Together with Oneself over Time"; den Hartogh, "Authority of Intention"; Ferrero, "Three Ways of Spilling Ink Tomorrow," "What Good Is a Diachronic Will?" "Decisions, Diachronic Autonomy, and the Division of Deliberative Labor," "Diachronic Constraints of Practical Rationality," "Diachronic Structural Rationality," and "Structures of Temporally Extended Agents"; Gauthier, "Assure and Threaten" and "Commitment and Choice"; Gold, "Putting Willpower into Decision Theory" and "Guard against Temptation"; Heeney, "Diachronic Agency and Practical Entitlement"; Hinchman, "Trust and Diachronic Agency," "Conspiracy, Commitment, and the Self," and "Narrative and the Stability of Intention"; Holton, *Willing, Wanting, Waiting*; Jaffro, "Weakness and the Memory of Resolutions"; McClennen, *Rationality and Dynamic Choice*; Morton, "Deliberating for Our Far Future Self"; Nefsky and Tenenbaum, "Extended Agency and the Problem of Diachronic Autonomy"; Paul, "Diachronic Incontinence Is a Problem in Moral Philosophy"; Raz, "Reasons for Action, Deliberation, and Norms"; Roth, "Agency and Time"; Rovane, *Bounds of Agency*, ch. 4; Smith, "Sovereign Agency"; Velleman, "Deciding How to Decide"; and Verbeek, "Rational Self-Commitment" and "On the Normativity of Intentions."

doing what I have decided and *doing something else instead*, and then resolve in favor of the latter. This second form of temptation is, therefore, different from the form that involves an unreasonable change of mind, and resisting it requires a different set of strategies. Or so I am going to argue in what follows.

To preview my argument, my central claim is that there is a distinctive form of temptation, which I call *temptation to violation*, in which a person is tempted to act contrary to a decision without undoing that decision or even calling it into question. This is possible, I argue, because the content of our decisions does not always settle exactly what is required to abide by them. This slack between the explicit content of our decisions and the specific acts by which we carry or fail to carry them out makes it possible for us to violate those decisions even as they remain in place. As such, temptation of this kind cannot be resisted simply by refraining from reconsidering our decisions or changing our minds about what to do.

Here is how my argument will proceed. Section 1 gives a general definition of temptation and then characterizes in more detail the two forms that I think it can take: the form that culminates in an unreasonable revision of a past decision and the form that culminates in a decision being violated without being taken back. Section 2 addresses a series of questions about this distinction. Section 3 explores recent work on temptation by Michael Bratman and Richard Holton, arguing that they both fail to recognize the possibility of temptation to violation and that this failure undermines their accounts of how temptation can be resisted. Section 4 diagnoses what I think is the source of this failure: that Bratman and Holton both focus only on decisions that determine *exactly* what must be done to act in accord with them, in contrast with ones that lack this kind of specificity. Finally, section 5 considers two puzzles that are generated by my argument, and section 6 discusses how temptation to violation can be resisted, arguing that this involves a crucial role for practical wisdom.

1. Two Forms of Temptation

Following Richard Holton, I understand succumbing to temptation as a way of manifesting weakness of will, where to be weak-willed is to be irresolute: it is to fail to persist in one's decisions, to be deflected too easily from the path one has chosen.[2] Temptation itself, then, is the mental process that culminates, if it does, in this kind of weakness or deflection, whereby a person does what is contrary to what she has decided.

[2] See Holton, *Willing, Wanting, Waiting*, 70. I quote this remark in full in section 3.

There is something normative in this definition: Holton says that we succumb to temptation when we are deflected from our chosen path *too easily*; and earlier I said that in succumbing to temptation, a person acts contrary to her past decision *without a good reason* for doing so. The point of this language is to set off, say, the case where I abandon my writing because I have to deal with a family emergency from the case where I abandon it in favor of watching the World Cup. Like Holton, I want to treat the latter cases as ones of succumbing to temptation and the former as a reason-responsive change in mind about what to do. One thing this means is that "temptation" as I am using it here is necessarily pejorative: it is an incitement to violate a past decision *unwarrantedly* and *unreasonably*.[3] In practice, of course, there is not always a bright line to be drawn between reasonably changing one's mind and unreasonably succumbing to temptation—not least because the person who does the latter kind of thing will often believe that she is being quite reasonable. But philosophers cannot draw brighter lines than the subject matter itself admits.

Here is what we have so far: a person succumbs to temptation when, without good reason, she does what is contrary to what she has decided. This characterization needs something more, for a person only succumbs to temptation, as opposed to acting merely foolishly or irresponsibly, if she violates her own decision *out of the desire to do what she knows to be contrary to it*. We need this condition to screen off the phenomenon of *involuntary* failure to act as one has decided to—as when, for example, I miss an appointment because I slept through my alarm or fail to stay sober because I did not know that the punch at a party was spiked.[4] (If the punch was secretly spiked and I drank it because I thought it looked tasty, then I acted out of the desire to do what was *in fact* contrary to my decision, but not what I *knew* to be contrary to it.) By contrast,

[3] Admittedly, this appeal to instrumental normativity might not be enough to do this concept justice, as shown by the following example. A person who has never thought one way or the other about stealing decides on a whim to steal a bottle from the liquor store. Just as she is about to hide the bottle in her bag, she sees a police officer walk into the store—and though there is no reason to think the officer will notice or apprehend her, the would-be thief gets cold feet and puts the bottle back. According to the definition I have given, putting the bottle back is a way of succumbing to temptation—while taking it from the store would not have been. And neither verdict is intuitive. Since, however, it would take another paper to work out what this reveals about the kind of normativity internal to the concept of temptation, for now I will employ the framework that has become standard in the literature. (I thank Timo-Peter Ertz and Anselm Müller for presenting me with this case, and Robert Audi for also raising an objection along these lines. For steps toward an account of temptation that might be able to resist the worry, see Blackburn, *Mirror, Mirror*, ch. 7.)

[4] I thank Marshall Bierson for prompting this clarification and suggesting the last two examples.

when I spend the morning in my office watching a soccer game, scrolling social media, or going out for coffee, it is out of my desire to do these other things—or, perhaps, out of a desire simply not to do my work—that I choose to act as I do, and so do not complete the work that I had planned. This idea will be important to my argument as it unfolds, and I will consider it in more detail just below.

With this in the background, let us look more closely at the low-grade drama in my office. I said there are two ways I could be tempted not to do what I decided to do—namely, complete the next section of my paper before I go to teach. One of these is the Hollywood way: the devil perches on my shoulder and preaches in praise of the other things I could do and of the relative unimportance of my work—and in light of this temptation, I *revise* the choice that I made this morning, thereby abandoning the decision to do my work.

But the other form of temptation, the one that I claim has been neglected by philosophers, works differently than this. Instead of attempting to change my mind outright, the devil works in a subtler manner by whispering persuasive-sounding justifications that often involve words like "only" and "just." *It's only a short break. It's just a way to clear your head. It's something that's got to be done eventually anyway.* In saying these things, the devil is trying to get me not to do my writing—but *not* by trying to undermine the decision that I made to do it. And so my morning unfolds: a bit after 9:00, I get to my office, stare out the window for a while, answer a few emails, quickly check social media, and then go out to get a cup of coffee. (Now it is about 9:20.) Back in my office, I read the first chapter of that book I had been waiting for and then use this as inspiration to bang out a couple of rough paragraphs that will need to be revised before I can go on. I go to the bathroom, then stare for a few minutes at my screen. (10:00.) The first new paragraph I revise to my satisfaction, but the second one is hopeless and has to be deleted. (10:20.) I stare at my bookshelf and think. I dig a bit further into the relevant literature, then go to get advice from a colleague who is more of an expert than I am. (11:05.) This leads to my writing a lengthy footnote full of citations that need to be added to my bibliography, which I then spend a few minutes reformatting. I stare out the window, have a snack, answer two emails, and check Twitter. Now the jig is up: my paper is only a paragraph and a footnote longer, and the start of my class is about twenty minutes away. It is in this way that I end up failing to do what I decided I would—where the failure is of my own choosing, but not because I have abandoned the decision to do my work.[5]

[5] This last phrase echoes G. E. M. Anscombe's description of Saint Peter's denial of Christ, in the closing pages of *Intention*, 93–94. For discussion of this passage, see my *Anscombe's Intention*, 207–10.

Crucially, in order for this case to be one of succumbing to temptation according to my working definition, it needs to be that it is the *desire* not to do my writing, or to do something else instead, that explains why I spend the morning as I do. And there are possible versions of my morning that do not have this character—say, if I spent several hours grading papers, which I abhor doing, out of a misplaced belief that this needed to be done right away. If this were what had happened, then the charge that I succumbed to temptation would seem not to stick—I could be worthy of criticism for failing to write but not for having given in to the *temptation* not to do so. However, in the version of my case that I think we will find more familiar, it is indeed because I give into temptation that I fail to get my writing done, though not necessarily because I change my mind and decide to do something else instead. In such a case, the desire to do things other than write—and also, perhaps, the simple desire *not to write* at all—will be the very thing that leads me to spend the morning in the way that I have described and so not to get done the writing I had planned. My claim, however, is that this need not involve any *decision* on my part that I will not do my writing after all.

For the sake of brevity, in what follows I will refer to the first of these forms of temptation, in which I am tempted to revise my decision and do something else instead, as "temptation to indecision," while the second, in which I am tempted to act contrary to my decision but without revising it, I will refer to as "temptation to violation." Neither label is perfect, but I hope they will work to elicit the corresponding notions. The next section will address several questions about this distinction.

2. Some Questions about This Distinction

1. Is the difference between these forms of temptation just that temptation to violation is always a temptation to *procrastinate*, or to delay the start of an activity one has decided to carry out?

If this were the case, then it would mean that I have not really identified a neglected phenomenon, as the topic of procrastination has received a great deal of fruitful philosophical attention.[6] Fortunately, though, the temptation to procrastinate is not always a temptation to violation, nor does this kind of temptation always involve putting off the start of a task. For example, suppose I have decided to get started on my paper as soon as I get to my office this morning, and when I arrive, I notice a book that has just been delivered by the library. In this case, I could start to think, unreasonably and out of the desire

[6] For a start, see the essays collected in Andreou and White, *Thief of Time*.

not to write, *either* that reading the book will be a good way of getting to work on my paper *or* that it is not that important to start on my writing right away, and therefore it can wait until after I have done some reading. If these thoughts are unreasonable, then both are temptations to procrastinate, but while the first takes the form of a temptation to violation, the second is a temptation to revise my decision and choose to do something else instead.

Likewise, succumbing to temptation to violation does not always involve putting off the start of a planned course of action. For example, even if I open up my document immediately when I get to my office, my subsequent "writing" might be mostly a matter of sipping coffee, fiddling over word choice, and staring at my bookshelf, none of which leads to my getting much done. If these choices are unreasonable, and if I made them out of the desire not to write, then in making them, I will have succumbed to temptation to violation—but not because I ever put off *starting* to do the thing I had decided I would do.

2. Is temptation to violation anything more than temptation to *akrasia*, or to action that is contrary to one's own best judgment of how to act?

Once again, if this were the correct account of temptation to violation, then it would undermine my claim to have identified a neglected phenomenon, as philosophers have written a great deal about akratic action.[7] But while there is something right in saying that a person who violates her own decision has thereby acted against her own best judgment, the phenomenon I am trying to highlight is quite different from *akrasia* as the latter phenomenon is usually understood. On the common understanding, a person who acts akratically does so while believing that *this* thing—that is, the very thing that she is doing, such as checking social media or watching a soccer match from her office computer—is something that she should not do.[8] By contrast, in succumbing to temptation to violation, we usually do *not* understand that we are thereby doing anything wrong or even that we are being irresolute. We saw this in my office drama: in getting coffee, going to the bathroom, staring at the bookshelf, and so on, I act under the belief that I am doing what is totally appropriate, at least as regards the decision to get my writing done. (If, instead, I got absorbed in reading professional gossip that I know I should ignore, then *that* might fit the

[7] Again, see for a start the essays collected in Stroud and Tappolet, *Weakness of Will and Practical Irrationality*.

[8] For example, according to Donald Davidson, a person acts incontinently (that is, akratically) in doing *x* "if and only if: (*a*) the agent does *x* intentionally; (*b*) the agent believes there is an alternative action *y* open to him; and (*c*) the agent judges that, all things considered, it would be better to do *y* than to do *x*" ("How Is Weakness of the Will Possible?" 22).

standard definition of *akrasia*.) Even if "at some level" I know that I am spending my time unwisely, my considered *judgment* may be that everything I do is entirely justifiable. I act in a way that is contrary to my own standing decision, but not by doing something that I judge I should not do.[9]

3. Does the distinction come down to whether the decision that is violated has a prescriptive character or a proscriptive one—so that temptation to indecision is always the temptation to revise a "shalt not," while temptation to violation always concerns a "shalt"?

I do not believe it does. For one thing, prescriptive decisions are clearly subject to temptation to indecision, as when I consider quitting my plan to write this morning because I prefer to watch soccer instead. Further, and as I will discuss in more detail below, there are lots of proscriptive or "shalt not" decisions that it seems possible to violate without revising. For example, someone who has decided to stop yelling at the children might justify his yelling on a given occasion by saying that really he is only raising his voice. Someone who has decided to stop checking social media during the workday might entertain the thought that it "doesn't count" if he does it while having a cup of coffee. Someone who has decided to refrain from drinking on weekday evenings might tell himself that not only does he "have" to calm his nerves this evening given how awful the children have been, but also that he isn't "really" drinking after all if he only has a small glass of wine (or two). And so on. (Enough with the autobiography, really.) All these are instances of temptations to violation, and each is in relation to the decision *not* to do a certain kind of thing.

4. Is the "violator" always *self-deceived* about her own intentions, professing to have a standing decision to do something when, in fact, she has already taken that decision back, if indeed she ever made it at all?

This is definitely a possible reading of my office drama. Maybe I would like to *think* that I have made the decision to work on my paper this morning and have not changed my mind about whether to do this, but in fact, this is only a story that I tell myself, and the reality is that I have decided to fritter away my day.[10] If this kind of diagnosis were correct in every case, it would undermine the description that I have given of what temptation to violation involves. But I do not believe this can be so.

[9] See, however, the discussion of "extended *akrasia*" in Tenenbaum, *Rational Powers in Action*, 191–92, for an account according to which this course of action comes out as akratic. I draw significantly on Tenenbaum's analysis in sections 4 and 6 below.

[10] I thank Mario Attie, Paul Blaschko, and Mike Rea for raising different versions of this objection.

One reason for this is that the pattern of behavior on display in my office drama could easily be the result, not of my having abandoned or never truly made the decision to write, but rather of my simply not *wanting* to act as I really have decided to, or of my more strongly wanting to do something else—just as, in a corresponding case of temptation to indecision, what explains why I fail to work on my paper is simply that I have more of a desire to watch soccer than to do my work and not that I never decided to do the latter thing at all. That is to say, if a person who has made a certain decision can *revise* that decision in light of contrary desires, then it seems possible also to *violate* that decision in the same way.

Second, while I grant that *sometimes* I might, for example, fritter away the morning in my office because I have not really decided, or have quietly taken back my decision, to work on my paper before I teach, in a given case there may be many things we can point to which would suggest the contrary—for example, that over breakfast I outlined the writing I was going to do; that when I got to my office I took some specific steps, such as canceling appointments and closing my office door, in order to limit distractions; that on several occasions I caught myself wasting time and made a concerted effort to get back to work; that most of the day was spent thinking about the topic of my paper with my document open on my laptop; and that when noon rolled around, I despaired at how little I had gotten done. In general, a person who does these things is a person who intends to get their morning writing done. In such a case, what explains why I do not end up doing this is not that I failed to persist in my decision but rather that I succumbed to temptation nevertheless.

5. Will any concrete case of succumbing to temptation usually involve a *mix* of these two forms rather than consisting wholly of one or the other?

Yes. In my office drama, for example, it is likely that I will have supplemented my general decision to work on my paper this morning with the further decision to employ some more specific measures, such as keeping my office door closed and not checking my email too frequently, in order to keep me out of tempting situations. And very often, if I fail to complete my writing, it will be because I failed to do some of these other things too. Further, this latter failure will often involve succumbing to temptation to indecision—such as when I tell myself that, contrary to what I decided this morning, it is okay to spend *some* time on social media as long as I have been making good progress.

One interesting question that this raises, which I will discuss in detail in section 6, is that of how to understand the relation between specific decisions like “do not check my email this morning” and general ones like “finish this section of my paper before noon” in cases where I adopt the former as a means of carrying out the latter. I will argue in that section that the achievement of our wider ends cannot always be reduced to the execution of narrowly defined policies. But the thing to see for now is that even if a specific case of succumbing to temptation does involve some unreasonable revision of a person's past decisions, it does not follow that the work of temptation will consist entirely of that. When I go back on my decision not to check my email, for example, this does not mean I have changed my mind about whether to do the writing I had planned. And that is because it is not strictly *necessary* that I eliminate all distractions if I am to get my writing finished—for just as I can get coffee, or stare out the window a bit, compatibly with or even as a means to writing productively, so it may be with spending a few minutes reading emails. As such, even if I do revise these specific decisions, the decision to get my writing done may nevertheless remain in place—though not, of course, in a way that provides any guarantee that I will end up doing as I said.

3. TWO INADEQUATE ACCOUNTS

Earlier, I claimed that the kind of temptation that is the focus of this paper—what I called *temptation to violation*, or the temptation to violate one's decisions without revising them—has been overlooked in recent philosophical discussions of temptation. Now I will substantiate this charge by exploring how temptation is construed in influential work by Michael Bratman and Richard Holton. In addition, I will show how the accounts that Bratman and Holton give of how a person can *resist* temptation, and of how this resistance can be instrumentally rational, fail to get traction in reference to temptations of this other kind.

3.1. Bratman

Let us begin with Bratman, whose analysis of temptation centers on cases like the following:

> Suppose I am a pianist who plays nightly at a club. Each night before my performance, I eat dinner with a friend, one who fancies good wines. Each night my friend offers me a fine wine with dinner, and—as I also love good wine—each night I am tempted to drink it. But I know that when I drink alcohol, my piano playing afterward suffers. And when I reflect in a calm moment, it is clear to me that superior piano playing in my evening performance is more important to me than the pleasures of wine with dinner. Indeed, each morning I reflect on the coming challenges of the day and have a clear preference for my turning down the wine. Yet early each evening when I am at dinner with my friend, I find

> myself inclined in the direction of the wine. If I were to go ahead and drink the wine, mine would be a case of giving into temptation.[11]

Bratman's example is a clear case of temptation to indecision. He begins the evening with a certain plan, then is tempted by the possibility of doing what that plan rules out. As Bratman presents the case, succumbing to this temptation would mean reconsidering and then revising his plan of refraining from drinking wine before his gig. By contrast, Bratman will *resist* temptation effectively if he refrains from revising this plan and so keeps his decision in place. And neither of these characterizations applies to the phenomenon of temptation to violation—first, because succumbing to such a temptation does not involve a revision of a prior decision, and second, because the action one is tempted to perform is not seen as incompatible with one's standing plans.

Is there a way, though, for Bratman to be tempted to violation in the situation he presents? Speaking for myself, the operative thoughts are all too familiar: *I'll order it just to be polite—I'll only have a sip or two—it's very low in alcohol anyway—I'll follow it up with a cup of coffee—and we're eating earlier than usual tonight, so I don't have to play for several hours*. Later on, I will allow that if Bratman's plan is *so* specific that it rules out any of these ways of getting around it, then it is a special case of a decision that cannot be violated without being taken back. What will matter, though, is to see that it *is* a special case—so if Bratman's decision were, by contrast, not to drink *so much* that it will interfere with his piano playing, then it would be easy to succumb to the temptation to do this without giving up the decision not to. (Section 5 will present a case of just this kind.) Further, as I will discuss in detail below, many of the decisions that relate us to relatively indeterminate ends or govern the structure of long stretches of our lives are such that they cannot be construed so narrowly.

This limitation in Bratman's understanding of temptation leads to a corresponding limitation in his account of how it can be resisted—an account that is, as he puts it, one of "mechanisms and strategies of reconsideration that sometimes block reconsideration of a prior intention in the face of merely temporary

[11] Bratman, "Planning and Temptation," 37–38. The basic structure of the case is a template for Bratman's later work on this topic. In "Toxin, Temptation, and the Stability of Intention," 74–77, Bratman, the pianist, is replaced with Ann, who is tempted to have a second beer that will interfere with her evening book reading. In "Temptation Revisited," 257–59, and "Temptation and the Agent's Standpoint," 154–56, the temptation is to have a second glass of wine with dinner even though this will interfere with your after-dinner work. And in "Rational Planning Agency," 217, Bratman considers the case of someone who resolves to have just one beer at a party while knowing that later on she will think it better to have many beers.

preference change."[12] For Bratman, the central thing that allows us to resist temptation rationally and effectively is the anticipation of the *regret* that we will feel later on if we revise our plans in the face of a tempting alternative to them.[13] This is explicit in the case above: the pianist's preference for a glass of wine is supposed to be *temporary*, since when it comes time for his gig, he will either wish that he had not had the glass (if he did) or be glad that he refrained (if instead he resisted the temptation). Bratman supposes, then, that a person who is being tempted can look forward to how she will feel later on about the choice she is tempted to make right now, and treat the prospect of her future regret as a reason not to reconsider. And even if we were to grant to Bratman that this strategy can do the trick in the kind of case that is his focus, it does not even get off the ground in the different kind of case that is mine.[14] Returning once more to the temptation that I face in my office, it is only insofar as I *recognize* how the tempting possibilities might keep me from doing my writing that I can anticipate how disappointed I will come to feel if I choose them, and use that as a reason to buckle down. As it is, when I choose to do the tempting things, it is never with the understanding that this will mean failing to do what I said I would. The anticipation of my future disappointment cannot motivate me to resist temptation, since I do not anticipate being disappointed at all.

3.2. Holton

A similar picture of temptation is laid out in Richard Holton's detailed treatment of this topic in *Willing, Wanting, Waiting*. Central to Holton's account is the idea that temptation often works by *corrupting* a person's judgment rather than *overcoming* her better judgment to the contrary. This makes Holton's notion of weakness of will, which is the focus of his discussion of temptation, different from the philosophical notion of *akrasia*. As I explained earlier, on standard

[12] Bratman, "Planning and Temptation," 53.

[13] Here is a characteristic formulation concerning the temptation to have a second glass of wine: "I know that this judgment shift will be temporary: at the end of the day I will stably revert to my judgment that what would have been best at dinnertime would have been to stop with a single glass of wine" (Bratman, "Temptation and the Agent's Standpoint," 154).

[14] It seems clear to me that we should not grant Bratman this much. Speaking from experience, often a person in the throes of temptation will be quite confident that the tempting choice will end up making her very happy—and sometimes she will be right! Related problems with Bratman's account are discussed in Andreou, "The Good, the Bad, and the Trivial"; and Holton, *Willing, Wanting, Waiting*, 156–60; and for further discussion, see Andreou, "General Assessments and Attractive Exceptions"; Bratman, "Planning, Time, and Self-Governance"; Gold, "Guard against Temptation"; Greene and Sullivan, "Against Time Bias"; Hinchman, "Narrative and the Stability of Intention"; and Tenenbaum, "On Self-Governance over Time."

accounts, a person acts akratically when she chooses to do what conflicts with her own best judgment. By contrast, on Holton's account, the person who succumbs to weakness of will is led by temptation to *revise* that judgment in an unreasonable way—paradigmatically, in the kind of case Holton considers in detail, by a psychic mechanism that leads our subjective valuations to tend to conform to what we expect ourselves to do.[15] Anticipating, for example, that I am likely to have a second glass of wine, I am led to judge having the glass to be worthwhile, since otherwise I would have to regard my own choice as stupid.[16]

As I will discuss in detail below, there are elements of this account that apply in turn to the phenomenon of temptation to violation, as the "corruption of judgment" can impair our thinking about which courses of action are compatible with doing what we have decided. But Holton himself does not consider this quite different form that temptation can take. Beginning from the idea I endorsed earlier—that "weak-willed people are irresolute; they don't persist in their intentions; they are too easily deflected from the path they have chosen"—which describes temptation to violation no less than temptation to indecision, Holton goes on to say that "Weakness of will arises . . . when agents are too ready to reconsider their intentions."[17] This latter phrase is a perfect description of temptation to indecision. And if, as I have argued, it is possible to succumb to temptation, and thus to be irresolute, *without* reconsidering or revising the intentions that we thereby fail to persist in, then Holton's definition draws the boundaries of temptation too narrowly.

As with Bratman, Holton's exclusive focus on the phenomenon of temptation to indecision leads to a corresponding limitation in his account of how temptation can be resisted. For Holton, the key to resisting temptation lies in forming *resolutions*, which he understands as "a specific type of intention that is designed to stand firm in the face of future contrary inclinations or beliefs."[18] The way that resolutions help us resist temptation is through the capacity to refrain from *reconsidering* the choices that they concern. Recognizing, for example, that from the warmth of my bed, I will fail to see the importance of going for an early morning run, the night before I go to bed I may form the intention,

[15] For this discussion, see Holton, *Willing, Wanting, Waiting*, 97–103.

[16] And likewise, I am led to judge that I will continue to think the same thing in the future. (This is relevant to the criticism of Bratman in note 14 above.)

[17] Holton, *Willing, Wanting, Waiting*, 70–71.

[18] Holton, *Willing, Wanting, Waiting*, 10. And again: "At the most intellectual level, resolutions can be seen as involving both an intention to engage in a certain action, and a further intention not to let that intention be deflected. . . . So, when I resolve to give up smoking, I form an intention to give up, and along with it I form a second-order intention not to let that intention be deflected" (*Willing, Wanting, Waiting*, 11).

not only to run when I get up, but also not to reopen the question of whether to do this.[19] That last step is important because, as we have seen, if I were to reconsider this question, then my ensuing judgment would likely be corrupted, leading me to judge it better to skip the run and remain in my warm bed. For Holton, then, "the effort involved in employing willpower is the effort involved in refusing to reconsider one's resolutions."[20]

There are, again, questions that can be raised about the adequacy of Holton's account as a description of how to resist temptation to indecision.[21] But even if the account were adequate on that score, it would be no account at all of how to resist the temptation to act contrary to our decisions *without* revising or even reconsidering them. When I give in to the temptation to fritter away the day, it is not because I reconsider the decision to do my work and decide it will be better to spend the day doing other things, thereby revising the decision to get my work done. Instead, that decision remains in place even as I succumb to the temptation to violate it. If there is a way to resist this kind of temptation, it is not by refusing to reconsider our decisions.

4. WHY THE ACCOUNTS FAIL

If not sheer oversight, then what accounts for the fact that philosophers like Bratman and Holton have failed to recognize the possibility of succumbing to temptation without reconsidering or revising the decision that one violates? The answer I will give is that it is because they have failed to recognize how the content of our decisions often does not specify exactly what we have to do, and refrain from doing, in order to follow through on them. It is, I will argue, the slack that exists between the content of our decisions and the specific acts by which we need to carry them out that makes for the possibility of violating our decisions without changing our minds about what to do.

To bring this out, let us first look more closely at Holton's case of the would-be morning runner:

> Homer has not been getting much exercise, and it is starting to show. He judges, and desires, that he should do something more active. He resolves to go for a daily run, starting next Saturday morning. But as his alarm goes off early on Saturday, his thoughts start to change. He is

[19] For this last case, see Holton, *Willing, Wanting, Waiting*, 138–40.

[20] Holton, *Willing, Wanting, Waiting*, 121.

[21] For some of them, see Bratman, "Temptation and the Agent's Standpoint"; Ferrero, "Diachronic Constraints of Practical Rationality"; and Paul, review of *Willing, Wanting, Waiting*.

> feeling particularly comfortable in bed, and the previous week had been very draining. He could start his running next weekend. And does he really want to be an early-morning runner at all? That was a decision made in the abstract, without the realization, which now presents itself so vividly, of what such a commitment would really involve.[22]

Holton uses this case to bring out the importance of a phenomenon he calls *rational non-reconsideration,* in which a person's resolution not to reconsider a decision makes it rational for them to persist in that decision even though, were they to reconsider it on a given occasion, they would rationally choose to revise it. (Rationally, since doing so would be in accordance with what would then be the person's best judgment.) In Homer's case, what makes it rational for him to run on a given morning is precisely the way that he does not reconsider his standing decision to do so; instead, Homer "springs out of bed . . . , brushing aside his desire to stay in bed, and any nagging thoughts about the worth of exercise, with the simple thought that he has resolved to run, and so that is what he is going to do."[23]

There are two things to say about Holton's presentation of this case. The first is that Homer's decision to go for a daily run cannot be a decision to do so *no matter what*—even if his ankle is injured, or he is very sick, or it is blowing wind and rain or snow outside, or he has been up all night tending to sick children, or he would need to start his run at 4:00 AM. because he has an early flight to catch. And because it is impossible to enumerate in advance all of the circumstances in which Homer would reasonably decide against running on a given day—that is to say, would decide this reasonably not from the warmth of his bed but rather from an appropriately impartial perspective—it would be madness for Homer to refrain *without exception* from reconsidering this decision when he wakes up. Instead, Homer's policy of not reconsidering his decision to go for a run has got to be somewhat flexible. This means, however, that there will always be some room for Homer to be tempted to indecision. *I've got a cold—I'm exhausted—the weather is awful—I'm sore from the hard workout I did yesterday—It's fine to skip today's run if I then double up tomorrow.* Sometimes, thoughts like these will be mere temptations. On other occasions, though, they will not be. Unfortunately, from the warmth of Homer's bed, it is not always easy to say which is which. For this reason, the resolution not to reconsider cannot make Homer invulnerable to the temptation to revise his decision.

Second, and more importantly for our purposes, the decision *to go for a run* every day includes a similar kind of flexibility that exposes it to the possibility of

[22] Holton, *Willing, Wanting, Waiting,* 138.

[23] Holton, *Willing, Wanting, Waiting,* 139.

temptation to violation. When Homer decides to go for a daily run, clearly he does not mean that each day he will do a lap around the living room or shuffle from the front door to the sidewalk and back. (A *run* must involve more than that.) But what if, on a given day, Homer finds that he only has the time, or the physical capacity, for an easy twenty-minute jog instead of the usual five miler? If that is okay—if sometimes that counts as "going for a run," depending on the circumstances—then how about running two or three times around the block? Or, again, if *sometimes* a twenty-minute jog is enough, then what if he did this for ten days straight? It seems impossible to rule such things out in advance. Yet as long as Homer's decision leaves room to consider such a possibility, it also leaves him vulnerable to temptation to violation—to choosing courses of action that he *represents* as belonging to the appropriately flexible articulation of his standing decision, but are actually quite incompatible with it.

The same lesson comes out in my office drama, though in that case, the room for slippage is even more obvious. This is because even the relatively specific decision that I made—that is, the decision to finish the next section of my paper by noon—could be executed in an enormous range of ways. I could, of course, arrive at my office first thing in the morning and not move from my desk, check my phone, or navigate away from my document until the morning's writing is complete—but as I will discuss below, this is not necessarily the best strategy for getting my work done. In any event, another *possible* way to finish the section involves doing quite a lot of the things that I actually did—things like reading a chapter from a relevant book (or even one that is not so relevant), getting a cup of coffee, staring at my bookshelf, clicking occasionally over to my email, and so on. And while clearly I should not have done all of this *so* much, at least not without getting much more done during other stretches of time, many of these things I did, considered in themselves, were still quite compatible with—or even conducive to!—the goal of completing my work. Yet all of this is exactly what made it possible for me to justify doing all the things that I did and to regard them as compatible with the decision to get my writing done. It is precisely in this way that I managed to *choose* to do what was contrary to that decision, without ever having to change my mind about it.

If this diagnosis is correct, then temptation to violation is similar to temptation to indecision in that both of them trade on a distinctive feature of our nature as finite and time-bound agents, but the features at work in each case are importantly distinct. As the case of Homer brings out, our vulnerability to temptation to indecision exploits the fact that we sometimes do have good reason to revise our decisions in light of changing circumstances or facts about our situation that we could not account for in our initial decision-making. Similarly, my suggestion now is that part of what makes us vulnerable to the temptation to act contrary to our decisions *without* revising them is the fact that these decisions often have the character I have just identified: they fail to determine in advance all of the things that one must do, or refrain from doing, in order to act in accordance with them—which means that we may fail to see how a given course of action is a violation of our own decisions.[24]

The final section of this paper will consider whether it is possible to close ourselves off to this vulnerability by adopting decisions whose content is more specific. Before that, I want to address a pair of further puzzles that are raised by this argument.

5. Two Puzzles

Suppose Homer decides, unreasonably and out of the desire not to run, that this morning he will just jog a couple of times around the block. While a wide range of activities could be enough to count as "going for a run" on a given day, on this particular day jogging twice around the block clearly does not. It seems right to say that, in deciding that this is what he will do, Homer decides *thereby* to act contrary to his decision to run that day.

The first puzzle I want to raise concerns how the case of my office drama seems to lack this simple structure, as in that case there is no *discrete* decision or action, or moment or series of moments of inaction or indecision, in which we can say my violation lies.[25] If this seems hard to swallow, consider first the stretches of the morning when a person who looked in my window might have said I was not writing, perhaps because I was in the bathroom or out to get coffee. Could these be singled out as the times when I violated my decision to get my writing done? Of course not—for some of the things that I was doing at those times may have been compatible with or even conducive to doing my work; and further, many of the things that I did while I *was* "writing," such as tinkering with my phrasing and adding entries to my bibliography, may have done as much as anything else to contribute to my eventual failure. Alternatively, consider the situation when 10:45 rolled around, and I was sitting with my colleague discussing the twists and turns of the secondary literature, despite having written only a paragraph to that point. While at this point my failure to buckle down and "really" get to writing might have been less forgivable than when at 9:15 I was sipping my coffee and reading a chapter from that book, this

[24] Here, I have learned a lot from the discussion of indeterminate ends in Tenenbaum, *Rational Powers in Action*, ch. 4. See also Andreou, "The Good, the Bad, and the Trivial" and "Temptation, Resolutions, and Regret"; and Tenenbaum and Raffman, "Vague Projects and the Puzzle of the Self-Torturer."

[25] I am very grateful to Nathan Helms for some spirited pushback against my argument here.

should not lead us to say that it is only at the later time that I acted contrary to my decision. For after all, it is only *because* of the way I had spent my time earlier—spent it, I would say, giving into temptations not to write—that I had so little leeway later to call on my colleague's expertise. Each stretch of my day takes on its character only in light of how I spend the others. And it is for this reason that we cannot locate where the moment of my violation lies.

This may seem surprising. Should it be? For one thing, this phenomenon is not limited to the violation of our decisions by commission rather than omission. Imagine, for example, that you are out on the town with your friends, and in light of what happened last weekend, you have decided not to drink too much this evening. Okay, then—having one drink is definitely not having too much. Nor is having a second. A third? Well, you are only going to sip it. At some point, you will be drinking to excess, despite never having taken back the decision not to. But is it *only then* that you do what is contrary to this decision? The problem with thinking so is not just that the "point" is really more of a region. It is, rather, that it keeps us from seeing how you approached the *entire* evening in the wrong way. Yes, you definitely should not have had that last drink—but nor should you have had the ones leading up to it, at least not without a better mechanism for cutting yourself off. It is, however, precisely the way that those earlier drinks were not *in themselves* violations of your decision not to drink to excess that made it possible for you to justify having them, and so to get yourself in a place where you drank as much as you did.

Further, this impossibility of pinpointing just where things go wrong (or right) pertains quite generally to a range of important virtue and vice descriptions. For while we can sometimes identify specific acts as ones of, say, justice or courage or intemperance, describing a stretch of a person's *life* with one of these words is not a matter of pointing to the various just, courageous, or intemperate acts they performed, nor of summing these up and considering the ratio between them. Rather, characterizing someone's life in terms like these is always a matter of seeing their particular deeds as instances of wider patterns.[26] And this is what explains how I could fritter my morning away. It is just insofar as I suppose that, in going out for coffee, staring out the window, tinkering with my wording, and so on, all of this belongs to a wider pattern that will culminate in the completion of my work, that I manage to violate this decision without ever taking it back.

But this raises a further puzzle. Earlier, I said that in order to succumb to temptation, as opposed to acting merely foolishly or irresponsibly, a person must violate her own decision *out of the desire to do what she knows to be contrary*

[26] Here I have learned a lot from Müller, "Acting Well."

to it. How, though, can this be true of my office drama as I have just described it or of the case when you are out on the town with your friends? When I read from that new book and then went to get a cup of coffee, I thought that I was thereby making progress in my work. Likewise for the time I spent revising my rough paragraphs, diving into the secondary literature, and talking about my work with my colleague. Likewise, even, for the bit of time that I spent scrolling Twitter ("just to give myself a break"). It is central to my description of the case that, as I was doing these things, it was always under the notion that I was getting my writing done.[27] There is no doubt that I desired to do each of the things that I did—but how can I then have been doing what I *knew* to be contrary to my standing decision, especially given that I *thought* I was acting in accord with it?

The answer to this question seems to turn on two things. The first is that the knowledge that is highlighted by this condition is partly a matter of self-knowledge—not just knowledge of how certain things are in the world, but knowledge of what *I myself* am up to. These two kinds of knowledge are related, of course—for example, without knowing that the punch in this bowl is spiked, I cannot know that I am drinking alcohol when I consume it. However, the cases under consideration do not turn on such purely factual ignorance. When you are out with your friends, perhaps you have lost track of just how many drinks you have had—but you *do* know that you have been sipping drinks all night without keeping count and without a clear plan to cut yourself off. Likewise, in my day at the office, I may have lost track of the time or of how long I have spent fiddling with word choice and staring out the window—but I *do* know that I have been taking a fairly relaxed approach to my work today, and I am under no illusion that the section I resolved to work on is just about complete. This makes these cases totally different from the one where I accidentally drink spiked punch. Each of us knows what we are up to, and it is no surprise to us that this is not a way of acting as we said we would. In the throes of temptation, however, such a thing can be difficult to appreciate.

The other thing we need to reflect on is the nature of the "thought" by means of which a person tempted to violation will tend to conceive of herself as following through on her decision. Holton's notion of corruption of judgment is helpful here: in the throes of a powerful desire, not only do my choices tend to conform to what I want, but so does the way that I think about what I am up to. This seems to happen in two ways.[28] First, it happens through the *avoidance*

[27] In the same way, during your night on the town, *each* drink is consumed in the belief that it is not too much.

[28] I thank Anselm Müller for helping me to see this.

of thoughts that would be ways of recognizing what has really been going on: so if my desire is to spend my morning doing things other than writing, then likely I will not keep close track of the time. And second, it happens through the *cultivation* of thoughts that provide justifications for going on as one prefers to. *It's just a short break*, I said to myself. *This paragraph needs revision. I need the coffee to clear my head*. One reason why temptation often seems genuinely demonic is that "thinking" like this is so patently insincere; its function is to *persuade* ourselves that we are doing one kind of thing when actually we are doing quite another. When this happens, it is *no accident* that we choose what is contrary to our own decisions, nor that we see these choices as compatible with them. It is from the desire to act as we do that we end up, not only doing what violates our own decisions, but thinking all the while that we are acting as we said we would.

6. Closing the Gap?

What follows from my argument about how temptation can be resisted? In particular, what ways might there be of resisting temptation *other* than by refraining from reconsidering or revising our decisions—strategies that are, as I have argued, generally ineffective in the face of temptation to violation?

We can identify an inadequate answer to this question by beginning from a natural reply to the argument of section 4. On my account, it is possible for us to choose what is contrary to our standing decisions to the extent that the decisions we thereby violate fail to identify the *specific* acts and courses of action that they mandate or rule out. Why, then, can we not immunize ourselves to this form of temptation simply by making decisions whose content is more specific? This is, after all, just the kind of transition that I made originally, from "Get my paper finished by the deadline" to "Complete the next section before noon today." Should it not be possible to continue this process further down the line, thereby ensuring that sheer willpower is enough to stay on task, since I will be unable to act contrary to my decisions without revising them?

Well, let us try to imagine how this might go. Suppose that instead of resting content with the decision to make good progress on my paper this morning, I adopt a number of subsidiary policies like the following:

1. Get home from the gym no later than 6:45.
2. Make my own lunch at the same time as I make the kids'.
3. Open up my document as soon as I arrive in my office.
4. No checking email or social media.
5. Turn off notifications on my phone.

6. Take just a ten-minute break for coffee.
7. No fiddling with the bibliography.

Without question, this is often a smart kind of planning to go in for.[29] It is smart because it increases the likelihood that I will finish my work: the policies from 1 to 3 do this by helping me to get started earlier, while those from 4 to 7 do it by limiting the number of occasions on which I will be tempted to unproductivity. Yet we know all too well that nothing in this kind of planning is enough to ensure that I will follow through on the decision to do my writing, nor that the only way not to follow through is by taking that decision back. And the reason for this is, of course, that there are countless ways I could violate this decision that do not appear anywhere on my list—nor could I, even if I tried, produce in advance a list of what they all might be.

Nor is this problem solved if, instead of a set of focused measures like these, I simply adopt a very general policy like:

> *X*: Do not check my phone, leave my office, talk to my colleagues, or navigate away from my document until I have a full draft of this section.

The first thing to recognize about *X* is that it is not, *in general,* the best way of trying to go about one's writing—first, because it describes a course of action so unenjoyable that one is likely to have a strong desire to go back on it, and second, because often we write more effectively when we allow ourselves some flexibility in the process, including the opportunity to take occasional breaks. Further, even if *X* is a wise policy to adopt on a given occasion, adopting it is still no *guarantee* that I will get my writing done—since I could, after all, still spend most of the morning staring out the window while I "formulate my thoughts," or decide that I have finished a draft when all I really have is a bunch of stream-of-consciousness remarks. Alternatively, to the extent that in sticking to a policy like *X,* I thereby *force* myself to complete my writing, this will not be because this is a magical sort of policy that makes it impossible to fail to get my writing done unless I take the policy back, but rather because the course of action it prescribes is so unenjoyable that I will have plowed through my writing as quickly as I could with an eye to getting back on my phone—which, again, is not a great way of getting one's writing done.[30]

[29] This planning falls under what Sergio Tenenbaum calls the "vertical" dimension of practical wisdom (*Rational Powers in Action*, ch. 8).

[30] Put differently, if I follow *X* too slavishly, then I will manifest what Tenenbaum calls the *vice of rigidity*, i.e., the vice "of performing the characteristic actions of [a] policy too often or at the wrong times" (*Rational Powers in Action*, 199). In this case, what makes my rigidity a vice is not just that it interferes with my extra-professional ends but that it leads me to act irrationally with respect to the very end that my policy is supposed to serve.

Well, here is one more thing we might try, perhaps in conjunction with the policies from 1 to 7:

> *Y*: Each hour on the hour, check that my progress is on schedule, and allow myself a snack and a five-minute social media break if it is.

Once again, policies like *Y* are often good to have in place. What makes *Y* good is not just that it provides positive reinforcement, but also that it invites me to notice where I have gone off course, to form further plans to prevent this from reoccurring, and to pick up the pace if I have fallen off schedule. Nevertheless, adopting *Y* as a policy, and keeping it firmly in place with no room to reconsider, is still no guarantee that I will get my writing done, and not just because I might forfeit the snack breaks or let myself backslide during the final hour. It is rather because I need *judgment* to apply *Y* in any given case—to say whether, for example, it counts as being "off schedule" if during the past hour I wrote only a bit because instead I was reading that chapter from a book that was pertinent to my topic. Perhaps it should count, if I am not one to be trusted with that much latitude. But then again, perhaps it should not, since applying *Y* that strictly means actively disincentivizing courses of action could be good ways of achieving my ends. More generally, the hourly opportunities for checking-in that are mandated by this policy are a forced and ultimately second-rate substitute for the kind of judgment that ideally I would be able to carry out "on the fly," recognizing from moment to moment what I am doing, what the motivations are for it, and how I should proceed from here.

All these lessons illustrate a much more general point that has been noticed by philosophers at least since Aristotle—namely, that success in practical reasoning cannot be reduced to the application of well-defined rules. It does not follow from this that general rules are useless in practical deliberation, nor that all substantive practical principles admit of exceptions.[31] However, it does have the consequence that, first, even the maximally prudent person will not be able to identify *in advance* all the things she must do in order to achieve a certain goal, and second, that even when correct practical principles have been adopted, the task still remains of identifying what falls under them. And the discussion above shows how these lessons apply even to stretches of activity that are governed by a single overriding end. For even if I rank getting my writing done definitively above things like being collegial, knowing what is happening in

[31] Compare Aristotle's list in *Nicomachean Ethics*, bk. 2, ch. 6, of actions "whose names directly imply evil": adultery, theft, and murder. To the extent that we can give noncircular definitions of which actions are of these kinds, there may be action-guiding principles that prohibit them without exception. In the case of my office drama, such a principle might rule out plagiarizing my section from someone else's work or having it drafted by ChatGPT.

the world, reading my colleagues' gripes about their students, or simply having a generally pleasant and relaxing morning, nevertheless my commitment to this singular end is not enough to decide what I should do at each moment, nor to guarantee that I will choose in accordance with this end as long as I do not revise or abandon it. And, further, it shows how there can be a trade-off between the success that a policy will have in screening off tempting courses of action and the success it will have in helping me to do *well* the thing that the policy is in the service of.[32]

What makes temptation an ever-present reality for us is that following through on our decisions depends on the exercise of practical wisdom. In practice, and especially for people who are far from perfectly virtuous, what does this exercise involve? One thing it may involve is the kind of thing I have just discussed: a strategy of attempting to anticipate the various ways we might fail to follow through on our decisions, in order to head them off as well as we can. Another is the kind of thing emphasized by Bratman and Holton: the capacity, in situations where we might be inclined to revise our decisions and choose to do something else, to shut down this process except where it is reasonable. Yet something more is needed, too: the ability to see ourselves aright, to recognize which courses of action would be ways of undermining our goals rather than fulfilling them, and to make and reevaluate our specific decisions in relation to our wider ends.[33]

University of Illinois Urbana-Champaign
jlschwenkler@gmail.com

REFERENCES

Andreou, Chrisoula. "General Assessments and Attractive Exceptions:

[32] For further discussion of all these points, see Tenenbaum, *Rational Powers in Action*, ch. 8.

[33] My work on this paper was supported by a sabbatical leave from Florida State University and a fellowship from the Notre Dame Institute for Advanced Study. I twice presented versions of it to audiences at Notre Dame: once to my NDIAS colleagues and once to the Department of Philosophy. I also presented later versions at Hillsdale College, Loyola University Chicago, the University of Illinois Urbana-Champaign, Georgia State University, and Blackfriars Hall, Oxford. I then gave a near-final version at a conference for Anselm Müller at the University of Erlangen-Nuremberg, where Anselm gave a characteristically perceptive reply. I thank the audiences on all these occasions, and especially Marshall Bierson, Blake Dutton, Timo-Peter Ertz, Nathan Helms, Gavin Lawrence, Beri Marušić, Erasmus Mayr, Juan Piñeros-Glasscock, Anthony Price, Sergio Tenenbaum, Maura Tumulty, and Neil Van Leeuwen for feedback and discussion.

Temptation in *Planning, Time, and Self-Governance.*" *Inquiry* 64, no. 9 (2021): 892–900.

———. "The Good, the Bad, and the Trivial." *Philosophical Studies* 169, no. 2 (June 2014): 209–25.

———. "Temptation, Resolutions, and Regret." *Inquiry* 57, no. 3 (April 2014): 275–92.

Andreou, Chrisoula, and Mark D. White, eds. *The Thief of Time: Philosophical Essays on Procrastination*. Oxford: Oxford University Press, 2010.

Anscombe, G. E. M. *Intention*. 2nd ed. Cambridge, MA: Harvard University Press, 2000.

Arruda, Caroline. "Sticking to It and Settling: Commitments, Normativity, and the Future." In Bagnolia, *Time in Action*, 149–72.

Bagnoli, Carla. "Hard Times: Self-Governance, Freedom to Change, and Normative Adjustment." In Bagnoli, *Time in Action*, 196–217.

———, ed. *Time in Action: The Temporal Structure of Rational Agency and Practical Thought*. New York: Routledge, 2022.

Betzler, Monika. "Inverted Akrasia." In Bagnoli, *Time in Action*, 243–63.

Blackburn, Simon. *Mirror, Mirror: The Uses and Abuses of Self-Love.* Princeton: Princeton University Press, 2014.

Bratman, Michael E. "Acting Together with Oneself over Time: Appendix to 'A Planning Agent's Self-Governance over Time.'" In Bagnoli, *Time in Action*, 95–107.

———. *Intention, Plans, and Practical Reason*. Palo Alto, CA: CSLI Publications, 1999.

———. "Planning and Temptation." In *Faces of Intention: Selected Essays on Intention and Agency*, 35–57. Cambridge: Cambridge University Press, 1999.

———. "A Planning Agent's Self-Governance over Time." In *Planning, Time, and Self-Governance*, 224–49.

———. *Planning, Time, and Self-Governance: Essays in Practical Rationality*. Oxford: Oxford University Press, 2018.

———. "Planning, Time, and Self-Governance: Replies to Andreou, Tenenbaum, and Velleman." *Inquiry* 64, no. 9 (2021): 926–36.

———. "Rational Planning Agency." In *Planning, Time, and Self-Governance*, 202–23.

———. "Temptation and the Agent's Standpoint." In *Planning, Time, and Self-Governance*, 149–67.

———. "Temptation Revisited." In *Structures of Agency: Essays*, 256–82. Oxford: Oxford University Press, 2007.

———. "Toxin, Temptation, and the Stability of Intention." In *Faces of Intention: Selected Essays on Intention and Agency*, 58–90. Cambridge: Cambridge

University Press.

Davidson, Donald. "How Is Weakness of the Will Possible?" In *Essays on Actions and Events*, 21–42. Oxford: Oxford University Press, 1981.

den Hartogh, Govert. "The Authority of Intention." *Ethics* 115, no. 1 (October 2004): 6–34.

Ferrero, Luca. "Decisions, Diachronic Autonomy, and the Division of Deliberative Labor." *Philosophers' Imprint* 10, no. 2 (February 2010): 1–23.

———. "Diachronic Constraints of Practical Rationality." *Philosophical Issues* 22, no. 1 (October 2012): 144–64.

———. "Diachronic Structural Rationality." *Inquiry* 57, no. 3 (2014): 311–36.

———. "The Structures of Temporally Extended Agents." In Bagnoli, *Time in Action*, 108–32.

———. "Three Ways of Spilling Ink Tomorrow." In *Rationality in Belief and Action*, edited by Elvio Baccarini and Snježana Prijić-Samaržija, 95–127. Rijeka, Croatia: Faculty of Arts and Sciences, 2006.

———. "What Good Is a Diachronic Will?" *Philosophical Studies* 144, no. 3 (June 2009): 403–30.

Gauthier, David. "Assure and Threaten." *Ethics* 104, no. 4 (July 1994): 690–721.

———. "Commitment and Choice: An Essay on the Rationality of Plans." In *Ethics, Rationality, and Economic Behaviour*, edited by Francesco Farina, Frank Hahn, and Stefano Vannucci, 217–43. Oxford: Clarendon Press, 1996.

Gold, Natalie. "Guard against Temptation: Intrapersonal Team Reasoning and the Role of Intentions in Exercising Willpower." *Noûs* 56, no. 3 (September 2022): 554–69.

———. "Putting Willpower into Decision Theory: The Person as a Team over Time and Intrapersonal Team Reasoning." In *Self-Control, Decision Theory, and Rationality: New Essays*, edited by José Luis Bermudez, 218–39. Cambridge: Cambridge University Press.

Greene, Preston, and Meghan Sullivan. "Against Time Bias." *Ethics* 125, no. 4 (July 2015): 947–70.

Heeney, Matthew. "Diachronic Agency and Practical Entitlement." *European Journal of Philosophy* 28, no. 1 (March 2020): 177–98.

Hinchman, Edward S. "Conspiracy, Commitment, and the Self." *Ethics* 120, no. 3 (April 2010): 526–56.

———. "Narrative and the Stability of Intention." *European Journal of Philosophy* 23, no. 1 (March 2015): 111–40.

———. "Trust and Diachronic Agency." *Noûs* 37, no. 1 (March 2003): 25–51.

Holton, Richard. *Willing, Wanting, Waiting*. Oxford: Oxford University Press, 2009.

Jaffro, Laurent. "Weakness and the Memory of Resolutions." In Bagnoli, *Time* *in Action*, 221–42.

McClennen, Edward F. *Rationality and Dynamic Choice: Foundational Explorations*. Cambridge: Cambridge University Press, 1990.

Morton, Jennifer M. “Deliberating for Our Far Future Selves.” *Ethical Theory and Moral Practice* 16, no. 4 (August 2013): 809–28.

Müller, Anselm Winfried. “Acting Well.” *Royal Institute of Philosophy Supplements* 54 (March 2004): 15–46.

Nefsky, Julia, and Sergio Tenenbaum. “Extended Agency and the Problem of Diachronic Autonomy.” In Bagnoli, *Time in Action*, 173–95.

Paul, Sarah K. “Diachronic Incontinence Is a Problem in Moral Philosophy.” *Inquiry* 57, no. 3 (2014): 337–55.

———. Review of *Willing, Wanting, Waiting*, by Richard Holton. *Mind* 120, no. 479 (July 2011): 889–92.

Raz, Joseph. “Reasons for Action, Decisions and Norms.” *Mind* 84, no. 336 (October 1975): 481–99.

Roth, Abraham Sesshu. “Agency and Time.” In Bagnoli, *Time in Action*, 133–48.

Rovane, Carol Anne. *The Bounds of Agency: An Essay in Revisionary Metaphysics*. Princeton: Princeton University Press, 1998.

Schwenkler, John. *Anscombe's Intention: A Guide*. New York: Oxford University Press, 2019.

Smith, Matthew Noah. “Sovereign Agency.” In *Reasoning: New Essays on Theoretical and Practical Thinking*, edited by Magdalena Balcerak Jackson and Brendan Balcerak Jackson, 248–60. Oxford: Oxford University Press, 2019.

Stroud, Sarah, and Christine Tappolet, eds. *Weakness of Will and Practical Irrationality*. Oxford: Oxford University Press, 2003.

Tenenbaum, Sergio. “On Self-Governance over Time.” *Inquiry* 64, no. 9 (2021): 901–12.

———. *Rational Powers in Action: Instrumental Rationality and Extended Agency*. Oxford: Oxford University Press, 2020.

Tenenbaum, Sergio, and Diana Raffman. “Vague Projects and the Puzzle of the Self-Torturer.” *Ethics* 123, no. 1 (October 2012): 86–112.

Velleman, J. David. “Deciding How to Decide.” In *The Possibility of Practical Reason*, 221–43. Ann Arbor, MI: University of Michigan Press, 2000.

Verbeek, Bruno. “On the Normativity of Intentions.” *Topoi* 33, no. 1 (April 2014): 87–101.

———. “Rational Self-Commitment.” In *Rationality and Commitment*, edited by Fabienne Peter and Hans Bernhard Schmid, 150–74. Oxford: Oxford University Press, 2007.

Journal of Ethics and Social Philosophy
vol. 27, no. 3 · may 2024

https://doi.org/10.26556/jesp.v27i3.3048

VALUE CAPTURE

C. Thi Nguyen

Here is a story about how metrics can change people. A relative of mine had been planning a long European vacation with some old friends, John and Shelley. My relative had been looking forward to seeing the sights with her friends—touring museums, seeing operas, having long dinners. But, she says, the entire vacation was dominated by John and Shelley's relationship with their Fitbits. John and Shelley would not go to the opera with her: not enough steps. They would cancel dinner dates because they had not met their daily step goals yet. My guess is that John and Shelley never consciously decided that step counts were more important than, say, art or friendship. The Fitbit just spoke more loudly in their internal deliberation, and there was no Artbit or Friendbit to compete. The clarity of those metrics just swamped quieter considerations.

And even if fitness was your main goal, the Fitbit can exert a narrowing influence. Exercise can be valuable in all sorts of ways that are not measured by a Fitbit. A Fitbit does not capture the ecstasy of complex, skillful motion. It does not capture the camaraderie of team sports, the meditative calm of paddling a canoe across a quiet lake, or the aesthetic loveliness of a delicate rock-climbing move. A Fitbit measures exactly one thing: steps. That limitation arises from its particular institutional and technological embeddedness. Fitbits are constrained by what mass-produced devices can easily measure and aggregate, given present-day technologies and institutional arrangements. We know how to make a watch that automatically measures steps, but not how to make a watch that automatically tracks your spiritual renewal.

Of course, you do not have to value what the Fitbit measures. You could just use a Fitbit as a source of data. But the Fitbit tempts us to do more. The Fitbit presents its output, not just as mere information but as an evaluation: a score. And when you buy into the Fitbit's preferred motivational scheme—when you adopt its scores as your values—you get all kinds of rewards. You gain the motivational benefits of having clear feedback about how well you are doing, of competing along a well-defined scale. All you have to do is give up on having fine and detailed control over your own values. Here is one way to put it: when you buy into a Fitbit's preferred value system, you are *outsourcing the process of value deliberation.*

Fitbit is just one example of a larger phenomenon that we can call *value capture*. Value capture happens when your environment presents you with simplified versions of your values, and those simple versions come to dominate your practical reasoning. Value capture offers you a quick shortcut—an opportunity to take on *prefabricated* values. You do not have to go through the painful process of value deliberation if you can get your values off the shelf.

I want to focus on one particularly clear, and quite common, form of value capture: when an institution presents you with some metric, and then you internalize that metric. You start exercising for your health, but you come to care about losing weight or optimizing your body mass index. Or you go on Twitter to connect to people and have fun but come to care more about maximizing your like, retweet, and follower counts. Or you go into philosophy graduate school for a love of wisdom but come out aimed at getting fancy grants, publications into highly ranked journals, and placement at a highly ranked institution. As anthropologist Sally Engle Merry puts it, the culture of indicators and metrics is "a form of governance that engages a person in governing himself or herself in terms of standards set by others."[1] I will focus for much of this paper on such institutional value capture. Metrics are the starkest case of value capture, and we are fortunate to have a rich empirical literature studying the social effect of metrics. But metrics are just a starting point; there are many other forms of value capture worth investigating.

Many of us feel an intuitive horror when contemplating cases of institutional value capture. But it is rather difficult to say in a principled way exactly why value capture is so horrifying. For one thing, value capture is often consensual. People buy Fitbits precisely because they know that those step counts will motivate them; they *want* to be captured because the motivational bump seems worthwhile. Such gamified technologies are frequently sold as a way to overcome weakness of the will and seem to succeed at doing so. The point of a Fitbit is to motivate you to walk more, and it does seem to work.

Why might this strike some of us as horrifying rather than as simply a useful and empowering tool? I will suggest that there is a problem with the *nature of the values* on offer. The problem with internalizing institutional metrics is not simply that we are getting our values from the outside. It is that such metrics are subject to the demand for a certain kind of stability and institutional usability. These institutional demands push our metrics away from the subtle, the dynamic, the sensitive—and toward what can easily be measured at scale, propagated across institutional units, and recorded in institutional memory. When we take on such metrics as our values—when we internalize them—we

[1] Merry, *Seductions of Quantification*, 33.

are imposing a narrowed filter on our values. We are letting the logic of institutions play a determining role in the articulation of our values.

Institutional value capture offers us a delightful reward. Once we have permitted ourselves to be value captured, our values become clear, coherent, and shared. Now we can be easily understood—unambiguously, almost effortlessly. But such clarity requires a degree of stabilization. Such clear, stabilized values arise from and are deeply embedded in external institutions and institutional processes. That stabilization has some benefits and some costs. Sometimes those costs may be worth paying—but we should at first get clear about what, exactly, they are.

In value capture, we *outsource the process of value deliberation*. And, as with other forms of outsourcing, there is a trade-off. You get the outsourced objects quickly and easily, and they fit neatly into a larger network of other standardized and modular parts. Somebody else has formulated our values for us and done the work of embedding them in readymade systems of measurement and technologies of motivation. When we adopt those values, we gain access to readymade methods for justification. It is easy to justify yourself in the language of metrics because metrics are easy to understand. They have, in fact, been engineered to be so. The cost of value capture is that we give up on the process of finely tuning our values to our own context: our personalities, our peculiar culture, our particular corner of the world. Outsourced values are not custom-tailored. In value capture, you are taking on prefabricated values.

1. A CASE STUDY: THE LAW SCHOOL RANKINGS

The social draw of quantification has been the subject of some extremely useful recent empirical studies from anthropologists, historians, and sociologists. My favorite is *Engines of Anxiety,* a study of the cultural effects of the *US News and World Report* (*USNWR*) law school rankings by sociologists Wendy Espeland and Michael Sauder.[2]

Before *USNWR*, they say, there were no law school rankings. Students often picked law schools through a complex process of evaluation, deliberation, and self-reflection. They got to know a school by reading about its mission, by talking to people, or by visiting. Importantly, different law schools pursued different missions. Some were tuned to academic legal research, others to the corporate world. Some law schools were devoted to social activism—toward supporting the local community or serving underrepresented populations. The process of choosing a law school often triggered a certain degree of soul-searching in the

[2] Espeland and Sauder, *Engines of Anxiety*.

students. The complex value plurality involved in the choice pushed students to reflect on what they wanted from their own legal educations and legal careers.

The rankings displaced all that. Espeland and Sauder studied online discussions between prospective law students. They found that once *USNWR* started publishing its rankings, those rankings came to dominate the choice process for most students. And the same is true for nonstudents: the public perception of law schools immediately reoriented itself along *USNWR*'s rankings. Espeland and Sauder say that the rankings drove value plurality out of the legal educational system. Many schools used to genuinely pursue their different missions. And many of those missions involved pursuing values that are not tracked by *USNWR*'s ranking formula—like, say, supporting local underserved minority communities. But following such a distinctive mission invariably meant dropping spots in the rankings, which promptly resulted in precipitous drops in donations and student interest. Most schools, report Espeland and Sauder, have since abandoned their original missions and reoriented their admissions process and educational methodology toward performance in *USNWR*'s ranking calculations. And what matters the most to that ranking is the grade point average (GPA) and Law School Admission Test score of the incoming class and the employment rate of the outgoing class.[3]

In the case of the law schools themselves, the change in goals could be understood as a case of perverse incentives. Law school administrators were forced to align their efforts with the rankings, even if their own values were unchanged. But with prospective law students, the problem seems to run much deeper. The rankings seem to exert a magnetic pull over students' values. Some students, of course, were merely responding to incentives—since potential employers also care about law schools' rankings. But a majority of students, say Espeland and Sauder, seemed to care directly about those rankings. Instead of exploring their own values and desires for their legal education, they seem to presume that the process of going to law school should be oriented toward getting into the "best" law school, where "best" is determined strictly by the rankings. The existence of that clear, vivid, objective-seeming list offers an easy substitute for the process of personal value deliberation.

The effect on students I take to be a clear example of value capture. The fact that value capture occurs I take to be an empirical matter—and its existence is well-documented.[4] My goal here is to think about the harms of value capture.

[3] Espeland and Sauder, *Engines of Anxiety*, 43.

[4] Beyond Espeland and Sauder, see Porter, *Trust in Numbers*; Scott, *Seeing Like a State*; and Merry, *Seductions of Quantification*, for good entry points into the literature.

2. VALUE CAPTURE

Value capture happens when:

1. An agent has values that are rich, subtle, or inchoate (or they are in the process of developing such values).
2. That agent is immersed in some larger context (often an institutional context) that presents an explicit expression of some value (which is typically simplified, standardized, and/or quantified).
3. This explicit expression of value, in unmodified form, comes to dominate the entity's practical reasoning and deliberative process in the relevant domain.

If you would like a portable version, try this: *value capture happens when a person or group adopts an externally sourced value as their own, without adapting it to their particular context.*

Let us take a moment to get clearer on what, exactly, counts as value capture. First, notice that value capture includes both voluntary and nonvoluntary adoptions of an external value. It certainly counts as value capture if, say, you were brainwashed, and an external value was somehow injected into you against your will. But it equally counts as value capture if you willingly and voluntarily adopted that external value—perhaps because it is easier or helps you to fit more easily within your profession or because it lets you avoid the painful process of value deliberation. The target of my criticism here is not simply those cases of involuntary value transformation. I am interested in the problem with letting externally sourced values dominate one's practical reasoning—even if that dominance was established knowingly and consensually.

Next, my definition of value capture is narrowly aimed at those cases where the entity uses the external expression of value precisely as given. It is aimed at those cases where we internalize and deploy an external value just as we found it, *without further adjustment*—without further contouring it, interpreting it, or fine-tuning it to ourselves. Value capture does not include cases where you get the seed of your values from the outside and then start fiddling with them. If you get the starting seed for your values from your family, your culture, your religion but then tweak them to fit your personality and place in the world—that is not value capture. Value capture is when an externally sourced value, like a metric, comes to dominate your practical reasoning in its given form—when your goal is simply to get to that higher ranking, those higher citation rates, those more likes.

I have been speaking so far about the value capture of individuals by large-scale institutions. Such examples are vivid and familiar. But they can invite a simplistic reading of the problem: that real values are somehow original inventions of the individual, and that socially generated values are somehow fake. It is tempting to think that what is going on here is, say, a battle over individual authenticity—some conflict between the solitary free spirit and the forces of social conformity. But the problem is much more complex than that. For one thing, our values are often acquired in their initial seed from social sources—parents, teachers, friends, colleagues. For another, we often develop our values in community with others.

And crucially, value capture is a problem that can afflict groups too. A philosophy department can be captured by the larger university's focus on student evaluation scores. In my own experience, the clarity of an institutional metric can quickly come to dominate the attention of a deliberating group. Even when a group agrees that they care more about some inchoate value—like, say, fostering curiosity—the actual day-to-day decisions end up driven by whatever clear metrics happen to be on hand. Merry, in *The Seductions of Quantification*, offers a good example.[5] The United Nations (UN) publishes the Human Development Index (HDI)—a quantified ranking of all countries in the world in terms of how they supported quality of life. Merry says that the committee behind the HDI published it with a very clear and loud set of qualifications. They published it with a lengthy report on the complexity and multidimensionality of "quality of life" measures and clearly stated that the HDI ranking was simply a gross oversimplification. Unsurprisingly, says Merry, the full report was largely ignored. Once the HDI was published, governments the world over became incredibly invested in advancing their ranking—even though the score was not attached to any concrete real-world incentives or rewards. Here is a case where entire governance cultures have been value captured by an external metric. So, in my account of value capture, I intend "agent" to be in a broad sense, including individual persons and group agents.[6]

Next: condition 3 in my account specifies that the external expression of value "comes to dominate" the entity's practical reasoning and deliberative process in the relevant domain. I mean the notion of "dominate" to be quite substantive here. Value capture occurs when an external value becomes the dominant source of reasons for action in a domain. It is not value capture if I adopt an external value in a controlled manner—as temporary instruments,

[5] Merry, *Seductions of Quantification*.

[6] I use "entity" rather than "agent" here because, while "agent" includes group agents, I think the category is not large enough. I suspect that some loosely organized communities can qualify as having values but not have sufficient internal cohesiveness to count as a group agent. See discussion of shared values in Hedahl and Huebner, "Sharing Values," and of loose community values in Nguyen and Strohl, "Cultural Appropriation."

accountable to my own richer values. I mean to exclude here from the category of value capture those cases where we use external values as proxies and heuristics under full reflective control—when we select, monitor, and adapt those heuristics in the light of our own richer values.

Suppose that I want to get healthier and more fit. By "healthy and fit," I mean something complex and textured and difficult to express—something about feeling good in my body, being more capable of comfortably executing complex physical tasks, and getting rid of this feeling of awkward clumsy brokenness that too much laptop time has left me with. But such inchoate and airy expressions of value are pretty hard to use in the rush of daily life. Beings like us need heuristics—simple and clear rules of thumb to use in the day to day. And I can pick a heuristic, like increasing my step counts, as a quick-and-easy decision procedure to use in my daily life as a way of pursuing that richer notion of health.

But such heuristics are not usually supposed to supplant our full values entirely. We are supposed to use them with the knowledge that they are mere proxies for our full values. They are supposed to serve our dominant values, which means they should be revisable and discardable under the light of our full values. What I really want is health in this richer sense, but I also know that I need an easier target to aim at on a daily basis in order to get myself motivated. So, I start using a Fitbit and just aim at getting step counts. But after a few months, I step back and reflect on my time with the Fitbit. Has pursuing step counts made me happier? Is my body performing better? Do I feel less broken and awkward? Perhaps the answer is affirmative and I keep going with the Fitbit; perhaps the answer is negative, and I abandon it and try some other proxy goal. Perhaps the answer is a qualified yes, and I modify my approach, adding a few more goals to the mix beyond just maximizing my steps. This *controlled use of a proxy value* is not a case of value capture since the externally sourced value does not dominate my deliberative procedure. The dominant value is not the Fitbit's step measures but something else, and this can be seen by the fact that I do sometimes adopt a reflective stance where I decide whether adopting the Fitbit's goals is serving my real values and decide whether to continue or discontinue my use of that simple proxy in the day to day. The Fitbit is not in charge.

Similarly, it is not value capture when I am merely *taking the metrics into account* in the pursuit of my own rich and textured goal. Suppose I want to be a legal activist working for immigration reform. I know that going to a high-ranked law school will be important for getting the influence I need for this kind of work. In that case, I will pay attention to the law school rankings—but they do not dominate my practical reasoning. I may see the instrumental value to going to a highly ranked law school, but I can also trade off that instrumental value against other things I value. I will not, for example, just go to the highest-ranked school possible. I will use the rankings for my purposes—trying to find a compromise between a well-ranked school that will get me the power I need and a school that will help me learn to do the activist work I want to do. That kind of instrumental awareness of ranking systems is a long way off from a real case of value capture—where somebody's primary goal was, say, simply to go to the highest-ranked law school.

Next: value capture can happen at different points in an agent's life arc with values. Sometimes, an agent has already established their values, and then they come to be *replaced* by some external metric. Other times, the agent does not yet have their own articulated values; they are in the process of figuring them out. But the existence of a prefabricated value offers them a *shortcut* in the process of value deliberation. They can simply adopt a ready-made value instead of going through the slow and oftentimes painful process of figuring out and adjusting their values to their own personality and circumstances. The definition of value capture is intended to include both replacement and shortcut cases.

Finally: value capture can happen at some different loci. One kind of value capture involves the *wholesale* capture of the entire value—such as when you got into this career for joy but came to care only about the money. In wholesale value capture cases, the agent systematically changes how they think of their values; they come to describe their values differently and report them differently. But just as common as these wholesale cases, I suspect, are cases of what we might call *application capture*. In such cases, an external expression of a value does not replace how the agent conceives of their original value—in how they would think about and report their values in the abstract. But the external value dominates how they act by setting the practical criteria in day-to-day applications of their values in particular decisions and evaluations. Say that I, an academic, care about the pursuit of truth, wisdom, and understanding. Across my career, if asked, I would describe my core values using those same terms. But suppose that in the course of my professionalization, the way I apply those terms changes. Now, whenever I try to evaluate the success of my articles, I turn to certain metrics, like the citation rate or the status of the publication venue on some ranked list. And when I evaluate my overall success as an academic, I turn to metrics like my total citation rate or the status of my institution on some ranked list. In that case, it is those institutional metrics, and not the vaguer values I report upon reflection, that effectively dominate my actual actions and self-evaluation. Here, the metric gains dominance by capturing, not the general terms in which I articulate my values, but the more specific application criteria I use when the values hit the ground. The metric fills out the process by which

I determine whether I have fulfilled my core values. And suppose that I guide my actions based on those evaluations: I start writing papers that are more like the ones that have succeeded, in these terms, and start taking actions that might advance my general success, in these terms. Then those external criteria have come to effectively dominate my practical decisions.

There is a crucial difference between controlled use of a proxy as I described it earlier, and application capture. When we *use* a metric as a mere proxy, our richer values are in charge. We will regularly reflect on the proxy from the perspective of our full values, and modify, discard, or adapt that proxy. In the application capture cases, we let the proxy take charge. It functions as the effective practical translator, connecting our abstract expressions of value into specific cases of evaluation—controlling how we apply our values to the world. (Though, at least from my own observations, many cases of value capture start as innocuous-seeming uses of a proxy. I have heard many people say that they put on a Fitbit in order to pursue some other goal, like health or happiness, but then years later they found that they had forgotten about that larger goal—that doing well in the Fitbit's terms had come to occlude all else.) From here on out, I will speak of "values" being captured, for brevity's sake—but I mean to indicate both wholesale value capture and application value capture.

To sum up: value capture does not include every interaction with rankings or metrics. It does not include the controlled use of proxies and heuristics, nor the informational use of metrics. Value capture occurs when an externally sourced value plays the dominant role in practical reason—when it gets put in charge for some domain. This looks like: people who pursue step counts even when it hurts their knees and exhausts their spirit; academics who pursue publications in the highest-ranked journals even when their work feels boring and meaningless; universities that pursue high rankings in the *USNWR* over richer understandings of education; newspapers that pursue clicks and pageviews over their own sense of newsworthiness and social importance. And, as I have noted: the empirical work indicates that this sort of robust value capture is actually quite common.

Value capture is distinctive because we do not change or adapt the particular externally sourced specification of a value to our particular context. Compare this with other, more open-ended and dynamic relationships we might have to externally sourced values. We often get our first grip on a new pursuit—and its value—with another's help. A friend shows me the wonders of horse riding, the beauty of jazz, the depth of haiku. They talk about what they find meaningful and rich in the activity; they guide me into the actions and attentions that will help me get onto its distinctive value. As Tal Brewer puts it, the values of activities are often obscure to the outsider or novice; it takes a long process of immersion in the activity to get onto its true value.[7] And we often need help to find our way in.[8] The friend who taught me to see jazz talked me into it—into the particular thrill of seeing a live improvisation. But I suspect she would have been very disappointed if, ten years down the line, the value I found in jazz still precisely mirrored her own. Like any good art friend, she hoped that I would eventually fly on my own wings and sharpen the details of my love of jazz in my own way.

And I did—I used her guidance to find my way in and then slowly began to develop my own relationship with jazz, finding out what thrilled and moved me in the music. This is not a case of value capture; I have used external guidance to get my first grip on the terrain of value in an activity, but then significantly tailored my sense of value in light of my own experiences. Value capture cases are the cases where I internalize, wholesale, an externally sourced value and permit it to dominate my reasoning in its unchanged form. This is where the "outsourcing" metaphor is particularly useful. The harms of outsourcing do not depend on any involuntariness. I can wholeheartedly consent to outsourcing. The harms come from the particular content and nature of outsourced objects—of their inflexibility and prefabrication. If you want a slogan: our values should be tailored to our particular selves and our particular context—but in value capture, we buy our values off the rack.

3. The Problem of Value Capture

What, then, is the cost of outsourcing one's values? First, to be clear: I am not trying to argue that value capture is always wrong. Value capture, as with any other form of outsourcing, involves a trade-off between efficiency and fine tuning. I think that we are often clear on the benefits of that trade-off but fail to plainly see the costs. My goal is to articulate more precisely the costs.

But it turns out to be rather hard to articulate the problem with value capture. First, what is wrong with getting our values from external sources? It seems utterly naive to think that our values need to spring fully formed from some magical inner place, wholly devoid of social origin. We are deeply social beings, and we often seem to get our values from our culture, our community, our social context. Second, how could value capture undermine autonomy? Many cases of value capture are entirely voluntary and consensual. We know that Fitbit motivates because it presents information in public and shared terms—though the

[7] Brewer, *The Retrieval of Ethics*.

[8] See Nguyen, "Trust and Sincerity in Art," for a discussion of how trust in others is often required to provide the motivation for attending to difficult or obscure art forms.

full implications of that publicity may not be entirely obvious. Consensual value capture can seem like an aid to autonomy. People often seek out such gamifications in order to overcome weakness of the will. People buy Fitbits or use Duolingo precisely because the gamified structure—in which they are awarded points and levels for progress—makes them more able to get certain things done, like start exercising or learn a language. They are *hoping* for value capture, and they are choosing the effect with some understanding of the basic mechanism. As Jane McGonigal puts it, gamification is a force for good because it can turn monotonous tasks into fun.[9] If value capture can help us overcome weakness of the will, then it helps increase our autonomy and agency. So, what is the harm?

There are at least three ways to think about the potential harm of value capture. First, it might be that *autonomous* participation in the formulation of our values seems good in and of itself—and not just mere one-off consent to a big package, but a fine-grained and ongoing autonomous control of the details of our values. If that were true, then value capture would undermine our autonomous control over our values.[10] Second, institutional values are subject to demands for *hyperexplicitness*, and hyperexplicit values seem unlikely to adequately capture the full richness and subtlety of human values.[11] Third, the kinds of external values we encounter are typically formulated according to the interests and perspectives of large-scale institutions. They are, we might say, *standardized* values. Such values seem unlikely to fit the varying and peculiar interests and situations of particular people and smaller-scale groups.

[9] McGonigal, *Reality Is Broken.*

[10] Of the three options outlined here, the autonomy option is the one I am most undecided about. While intuitively appealing, developing such an account depends on walking the tightrope between specifying a substantive condition of autonomy, while keeping a grip on the social sourcing of many of our values, even in the most autonomous cases. An earlier, and much simpler, version of this paper attempted to offer an analysis of the harms of value capture in terms of a violation of autonomy, but that version of the argument could not survive contact with the insights from the literature on relational autonomy, especially from feminist critics of idealized pictures of autonomy. See Buss, "Valuing Autonomy and Respecting Persons"; Christman, "Relational Autonomy, Liberal Individualism, and the Social Constitution of Selves"; Khader, "The Feminist Case against Relational Autonomy"; Superson, "Deformed Desires and Informed Desire Tests" and "Deferential Wife Revisited"; and Westlund, "Selflessness and Responsibility for Self." It remains to be seen whether there is some more refined version of the autonomy worry that can be made in light of this discussion.

[11] I am exploring this possibility in other work. See Nguyen, "Value Collapse," for a discussion of the possibility that hyper-explicit values represent a bad epistemic attitude toward the world of value—that they discourage exploration of the space of value by making it easy to dismiss new candidates for value.

A full account of value capture would need to address at least these three approaches, and I hope one day to provide such a full account. Here, I can only take a first step. I will concentrate on the issue of *standardized values*—in part, because I think it highlights the unique problems of value capture.

Here is the worry in a nutshell. Value self-determination is important for all sorts of reasons. Here is one: value self-determination yields values that are finely tuned to our particular context. By substantively participating in the detailed process of formulating our values for ourselves, we can get values that nicely fit our particular circumstances—our individual psychologies and phenomenology, our group culture, our local context. Value capture intrudes on that process of value self-determination, *substituting prefabricated and standardized values for finely tailored ones*. Note that this is not an argument that autonomous value formulation is a good in and of itself, but rather an argument that substantively participating in the process of shaping one's values is instrumentally good in that it yields better, more finely tailored values.

Here is an example from my own life. For the first two decades of life, I avoided most physical activity. I had an incredibly simplistic conception of the value of exercise. I thought that exercise was basically pounding out some miles on a treadmill to burn some calories. Eventually, I came to see the vast and varied joys of athleticism. But in order to get there, I took a long meandering journey through many different sports, each of which paid off in profoundly different ways. Long-distance running turns out to be Zen-like and calming. Trail running requires more attention but offers this thrilling sense of reactive flow to the difficulty of the trail. Deadlifting is brutal and intense, a pure shot of grueling focus. And rock climbing turns out to be a fascinating fusion of bodily aesthetics and puzzle solving, where you solve thorny movement puzzles through elegant motion.[12] And even inside one of these activities, there is not some singular value on offer. Rock climbing can be pursued in radically different ways, each of which offers very different rewards. You can seek out thrills and risk; you can do easy climbs in rapturously beautiful terrain; you can focus on finding climbs with graceful movement; or you can go for gruelingly athletic climbs on a cave roof just eight feet off the ground. Each of these different ways of valuing rock climbing suggests a different way of approaching it, which in turn yields richly different textures of activity.[13] This is a process of *exploration,* where you try things out, figuring out how they fit with you, and

[12] For more on the aesthetic qualities of movement in rock climbing and other games, see Nguyen, *Games* and "Arts of Action."

[13] This description has been deeply influenced by Tal Brewer's account of how the formulation of the value of an activity and the way we do an activity form a feedback loop as we explore and refine our understanding of the activity (*The Retrieval of Ethics*). Agnes Callard

changing around your approach in response, seeing how it goes in an ongoing loop of feedback and adjustment.

The worry then is that when you are value captured by a Fitbit, you do not go through that process of exploration and fine-tuning. My claim here is not that one puts on a Fitbit and is automatically value captured. One could simply use a Fitbit as a data-gathering system to pursue one's own values. But Fitbit does present its step counts as a *score*. It is a gamified system, which openly employs design features from games.[14] A Fitbit does not force value capture, but it certainly invites it.

We might even call this an extended value system. Some philosophers have been very excited to claim that our minds are extended beyond our bodies—that our minds can include various technologies as parts of their internal functioning.[15] Most of the discussion of extended mind has focused on adopting, as part of our extended mind, various value-neutral cognitive resources—like using a notebook or Google Docs as an extended memory. Some of the discussion has gone so far as to suggest that we can extend our mind to use various technologies as part of our emotion regulation system, such as Joel Krueger's suggestion that we use our portable music devices for mood regulation.[16] My suggestion is one further step: in some cases, a *standard for evaluation* is embedded in a technology, as in Fitbit's step counts or Twitter's likes. When we integrate that technology into our cognition, our extended mind now includes a *value system* which was created externally, and which is sustained through external technologies.

Of course, one might respond, we get our value systems from external sources all the time— from our parents, our community, our culture. But, in the unproblematic cases, we can use external values as a starting seed, which we can adapt and tailor to ourselves. The worry is that we simply plug in these external values and use them as is. In particular, some external values can *resist further tailoring*. This is especially likely when adopting a particular prefabricated value is appealing precisely *for the standardization.* Then we will be quite tempted to leave them as they are, or else lose out on the promised efficiency.

has also written on such proleptic ends, though her account adds a requirement that the process is triggered by a desire to become the kind of person who so values (*Aspiration*).

[14] See Nguyen, "How Twitter Gamifies Communication," for a discussion of how certain user interfaces can present metrics as scores, and thus as forms of evaluation.

[15] Alternately, if we want to avoid the endless tussles about what exactly is the line between mind and not mind, we can use Kim Sterelney's locution: that various technologies are scaffolds for agency ("Minds"). Either locution takes us to the same worry.

[16] Krueger, "Music as Affective Scaffolding."

Value standardization is like any other kind of standardization: we gain in efficiency but in exchange for giving up localized tailoring.[17]

To really understand the problem here, we need to have a good grip on why we should want to tailor our values to fit. Elijah Millgram offers a useful account of how we adapt and improve our values in *Practical Induction*.[18] He is not talking about abstract, generic renderings of value, like, say, "happiness" or "flourishing." He is interested in the specific, grounded articulation of our values and goals by which we conduct our day-to-day lives, such as a runner's pursuit of a better marathon time, or a filmmaker's interest in playfully subverting genre conventions, or a philosopher's interest in writing deep, rigorously argued papers. Crucially, says Millgram, we do not derive these specific articulations of value by deriving them from some abstract specification of the good. Rather, we acquire our particular values and goals via a process of practical induction. We *try on* particular goals and values for a while. We might enter a profession and try on the goals associated with that profession: a literary fiction writer might start caring about achieving realism of character and setting; a Montessori teacher about fostering autonomy in very young children. And then the person gets *feedback* from the experience of living life under that particular value system. They might get positive feedback, like feeling engaged, happy, or interested; they find themselves savoring the details of their life. Or they might get negative feedback: they feel bored, listless, disengaged. This is feedback about the *fit* between the values we have adopted and the particular circumstances of our lives: our personality, our culture, our place in society. To flourish, we need to be sensitive to that feedback, and use it in fine-tuning our values to fit.

Millgram's own discussion focuses on large-scale value shifts which accompany things like, say, having a midlife crisis and changing careers. But his argument leaves room for smaller-scale adjustments in the articulations of our values. Say I start rock climbing and take up the most obvious standard of success in that hobby. There is a generally agreed upon difficulty scale for climbs; most new climbers just start by trying to advance up that scale. Some climbers flourish under that goal; others do not. When I focused on advancing on the difficulty scale, I found myself miserable, tormented by my sense of inadequacy and my inability to progress. My climbing days were filled with exhaustion and dread. So, I began to change my sense of my goal in my climbing, started pursuing a slightly more personal vision. I started looking for the most elegant,

[17] Bowker and Star, *Sorting Things Out*.

[18] Millgram, *Practical Induction*. Millgram has since developed some of the ideas in *Practical Induction* further, most notably in an argument that boredom and disengagement is a signal that one's values and chosen roles are a bad fit ("On Being Bored out of Your Mind"). My treatment here relies on both of his discussions.

interesting climbs, and my goal became to climb them as delicately and with as much control as I could. And under that goal, I flourished: I became more constantly sensitized to the details of my movement and more at peace with simply enjoying a bit of lovely climbing; rock climbing trips now left me feeling restored and happy.

So here is a first pass at the *tailoring argument* for the harm of institutional value capture. If Millgram is right, then we will flourish when we have the capacity to adjust and tailor our values in light of our rich experience of the world living under them. When we tailor our values to ourselves in light of those rich experiences, then our values will be better fit to promote our flourishing, as the very specific people we are, in our very specific circumstances.

Perhaps you do not like the references to some ill-defined sense of "human flourishing" or "well-being." We can put the same thought in less mysterious terms. When we adjust our values in light of our rich emotional experience of the world, then those adjusted values will be better suited to support a more emotionally positive life. If we adjust our values, taking interest and engagement as a positive sign and boredom and ennui as a negative sign, then our values are more likely to give us a rich, interested, and engaged life, rather than a bored and listless one. But in institutional value capture, we do not adjust our values in light of our particular experiences. We take values as provided by some large-scale institution and live under them as given. Those values will have been formulated to take deeply into account various institutional interests: like the ability to be counted in a reliable way across a large institution and the ability to be readily aggregated in an institutional bureaucracy. They will not have been formulated in light of the rich feedback of how our particular lives have gone when we live under these values. In value capture, we adopt values that have been formulated in a way that is *insensitive to* and therefore *less able to support* our rich, subtle, and personal emotional experiences.

The first pass emphasized the problems of value capture for the individual. This reflects Millgram's own version of the argument, which emphasized individual values and individual phenomenology. But we can also easily extend the argument to encompass group value capture. Groups, too, have particular articulations of their values.[19] This can look like, say, the community of analytic philosophers' value in rigor or the creative writing community's value in personal expression.[20] The contemporary community of improv comedians, for example, have come to put a strong value in collaboration via automatic agreement.

[19] See Nguyen, "Monuments as Commitments," for a further discussion of group valuing, and how public art may function as an alternative to metrics for expressing group values.

[20] For an excellent discussion of the scientific community's interest in enlarging the collective data supply, see Strevens, *Knowledge Machine*.

The core rule is "Yes, and . . .": you always accept other people's suggestions and build on them. This expresses a value, we might say, of *radical acceptance*—of never refusing ideas, and always integrating every proposal and building upon it. This value works extremely well in the context of improv comedy. And the precise articulation of this value has obviously evolved over the years through trial and error in countless acts of improv comedy. But its success is context-dependent. (Imagine trying to center such a value in analytic philosophy. Analytic philosophy, one might think, is a century-long social experiment in the value of harsh criticism and the radical refusal to accept anything.)

So long as there are accessible signs of a groups' flourishing or of a community's well-being, then Millgram's account of practical induction should also apply to the development of group values. Such group-level value tailoring is unlikely to center internal emotional phenomenology as strongly. But groups can tailor their values in response to their particular nature and context. Jane Jacobs offers a particularly vivid example of how we might tailor a specific value to a specific context.[21] Dwellers in dense urban environments, she says, have learned to cherish *privacy* in a way that suburban and rural people do not. So much of one's life is conducted in dense public environments, that city dwellers have developed a profound devotion to maintaining privacy: of not making unnecessary eye contact, of not intruding into nearby conversations. Valuing a certain kind of eager "friendliness"—easy eye contact, being willing to start conversations with anybody at anytime—makes perfect sense in lower population density areas without that constant press of humanity. But in a dense city, without that collective devotion to the practice of privacy, city dwellers would be utterly overwhelmed by constant social interactions and demands.

To sum up: it is good for agents—individual and group—to tailor their values to their particular context. Those values will be better suited to support the well-being and flourishing of individuals, groups, and communities by being adapted in their formulation to the particular nature of the agents and to their particular context. Value capture interferes with that tailoring. It does so even when the value capture is the result of a fully informed and consensual process since the problem lies in the *content* of the values and not in the bare fact of their voluntary adoption.

Here is another way to put it: value capture, even when consensual, involves a low degree of granular control over the details of the contents of one's value. It puts you in the same relationship with your values as you have with, say, your iPhone's end-user license agreement (EULA). When you clicked to sign a EULA, you did, technically, consent, and you are, technically, responsible. But

[21] Jacobs, *The Death and Life of Great American Cities*.

you only have one binary choice: accept the whole package or not. When we permit ourselves to be value captured by institutional values, we have the same low granularity of control over our values: we either accept the whole package, or not. You cannot get control over how your Fitbit counts steps or how the edifice of higher education counts citation rates and impact factors.

This low granularity arises directly from the core functioning of large-scale collective values. To better understand why, let us look at the processes that drive the creation of institutional metrics.

4. METRICS AND THE STANDARDIZATION OF VALUE

Let us focus on value capture by institutional metrics. I think this is the starkest case of value capture and a good starting point for thinking about other forms of value capture. My goal in this section is to make clear why institutional metrics resist value tailoring as part of their essential functioning.

One response to the case studies I have offered so far—the *USNWR* case and the UN's human rights metrics case—is that the particular metrics are bad. Perhaps there is nothing wrong with value capture per se; it is just that we need to pick good metrics. But I will suggest that we are unlikely to find any institutional metrics that are good to take on as individual or small-group values. Metrics are formulated to serve certain key institutional interests—to work at large scale—and they need to be relatively inflexible to play their role. Institutions want metrics that are narrowly specified, standardized, and inflexible. Precisely what makes a metric good in the institutional context will make it problematic to internalize as a value for individuals and small-scale communities.

Here, we can turn to a rich and useful empirical literature on the place of quantification and standardization in bureaucracy and political life. Here, we are the beneficiaries of decades of empirical study of quantification culture, performed across a number of fields, including history, sociology, anthropology, and communications.[22] What follows may sound familiar to some philosophical ears; the empiricists I will be discussing are often working in a Foucaultian mode. The field, in particular, has been highly influenced by philosophical figures such as Ian Hacking, Bruno Latour, and Martha Nussbaum.

A foundational work here is Theodore Porter's 1995 history of quantification culture, *Trust in Numbers*.[23] Porter is particularly interested in how quantified forms of justifications, like the cost-benefit analysis, came to dominate politics

[22] The study of quantification culture is often associated with the interdisciplinary field called Science and Technology Studies.

[23] Though Porter is a historian, he was significantly influenced by Ian Hacking's work in the philosophy of science on the formation of categories and measures.

and management. He is not arguing that quantification is always bad. Rather, his goal is to get clear on the relative advantages and disadvantages of qualitative and quantitative ways of knowing. Porter argues that qualitative ways of knowing are nuanced and context-sensitive. But qualitative information is difficult to manage en masse and difficult to transfer across contexts. Qualitative evaluations usually require significant shared background knowledge to adequately interpret. When we transform information from a qualitative to a quantitative format, we strip off much of the nuance, texture, and context-sensitivity. By doing so, we create a *portable* package of information, which can be easily sent across contexts and understood by people with little shared background.[24] Quantified evaluations can be easily transmitted between people with little shared background, precisely because they have been stripped of context-dependent features. And quantification isolates the more invariant parts of that information so that the results can be readily aggregated. For this reason, quantitative methods are preferred by large-scale institutions, which must pass information across many levels of hierarchy—between distant administrators with low shared context.[25] In other words, quantifications are preferred in large-scale institutions precisely because of their narrowness and their context-invariant stability.

And quantitative evaluations themselves vary according to their nuance and context sensitivity. Once, land in England was measured in *hides*. A hide is the amount of land required to support the average family. The hide is a measure that highlights a highly relevant functional quality. The acre is a measure of land size, rather than land function. Similarly, says Porter, older Polish land measures varied by soil quality, so a given unit of land would approximately represent a similar productive value.[26] When a ruler attempts a fair distribution, the measure they use will determine which quality is evenly distributed—in this case, land size versus land functionality. Think about the difference between, say, a king's giving each of his soldiers a hide of land, versus his giving them each ten acres of land. One might think that the hide is a superior measure of functional worth, and so a vastly preferable measure for providing fair compensation. But hides are highly variable in size, and determining what counts as a hide requires the application of detailed local knowledge. A hide in a fertile river valley is smaller than a hide in a desert. Hides also vary depending on

[24] This notion of "portability" as the center of quantified information is alive in more contemporary work in this space. In Sabina Leonelli's crucial work on the philosophy of data, she defines "data" as information that has been prepared to travel to new and unexpected contexts ("What Counts as Scientific Data" and *Data-Centric Biology*).

[25] Porter, *Trust in Numbers*, 3–86.

[26] Porter, *Trust in Numbers*, 24.

local weather patterns, game animal migration patterns, and more. The hide is a measure that can really only be effectively managed at the administrative periphery—by locals, who know their environment and its inhabitants' typical needs and usage patterns. The hide is impossible to administrate from any sort of distant bureaucracy. So, says Porter, when we shift from small, local, distributed governance to large-scale centralized governance, we inevitably shift from informationally rich—but difficult to manage—measures like the hide to more standardized, but informationally impoverished, measures like the acre.

James Scott calls this the state's view of the world. By "states," Scott means any large-scale institution, including governments, corporations, and the emergent networked institution of globalized capitalism. States, says Scott, can only manage what they can see, and they can only see that information which has been rendered into a form which can be processed bureaucratically—information that has been standardized and quantified. States can only see those parts of the world which have been rendered *legible* to them.[27]

Student grades provide a familiar example. In the modern educational environment, student grades are almost always quantified. But there are other modes of educational assessment. Imagine an educational environment where we only offered qualitative evaluation of their students' work, like written feedback describing its good qualities and its problems. Such evaluations can easily pivot to address different dimensions—like the writing clarity, the originality, the argumentative clarity—without any demand to compress that all down to a single dimension of evaluation. Such evaluations can also be tailored to each student's own particular goals. I might give very different suggestions to a nursing student interested in the practical implications for their work than I would to, say, a future lawyer or future journalist. If our goal is simply to educate the student, we do not necessarily need to provide an overall rating of all our students on some single common scale.

But in our actual world, we must offer a quantified measure of each student's success—a measure which permits us to instantly compare any student with any other: their grade. This quantified ranking of students is extremely useful to administrators. All of a student's efforts in a class can be expressed in a single number. This also enables a further aggregation: all their class grades can be averaged to generate a single number, which represents their entire educational career—a GPA. And the existence of GPAs is enormously useful for the project of administrating a large-scale educational bureaucracy. They enable all kinds of fast, easy, and objective-seeming manipulations. An admissions officer can arrange the data from every single student application into a spreadsheet and

[27] Scott, *Seeing Like a State*, 11–83.

quickly sort them by GPA. They can create an automatic cutoff point below which student applications are automatically discarded. Sets of student GPAs can be aggregated in order to yield a single number that can be used as a metric of performance for a particular teacher or a whole school district.

In their study of the history of American grading, Jack Schneider and Ethan Hutt argue that standardized grading schemes were implemented to make grades more legible and usable to administrators and employers.[28] Before grading, there was no communicative "shorthand." Evaluations required intimate communication between teacher and student. Early systems of grading were "low-stakes" affairs; they were set up differently in different schools and built to encourage student learning. But the modern system of grading serves not a pedagogical purpose but an organizational purpose. It enables students to easily transfer between different institutions. Perhaps most importantly, it standardizes a product for future consumption on a market. Standardized grades make possible standardized educational certificates, which are extremely useful for potential employers. It was administrators and employers who "placed a premium on readily interpretable and necessarily abstract grading systems."[29] Qualitative evaluations of student might be nuanced and context-sensitive—but they are illegible to the large-scale administrative institution.

Finally, these various procedures—data collection, transformation into standardized inputs, and aggregation—need to be codified into a set of policies that can reliably executed by very different people. Large-scale institutions need to train up people from different backgrounds to perform the same sorts of tasks. And their performance needs to be assessable and auditable by others—where those auditors also come from different contexts, and their audit procedures themselves are subject to the same demands of explicability and transmissibility.[30] These various procedures need to be *standardized*. That means that the inputs and processing rules of these procedures need to be regulated across many contexts.[31]

We can draw from this mess of observations some underlying themes. Institutions share a basic functional interest, inherent to the functioning of large-scale administrative systems. They need to manage information across a vast domain. This need arises intrinsically from the need for an institution to function as a coherent whole. Notice here that I am not presuming that the institution has some interest in controlling or manipulating individuals. Even

[28] Schneider and Hutt, "Making the Grade," 203.

[29] Schneider and Hutt, "Making the Grade," 217–18.

[30] Du Gay, *In Praise of Bureaucracy*.

[31] Bowker and Star, *Sorting Things Out*, 13–16.

the most well-intentioned of organizations—like, say, a charitable nonprofit—has this same functional interest in information management. The interest arises from the basic conditions of coherent group agency, as instantiated in a policy-based, centralized bureaucracy.

This functional interest is served by two standard mechanisms—quantification and standardization. Institutions need to render the world into a format legible to large-scale institutional information processing procedures. So, institutions need information in quantified and standardized format. Because of their institutional function, these mechanisms—quantification and standardization—tend to share some specific features that make them problematic to internalize as personal values. First, quantified metrics are *narrowed* by design. Only certain things count. Institutional measures need to be usable across different contexts. This requires that the measures leave aside highly context-dependent forms of understanding and focus for their inputs on context-invariant qualities. As Scott says, the narrowness of the metric creates a narrowness of institutional vision. Institutions can only see, process, and act on parts of the world that are counted by their metrics. Anything that does not impinge on those metrics is invisible at an institutional level.

In value capture, we internalize those narrowed metrics, thus narrowing our values. And insofar as our values drive our attention, then the value captured will be subject to an analogous effect to Scott's narrowed institutional vision. It is not that we literally do not see things that fall outside our narrowed values, but we will not devote much energy to them or dismiss them as unimportant. Think here of the businessperson who thinks that only money matters and who immediately dismisses from mind any unprofitable ventures—like art or philosophy.[32]

Next, such institutional metrics typically present values in highly explicated, *finished* form. They resist reinterpretation. Preinstitutionalized values are often expressed in an open-ended manner. A concept like "health" or "fitness" or "a good education" admits of different interpretations. Different people may work out their own interpretation of what counts as a good education—and so evaluate their understanding of the term. You want to know more useful things; I want to indulge my sense of curiosity—both are viable understandings of what one might want out of an education. But step counts and law school rankings do not admit of such variability. The method of assessment is rigid. Says Porter, the process of quantification is useful to large-scale institutions, in significant part, precisely because it reformulates information so as to remove the need

[32] I further develop this line of thinking in Nguyen, "Value Collapse," which explores the possibility that overly explicit articulations of value can narrow our attention and exploration.

for interpretation.[33] Standardization is required for informational portability—and standardization requires rigidity.

And those off-the-rack values usually come *embedded* in institutional infrastructures, institutional language, and mass technologies so as to resist further tailoring. We do not have the power to fine-tune the innards of such institutional values. We cannot tinker with the way Twitter counts likes, nor adjust with *USNWR*'s ranking algorithm. They are hardwired into external systems. This rigidity and uncustomizability of measures and metrics is no accident. It is essential to their institutional function. Standardization enables easy communication and ready aggregation—but to do so it must resist individual customization.

To summarize: institutional metrics are designed according to *alien interests*.[34] They have, in fact, been fine-tuned and adjusted—but to satisfy interests that are not our own. I am not presupposing here that institutions must have malevolent intent, like an interest in domination, control, or power. To put Scott's discussion into philosophers' terms: all we need to attribute to an institution is a basic interest in *agency at scale*—an interest in gathering information about the world, managing that information, and using it to inform actions. But the scaled-up nature of bureaucratic institutions imposes certain distinctive requirements on that information-gathering process. Institutional metrics are typically formulated to fit the demands of scaled-up informational agency: for easy recording in institutional memory, for transmission across bureaucratic layers, and for manipulability by institutional methods. When we internalize institutional values, we are letting such interests play a powerful role in the formulation of our own values. Value capture gets us to take an institution's eye view on ourselves—to evaluate ourselves and our activities in institutional terms.

We have much to gain by fine-tuning our values, fitting them with our psychology and world. Institutional metrics are tuned, not to an individual's rich and particular experience of the world or a small community's particular context, but to the needs of information processing at a mass scale. In value capture, by institutional metrics, our values become rigidly tied to an external

[33] Porter, *Trust in Numbers*, 21–29.

[34] Owens and Cribb make a similar point in their analysis of Fitbit technologies ("My Fitbit Thinks I Can Do Better," 32–35). They distinguish, however, between procedural autonomy—which involves internal deliberative processes—and substantive autonomy, which involves one's ability to actually act on and bring to fruition one's decisions. They say that by and large, self-tracking technologies like Fitbit may aid procedural autonomy but cannot aid substantive autonomy since such technologies cannot fix large-scale social inequities. My argument is that such technologies also significantly undermine procedural autonomy.

expression. That rigidity arises, in significant part, from the institution's interest in large-scale informational management. Metrics, by design, resist attempts at digestion and customization by the agent—and they usually come embedded in large-scale institutional infrastructures that make them even more inflexible. Their alienness resists adaptation.[35]

5. The Seductiveness of Metrics

Of course, we often have to use such metrics when we work within, and next to, institutions. But we could use them while also keeping them at emotional arm's length. We could employ them in our reports and our requests for funding, but only as the trade language of bureaucracies. Why might we ever take the further step and internalize them? The answer comes in several stages.

First, quantifications, in and of themselves, are seductive in their clarity and crispness. Many people seem to trust quantified data simply because it is quantified. And we should certainly trust data when it has been generated using reliable methods. However, the mere quantified format itself often seems to generate trust, regardless of quality of the underlying methodology. But obviously, mere presentation in a quantified format does not offer any guarantee of reliability. So, insofar as we trust from the bare fact of quantified presentation, then that trust is unwarranted. And Porter, Merry, and Espeland and Sauder provide evidence aplenty that bare fact of quantification actually does, in fact, generate such unwarranted credibility. To put in the contemporary parlance, the excessive credibility given to quantified data counts as a form of epistemic injustice or epistemic oppression.[36] It harms those who are unwilling or unable to present their information in such quantified form, preventing them from being appropriately recognized as sources of information. And insofar as

[35] One might ask what relationship this view has with various forms of alienation critique. Though my analysis here is obviously similar, in spirit, to the general themes of alienation critique, I avoid use of the term "alienation" because my analysis here differs, in key respects and in many details, from traditional alienation critiques. As Rahel Jaeggi says, many forms of alienation critique involve views of the alienated agent as divided against themselves, as unable to identify with their work, as diffident and depressed (*Alienation*). But the value captured agent can be wholehearted (think of the capitalist all-in for money), fully identified with their work, energized, and motivated. They are not divided against themselves; rather, they are simplified, where that simplification has been guided along institutional lines. Notice, furthermore, the difference between my analysis and the traditional Marxist alienation critique. It is possible to be value captured by a fully socialist bureaucracy. Here, I am aligned more with Scott's particular version of neo-Foucaltian critique than with Marx. For Scott, both globalized capitalism and centralized communism share an interest in rendering the world legible into the terms that they can process and act upon.

[36] Fricker, *Epistemic Injustice*; Dotson, "Conceptualizing Epistemic Oppression."

quantified data tends to emerge from certain sorts of institutions, then those institutions themselves are the beneficiaries of epistemic injustice.

Why might the mere presentation of information in quantified form invite such excess credibility? One familiar suggestion is that numbers carry with them, through their association with the sciences, an aura of authority. I would like to suggest another mechanism: our use of *cognitive fluency*, a phenomenon well-documented by cognitive psychologists. Cognitive fluency is the "subjective experience of ease or difficulty with which we are able to process information."[37] As it turns out, we often use cognitive fluency as an epistemic heuristic. The easier an idea is for us to comprehend, the more likely we are to accept it as true. This is sometimes a useful shortcut. We are typically better at processing information in domains where we have expertise, so ease of comprehension is somewhat correlated with correctness. But the heuristic is far from perfect, as cognitive psychologists have amply demonstrated. First, we seem more willing to accept an idea simply because it is familiar. Second, we are more likely to accept claims presented in a more legible font. But, obviously, the bare fact of repetition or graphic legibility has no direct bearing on truth. In both cases, using a cognitive fluency heuristic results in a mistaken degree of trust.[38]

We should, then, experience a cognitive fluency effect with anything with which we are familiar. And we are extremely familiar with numbers. They are the universal abstraction. Information presented in quantified form thus wears an extremely familiar face. So, the fluency heuristic can lead us astray with quantification, just as it does with fonts. This offers an explanation for the unwarranted credibility of quantified values. Fluency may bring somebody to accept a quantified evaluation of value over a more inchoate one—like accepting the *USNWR*'s clear presentation of a ranking over one's own internal sense of, say, fit with a law school's culture. And insofar as the quantified presentation is more likely to emerge from external and institutional sources, then the fluency effect gives an unwarranted credibility boost to such sources.

But it is not just that metrics are quantified; it is that they are *standardized.* Once our values are standardized, then we can easily explain our actions and justify our decisions to others. Metrics offer an engineered communicability for values. This engineered communicability grants a further credibility advantage to claims made in the evaluative language of those metrics. After all, our ability to make ourselves understood to others can be a sign that our own

[37] Oppenheimer, "The Secret Life of Fluency," 237.

[38] Reber and Unkelbach, "The Epistemic Status of Processing Fluency." I offer a sustained discussion of how cognitive fluency plays into our attraction to seductively clear systems in Nguyen, "The Seductions of Clarity."

understanding is good.[39] And metrics are, by their very nature, easier to understand across contexts. But there is a gap between communicability and epistemic worth, and that gap can be exploited. As Porter makes clear, institutional metrics trade away informational nuance, richness, and contextual sensitivity for the sake of easy portability across communicative contexts. Metrics, then, in virtue of their basic institutional function, also function to precisely exploit the gap between communicability and epistemic worth. When we standardize metrics, we engineer in broad-ranging comprehensibility by removing contextual nuance. Because the very act of removing contextual nuance increases communicability and cross-contextual comprehensibility, metrics—by their essential nature—invite excess trust.

Quantifications can also be seductive because they offer us the pleasures of *value clarity*. When we internalize them, our value landscape becomes simpler and easier to navigate. We are tempted to take them on because they offer us hedonistic rewards in exchange for simplifying our values along certain lines. This line of argument draws on my account of the motivational structure of games, which I have developed at length elsewhere.[40] In games, we take on artificially constructed goals. In ordinary life, our goals and values are often complex and subtle. It is often hard to explain our values clearly, hard to adjudicate conflicts between values, and hard to figure out if we have actually achieved what we value. But in games, values are easy. They are clearly articulated, with explicit criteria for application. In games, we know exactly what we should be doing, and exactly how well we have done. Games offer us a momentary refuge from the nauseating complexity of real-world values. They are an existential balm.

This offers us a second mechanism for the seductiveness of quantification. We can gain a hedonic reward for internalizing simplified values. When we come to value a simplified goal in a non-game activity, we bring the pleasures of value clarity into the real world. Our purposes become clearer, our degree of success becomes more obvious, and our achievements become more readily comprehensible—and it becomes easier to compare and rank our respective achievements. But to get those pleasures, we need to simplify the target. And this helps to explain why it can be so tempting to internalize institutional metrics. Metrics are narrowed and finished. When we internalize such clarified metrics, we can eliminate our struggles with the ambiguity and complexity of our values. Metrics may

[39] I am relying here on the literature from the philosophy of science's investigation of understanding. According to the standard account, one of the signs of understanding is the ability to communicate that understanding to others (Strevens, "No Understanding without Explanation"). I offer a discussion of how engineered simplicity can hijack our sense of understanding in Nguyen, "The Seductions of Clarity."

[40] Nguyen, "Games and the Art of Agency" and *Games*.

not have been explicitly made for gamification, but the institutional pressures on the generation of metrics make them function as pleasingly game-like goals.

And the value clarity effect becomes even more powerful when that clarity is standardized. After all, the existential burden of our complex values is not merely a personal affair; we have to deal with the buzzing tangle of everybody else's values too. Navigating this overwhelming plurality—understanding other people's values and explaining our own—can be grueling. There is, so often, a vast gap between our values. Try explaining to another person your profound love of some weird old comedy, or why a sour cabbage casserole makes you feel so comforted on the bleakest of days. Try explaining why a particularly acid passage of Elizabeth Anscombe's fills you with such glee, or why you never quite got along with running, but rock climbing makes you feel so amazing. Sometimes we can make ourselves understood, but often we cannot. So much of our sense of value arises from our particular experiences—the long life we have led, our twisty paths to self-understanding and world-loving—that explaining the whole mess to others is often beyond our capacities.

But institutionalized values offer us an experience of social value clarity. If an institution offers us a prefabricated metric for some value, and we collectively internalize it, then we will make easy sense to each other. Perhaps that prefabricated metric is citation rates, or Twitter followers, or GPA, or your university's ranking. In any case, once we internalize that value together, much of the existential friction of social life suddenly disappears. Metrics create a common currency for justification. I no longer need to struggle to explain my way of valuing to others, or to understand their way of caring about the world. Justification becomes easy because metrics can function as a preengineered system of aligned value. Metrics offer, not just a personal form of value clarity, but *social value clarity*. When we converge on the same simple, public value system, it becomes so much easier to communicate our values and our justifications. We have gone on the same value standard.

Let us take a step back. Is it some wild accident that institutional metrics turn out to be so seductive? I suspect not, though I can only offer a brief sketch here. Rational agents often need clearly articulated policies to function—including policies about what our goals are and how we are to evaluate our progress toward those goals. Clearly articulated policies ensure reasonably fast decision-making that is consistent over time. As Michael Bratman argues, such policies play an integral part in our being able to maintain coherent agency over time.[41] Policies are desirable for large-scale institutions for similar

[41] Bratman, *Intention, Plans, and Practical Reason* and *Shared Agency*; Holton, *Willing, Wanting, Waiting*; Andreou, "Coping with Procrastination."

reasons, since institutions also need to ensure relatively quick and consistent decision-making across a large and scattered structure, in order to enable cohesive collective action.[42]

But the nature of large institutions requires that we heighten the clarity and explicitness of those policies in order for them to function consistently across the whole. The policies I set for myself can hinge on my own peculiar sensitivities and ways of understanding the world. A coherent policy for me is: "Exercise every day until I start to get that pleasant, warm, cheerful feeling." This works for me because I can consistently recognize that pleasant warm feeling. (Another coherent policy for me is, "Cocktails before 6:00 PM only when I really, really, really need it.") But such policies will not work for large-scale institutions because criteria like "a pleasant, warm, cheerful feeling" cannot be written into institutional policy, nor could they be reliably applied by different people across the institution. Institutional policies need to be hyperexplicated so that they may be executed by a wide variety of people hired from a variety of backgrounds. They need to be, to adapt Porter's language, procedures that are *portable* between many contexts. In order to function in institutions, policies need to be easier to apply—and so they can be appealing to internalize. It is very easy to act clearly and consistently when we adopt such hyperexplicit policies. However, in adopting them, we are giving up on the kinds of policies that hinge on sensitivity to subtle internal phenomena.[43]

6. CONTEXT LOSS

Let us step back and summarize the action so far. Values, I have argued, benefit from being tailored to an agent's particular context. In individual cases, that context can involve all the particular details about the person—their personality, their subtle emotional responses to the world. In group cases, that context can involve details about the particular people who make up the group, or the group's ambient culture. It can involve the kinds of subtle considerations that are only adequately comprehensible to a particular community who have gone through a particular set of struggles together.[44] And in all cases, it involves

[42] This comment relies on the extensive recent literature on group agency, including List and Pettit, *Group Agency*; Gilbert, *Joint Commitment*; and Rovane, *Bounds of Agency*.

[43] For a further discussion of problems with the explicit policies in group agents, see Nguyen, "Games and the Art of Agency."

[44] For an excellent recent overview of standpoint epistemology, see Toole, "Demarginalizing Standpoint Epistemology." For a discussion of the tension between standpoint epistemology and the demands of bureaucratic transparency, see Nguyen, "Transparency Is Surveillance."

particular details of the specific context of the agent: the location, the surrounding culture, the environment.

The argument here is an instrumental one, and epistemic in nature. I can easily imagine very different accounts of the harm of value capture. One might wish to argue that autonomous control over one's value was a good in and of itself and that something was intrinsically wrong with ceding that control via value capture. And this argument might also be a good one. But that is not the argument I am making here. Here, I am arguing that fine control over the expression of one's values is instrumentally good. It promotes well-being and flourishing and other such ways of indicating a good life in individuals and communities. It does so because substantively participating in the process of adjusting and fine-tuning one's own values yields more nicely tailored values. Notice this argument works irrespective of whether or not we conceive of the value capture process as consensual or voluntary. It has to do with how much one substantively tailored that value to one's context. One may have voluntarily undergone value capture, but in so doing, one has withdrawn from the process of finely tailoring one's values to fit.

And fine control leads to better well-being and flourishing for epistemic reasons. Individuals and smaller-scale groups have better access to the details of their specific context. There is a useful analogy here to a discussion about the epistemic value of democracy. One reason that democracy is important, one might think, is that self-determination is an intrinsic good. But another reason that democracy is important is that, when appropriately structured, it is the best way to integrate the *epistemic access* of the governed. This is the epistemic defense of democracy. As Helene Landemore puts it, epistemic democracy functions well when it employs a deliberative process that takes into account the specific details known to the relevant communities. Democratic deliberation, done properly, is sensitive to the special understandings of the deliberating citizens. Importantly, Landemore's argument is not that democratic participation is an intrinsic good or constitutive of an authoritative government. Landemore's argument, rather, is that democratic participation is instrumentally good since it yields laws and policies that better fit the circumstances and take better advantage of the various perspectives, expertises, and understandings of the entire citizenry.[45]

[45] Interestingly, Landemore does not consider the problems of scale (*Democratic Reason*). This is worth a wholly separate discussion, but I can briefly say: Landemore's argument presumes that the process of democratic deliberation will always preserve the knowledge of the participants and aggregate them. I think there is significant reason to be skeptical about that. I have offered some reasons to be skeptical in my discussion of the epistemic problems with public transparency metrics (Nguyen, "Transparency Is Surveillance"). I think

The tailoring argument works similarly. When an agent tailors its own values, this yields an instrumental good for epistemic reasons. The agent has more access to their context—their psychology, culture, the local details—and so can tailor their way to better-fitting values. Value capture involves adopting values from an external source—typically a massive institution. Such an external source has far less fine-grained access to the local details. The large scale at which such an institution operates imposes a specific demand: that the information they use can be transferred easily across very different contexts. The general insight from the empirical work on bureaucracies and metrics is this: the larger the scale, the less the sensitivity to the details of a particular context. As F. A. Hayek puts it, central decision makers cannot serve each particular person, but only the average person.[46] And, I might add, central decision makers cannot serve local communities but only the average community.

7. Value Swamping

You might have started to suspect that there are actually two distinct problems running side by side here: a problem involving externality and a problem involving scale. To disentangle them, let us consider a different phenomenon, right next door to value capture, which will isolate the problems of scale. Consider a case where we actively participate in specifying some shared value—but the efforts of coordination at scale color the formulation of that value. Let us call it value swamping.

Value swamping happens when:

1. An agent's values are rich and subtle (or in the process of developing in that direction).
2. The agent participates in a large-scale social process that yields a specification of shared values.
3. Those specifications of shared values come to dominate the agent's practical reasoning (in the relevant domain).

Landemore underestimates the importance of federalism and local governance because of her unwarranted optimism in the possibility for lossless information aggregation at scale. What the empirical work I have discussed on quantification and bureaucracy—especially Scott's discussion—demonstrates is that information aggregation at scale always leads to massive data loss. This, Scott suggests, is a reason to strongly prefer local governance in most situation.

[46] Hayek, "The Use of Knowledge in Society." I take Hayek to be a major influence on Scott's analysis. In fact, I take Scott's analysis to be offering a synthesis of Marxist criticisms of capitalism with Hayekian criticisms of central planning.

Value swamping is just like value capture, except for step 2. In value swamping, the agent does not get their values from some wholly external source; instead, they participate in a large-scale social process that yields a specification of a value.

Here is an example, slightly fictionalized from my own life. I took part in a large effort, across the humanities departments of my university, to defend the humanities from constant budgetary incursions from the STEM departments and the business school. We wanted to help the humanities survive and thrive. We ended up in a long discussion about realistic goals, and we decided that we most wanted to push for increasing the number of lines of humanities faculty and increasing faculty diversity. We ended up settling on some targets: we were going to push for a fast increase in the total number of humanities lines by 5 percent and embark on a long-term project to increase the representation of people of color in the faculty by 20 percent.

We needed such clear targets—and such a small number of them—because we needed some specific demands to bring to the upper administration. We also needed highly legible targets—the kind of targets that could be coherently targeted and tracked over the coming years by a revolving set of faculty representatives. Notice what is not on the list, however. I would have loved to push for creative work in new hires and for diversity in intellectual interests—and not just race. But how could we track such fuzzy, inchoate targets over the years? It might be that every humanities faculty cares about "creativity," but since we lacked a readily accessible and scalable measure of creativity, we cannot easily make it a group target.

What happened? Bowker and Star say that any attempt at large-scale collective action creates a demand for cross-contextual informational categories and for data that is readily aggregable.[47] This, in turn, creates a demand for publicly accessible, standardized procedures of measurement, such as metrics. Notice that the pressure for standardization here does not arise from the external sourcing of the metrics but from the demands of large-scale collective action itself. In other words, for some of the pressures, *it does not matter if the metrics are generated by an external source*.

Suppose that our case of value swamping is ideally participatory. Still, as the size of the relevant community scales up, the values that get generated are more and more subject to the demands of cross-contextual communication and consensus. Value swamping admits of more tailoring than value capture. Since a group's values are generated by the group itself, they can still be moderately tailored to the group's experiences. But there are still formidable constraints

[47] Bowker and Star, *Sorting Things Out*, 53–161.

on the kinds of values the group can use. The group can only adopt the sorts of values that can be understood in all of the varying contexts across which the group operates. The processes of context stripping and de-nuancing are problems of scale, not of externality. In value swamping, the state-level interests, in information that is aggregable and portable across contexts, are not the alien impositions of an external force. They are necessitated by the process that we signed onto, and for very good reasons. But they do pull us toward less nuance in the specifications of our values as part of the drive for larger-scale cooperative action.

What we have learned is that there are two problems that can lead to badly tailored values. The first problem is that the values are generated by an external source. The second problem is the values are subject to the pressures of scale. In value capture by institutional metrics, we are exposed to both problems: externality and scale. In value swamping, the values may be our own, but they are still subject to the demands of scale.

Something is still lost in value swamping cases. But we cannot quite say that the values are not our own. We consented, we participated, we actively formulated, and we approve of the outcome. What is going on here is not exactly outsourcing. But we are sending our values out for processing at a larger scale and getting them back *filtered*. What comes back is what can survive the large-scale deliberative process intact; the more private, intimate, or small-scale communal reasons get filtered out.

8. THE SCALE PROBLEM OF VALUES

This suggests a larger picture. What we are starting to expose here is an essential problem with group agency at scale—or at least, a deep tension between smaller-scale agents and the demands of larger-scale agency.

We have lots of reasons to participate in large-scale collective efforts, and some efforts are far more effective when scaled up. Some things are best pursued collectively: reducing carbon emissions, increasing vaccination rates. In many cases, we can pursue those targets most efficiently by agreeing on a precise and shared specification of that target. In those cases, the upsides of having a precise, stable, shared specification of value may outweigh the cost.

But when we scale up our target-setting process, we lose sensitivity, contextual nuance, and granularity. And I take figures like Porter, Scott, and Bowker and Starr to have shown that this is no accident; it is an inevitable cost of scaling up organization for beings like us, using the methodologies we are presently using for informational aggregation. When we need to achieve agreement across a vast scale—across people who do not share all the same context, who do not share the sensitivities—then what we can agree on will need to be subject to that filter of low-context comprehensibility. And insofar as collective efforts require some kind of shared stability, then these collective values will, of necessity, not admit of tailoring to small-scale or individual contexts.

The tensions between the value swamping and contextual tailoring are not the result of some sloppy process of coordination. They are baked into our very nature as limited beings with varying personalities and contexts who need to coordinate our actions across those contexts. They arise, in particular, from one feature of our finitude: that we each have a special understanding of our own patch and our particular context and a far weaker grasp of distant contexts. So, any attempt to render anything comprehensible across scale involves eliminating those details that require special understanding or contextual sensitivity. The tension between small-scale and large-scale valuing is ineliminable; our lives as both individual and social beings will always involve some kind of tension between our small-scale and large-scale commitments. We are the kind of beings that are perpetually stuck in a painful compromise—between the intimacy of small-scale understanding and the de-contextualized comprehensibility demanded of large-scale shared understanding.

None of this shows that we should not scale up our activities sometimes. Some things are best pursued collectively, as shared projects on the largest scales: reducing carbon emissions, increasing vaccination rates. And the demands of large-scale organizations clearly require clear, legible targets. But there is a cost to scaling up. Some goals—like stopping climate change—are worth the cost. And in other cases, we care more about the goods of local tailoring than the goods of large-scale cooperation. I can see plenty of good that can come from collectively pursuing a clear target that we all understand in the case of climate change or public vaccination. It is much harder to see the goods that come from collectively pursuing the same specification of, say, values in fitness, or musical values, or values in family organization. Some things are best managed at a personal or local scale. There is a trade-off between collective coordination and local specificity—and we may want to make that trade-off quite differently in different domains.

Here is an analogy: in law, we want federalism. That is, we want some of our laws set at the national level, some at the state level, and some at the county or city level. And the explanation is that for some kinds of laws, it is better to coordinate across a vast realm because the goods of standardization and sameness are worth the cost of low local tailoring. And for other kinds of laws, it is best to let them be set at smaller and smaller scales.

What this suggests is that we should want *value federalism*. Some values are perhaps best pursued at the largest-scale level, some at smaller community levels, and some individually. And the upshot here is not that we should reject all large-scale values. It is that we should maintain a variety of differently-scaled values. There are many cases in which it might be useful to participate in a larger collective effort and so to accept, as part of that collective effort, less finely tailored goals. But, at the same time, we can confine those large-scale, standardized goals to our life inside those collectives and not let them swamp the rest of our values. The problem occurs when we exhibit an excess preference for the largest-scale values and let the largest-scale values swamp too many of our smaller-scale values. The problem comes when we let the demand for large-scale legibility intrude into every aspect of our lives, even the most intimate ones.[48]

University of Utah
c.thi.nguyen@utah.edu

REFERENCES

Andreou, Chrisoula. "Coping with Procrastination." In *The Thief of Time: Philosophical Essays on Procrastination*, edited by Chrisoula Andreou and Mark D. White, 106–15. New York: Oxford University Press, 2010.

Bowker, Geoffrey C., and Susan Leigh Star. *Sorting Things Out: Classification and Its Consequences.* Rev. ed. Cambridge, MA: MIT Press, 2000.

Bratman, Michael E. *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press, 1987.

———. *Shared Agency: A Planning Theory of Acting Together*. New York: Oxford University Press, 2014.

Brewer, Talbot. *The Retrieval of Ethics*. Oxford: Oxford University Press, 2009.

Buss, Sarah. "Valuing Autonomy and Respecting Persons: Manipulation, Seduction, and the Basis of Moral Constraints." *Ethics* 115, no. 2 (January 2005): 195–235.

[48] This paper has had a long, tortured, and meandering history. I owe an enormous debt of gratitude to an enormous number of people, and I am sure that more people have been deeply helpful than I have managed to recall and record. But as a start, I would like to thank Natalie Ashton, John Basl, Peggy Battin, Sarah Buss, Elizabeth Camp, Elizabeth Cantamalessa, Mercy Corredor, Steven Diggin, Steve Downes, John Dyck, Melinda Fagan, Will Fleischer, Carolina Flores, Rachel Elizabeth Fraser, August Gorman, Joyce Havstad, Max Hayward, Aaron James, Paul Katsafanas, Tim Kenyon, Eliot Michaelson, Elijah Millgram, Paul De Font Reaulx, Ronnie Sandler, Carlos Santana, Alexander Sarch, Karl Schafer, Jonah Schupbach, Nick Stang, Dustin Stokes, Olufemi Taiwo, Greta Turnbull, Elise Woodard, and the reviewers and editors of *JESP*.

Callard, Agnes. *Aspiration: The Agency of Becoming*. New York: Oxford University Press, 2018.

Christman, John. "Relational Autonomy, Liberal Individualism, and the Social Constitution of Selves." *Philosophical Studies* 117, nos. 1–2 (January 2004): 143–64.

Dotson, Kristie. "Conceptualizing Epistemic Oppression." *Social Epistemology* 28, no. 2 (2014): 115–38.

du Gay, Paul. *In Praise of Bureaucracy: Weber, Organization, Ethics*. London: SAGE, 2000.

Espeland, Wendy Nelson, and Michael Sauder. *Engines of Anxiety: Academic Rankings, Reputation, and Accountability.* New York: Russell Sage Foundation, 2016.

Fricker, Miranda. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Clarendon Press, 2007.

Gilbert, Margaret. *Joint Commitment: How We Make the Social World*. New York: Oxford University Press, 2013.

Hayek, F. A. "The Use of Knowledge in Society." *American Economic Review* 35, no. 4 (September 1945): 519–30.

Hedahl, Marcus, and Bryce Huebner. "Sharing Values." *Southern Journal of Philosophy* 56, no. 2 (June 2018): 240–72.

Holton, Richard. *Willing, Wanting, Waiting*. Oxford: Oxford University Press, 2009.

Jacobs, Jane. *The Death and Life of Great American Cities*. New York: Random House, 1961.

Jaeggi, Rahel. *Alienation*. Edited by Frederick Neuhouser. Translated by Frederick Newhouser and Alan Smith. New York: Columbia University Press, 2014.

Khader, Serene J. "The Feminist Case against Relational Autonomy." *Journal of Moral Philosophy* 17, no. 5 (October 2020): 499–526.

Krueger, Joel. "Music as Affective Scaffolding." In *Music and Consciousness 2: Worlds, Practices, Modalities*, edited by Ruth Herbert, David Clarke, and Eric Clarke, 48–63. Oxford: Oxford University Press, 2019.

Landemore, Hélène. *Democratic Reason: Politics, Collective Intelligence, and the Rule of Many*. Princeton: Princeton University Press, 2012.

Leonelli, Sabina. *Data-Centric Biology: A Philosophical Study*. Chicago: University of Chicago Press, 2016.

———. "What Counts as Scientific Data? A Relational Framework." *Philosophy of Science* 82, no. 5 (December 2015): 810–21.

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

McGonigal, Jane. *Reality Is Broken: Why Games Make Us Better and How They Can Change the World*. London: Penguin Books, 2011.

Merry, Sally Engle. *The Seductions of Quantification: Measuring Human Rights, Gender Violence, and Sex Trafficking*. Chicago: University of Chicago Press, 2016.

Millgram, Elijah. "On Being Bored out of Your Mind." *Proceedings of the Aristotelian Society* 104, no. 1 (June 2004): 165–86.

———. *Practical Induction*. Cambridge, MA: Harvard University Press, 1997.

Nguyen, C. Thi. "The Arts of Action." *Philosopher's Imprint* 20, no. 14 (May 2020): 1–27.

———. *Games: Agency as Art*. Oxford: Oxford University Press, 2020.

———. "Games and the Art of Agency." *Philosophical Review* 128, no. 4 (October 2019): 423–62.

———. "How Twitter Gamifies Communication." In *Applied Epistemology*, edited by Jennifer Lackey, 410–36. Oxford: Oxford University Press, 2021.

———."Monuments as Commitments: How Art Speaks to Groups and How Groups Think in Art." *Pacific Philosophical Quarterly* 100, no. 4 (December 2019): 971–94.

———. "The Seductions of Clarity." *Royal Institute of Philosophy Supplements* 89 (May 2021): 227–55.

———. "Transparency Is Surveillance." *Philosophy and Phenomenological Research* 105, no. 2 (September 2022): 331–61.

———. "Trust and Sincerity in Art." *Ergo* 8, no. 2 (2021): 21–53.

———. "Value Collapse." *Royal Institute of Philosophy Supplements* (forthcoming).

Nguyen, C. Thi, and Matthew Strohl. "Cultural Appropriation and the Intimacy of Groups." *Philosophical Studies* 176, no. 4 (January 2019): 981–1002.

Oppenheimer, Daniel M. "The Secret Life of Fluency." *Trends in Cognitive Sciences* 12, no. 6 (June 2008), 237–41.

Owens, John, and Alan Cribb. "'My Fitbit Thinks I Can Do Better!' Do Health Promoting Wearable Technologies Support Personal Autonomy?" *Philosophy and Technology* 32, no. 1 (March 2019): 23–38.

Porter, Theodore M. *Trust in Numbers: The Pursuit of Objectivity and Public Life*. Princeton: Princeton University Press, 1995.

Reber, Rolf, and Christian Unkelbach. "The Epistemic Status of Processing Fluency as Source for Judgments of Truth." *Review of Philosophy and Psychology* 1, no. 4 (December 2010): 563–81.

Rovane, Carol. *The Bounds of Agency: An Essay in Revisionary Metaphysics.* Princeton: Princeton University Press, 1997.

Schneider, Jack, and Ethan Hutt. "Making the Grade: A History of the A–F

Marking Scheme." *Journal of Curriculum Studies* 46, no. 2 (2013): 201–24.

Scott, James C. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. New Haven, CT: Yale University Press, 1998.

Sterelney, Kim. "Minds: Extended or Scaffolded?" *Phenomenology and the Cognitive Sciences* 9, no. 4 (December 2010): 465–81.

Strevens, Michael. *The Knowledge Machine: How Irrationality Created Modern Science.* New York: Liveright, 2020.

———. "No Understanding without Explanation." *Studies in History and Philosophy of Science* 44, no. 3 (September 2013): 510–15.

Superson, Anita. "The Deferential Wife Revisited: Agency and Moral Responsibility." *Hypatia* 25, no. 2 (Spring 2010): 253–75.

———. "Deformed Desires and Informed Desire Tests." *Hypatia* 20, no. 4 (Fall 2005): 109–26.

Toole, Briana. "Demarginalizing Standpoint Epistemology." *Episteme* 19, no. 1 (March 2022): 47–65.

Westlund, Andrea C. "Selflessness and Responsibility for Self: Is Deference Compatible with Autonomy?" *Philosophical Review* 112, no. 4 (October 2003): 483–523.

Journal of Ethics and Social Philosophy
VOL. 27, NO. 3 · MAY 2024

https://doi.org/10.26556/jesp.v27i3.3092

PROBABILITY, NORMALCY, AND THE RIGHT AGAINST RISK IMPOSITION

Martin Smith

MANY PHILOSOPHERS accept that, as well as having a right that others not harm us, we also have a right that others not subject us to a *risk* of harm. And yet, when we attempt to spell out precisely what this "right against risk imposition" involves, we encounter a series of notorious puzzles. Existing attempts to deal with these puzzles have tended to focus on the nature of *rights*—but I propose an approach that focuses instead on the nature of *risk*. The key move is to distinguish two different ways in which to conceptualize the risk that a given action presents—one of which is linked to the notion of *probability* and the other to the notion of *normalcy*.

1. THE RISK THESIS AND THE HIGH RISK THESIS

Consider the following case of "pure" risk imposition.[1] Suppose *A* plays Russian roulette on *B*. That is, suppose *A* takes a revolver, inserts a bullet into one chamber, spins the cylinder, aims at *B*'s head, and pulls the trigger. Suppose that, as it happens, the chamber that rotates into the firing position when the trigger is pulled is empty, and the gun does not discharge. Suppose, further, that *B* is asleep or otherwise unaware of what is happening and, as a result, experiences no fear or distress.

We can all agree that, other things being equal, *A*'s action is morally impermissible. But surely *B* would not merely regard this as an "impermissible" action—an action that *A* "ought not to have performed"—he would see this as a *violation of his rights*. And *A*'s action has many of the telltale signs of a rights infringement: *A*'s action could never be justified on purely hedonistic grounds—it could never be justified on the grounds that it would bring *pleasure* to *A* or to others, irrespective of the amount of pleasure that might be derived. Given the opportunity, *B*, or a third party, would have been morally permitted to use force—even extreme force—against *A* in order to prevent him from undertaking the action. Finally, in the absence of a strong reason or excuse, it

[1] Thomson, "Imposing Risks," 126.

would be legitimate for *A* to be punished—perhaps even severely punished—for his conduct.[2]

But what right of *B*'s could have been infringed by *A*'s action? It is natural to think that we each have a right not to be harmed by others.[3] However, by stipulation, *B* has not suffered any actual harm at the hands of *A* (that is why the risk imposition is described as "pure").[4] Perhaps the most obvious suggestion is that, in addition to the right that others not harm us, we also have a right that others not subject us to a *risk* of harm—and this is the right that *A* infringes. Call this the *Risk Thesis*.[5]

Although it provides a straightforward treatment of this example, the Risk Thesis faces an immediate problem—many of the ordinary activities we engage in every day will impose *some* risk of harm on others. If, for instance, *A* drops a piece of bread into his toaster and presses down the lever, there is some risk that this could cause a fire in which his neighbor *B* dies.[6] But presumably, *B* has

[2] On the permissibility of preventive force and punishment in this kind of case, see Bergelson, "Self-Defense and Risks," sec. 3; Thomson, "Some Questions about Government Regulation of Behavior," sec. 4. Many theorists agree that one of the crucial roles of a right is to legitimize defensive and punitive actions when the right is threatened or infringed (see, for instance, McKerlie, "Rights and Risk," 241–42; Thomson, "Some Questions about Government Regulation of Behavior," sec. 2, and *The Realm of Rights*, 2 and ch. 14, sec. 5). If it is legitimate for an action to be punished by the state or for a person to use force to prevent it, this is treated here as defeasible evidence that the action infringes the rights of another. Furthermore, many theorists have claimed that rights serve to *trump* certain justifications for action—including justifications that cite personal pleasure or preference (Dworkin, *Taking Rights Seriously*, ch. 4, sec. 3, and ch. 7; Nozick, *Anarchy, State, and Utopia*, ch. 3; Railton, "Locke, Stock, and Peril," 189). If an action could never be justified on these grounds, then this will also be taken as a defeasible indication of a rights infringement.

[3] If "harm" is construed broadly, then there may be certain harms that one can inflict on others without infringing their rights (see, for instance, Thomson, "Some Questions about Government Regulation of Behavior," sec. 2). I put this issue to one side here—"harm" in the main text can be read as restricted to physical injury and death.

[4] Some have argued that even pure risk impositions constitute harms on the grounds that they frustrate one's interests or diminish one's autonomy. (See Finkelstein, "Is Risk a Harm?"; Oberdiek, "The Moral Significance of Risking." For related discussion, see Rowe, "Can a Risk of Harm Itself Be a Harm?"; Thomson, "Some Questions about Government Regulation of Behavior," sec. 3, and *The Realm of Rights*, 244.) This opens up a different way of thinking about a right against risk imposition—and offers the potential of subsuming such a right within a broader right not to be harmed. While this view would require us to reformulate the problems that I will consider, it does not, as far as I can tell, offer any immediate solutions.

[5] See Holm, "A Right against Risk Imposition and the Problem of Paralysis," 918; McCarthy, "Rights, Explanation, and Risks," 208; Song, "Rights against High-Level Risk Impositions," 765; Thomson, *The Realm of Rights*, 243.

[6] *B* will, of course, be subject to a base-level risk of dying in a fire even if *A* does not make toast. The risk that *A* imposes is that of *B* dying in a fire as a result of, or in a way that is

no right that *A* refrain from making toast. The risk that *B* would die as a result of *A* making toast is, of course, very low—at least in the order of one in billions—and this suggests an obvious fix: perhaps our right against risk imposition only applies to risks that are relatively *high* or significant, like the risk imposed by Russian roulette. The idea, more precisely, is that we have a right that others not subject us to a high risk of harm—that others not act in such a way that the risk of our being harmed as a result of their action exceeds a threshold *t*. This is sometimes referred to as the *Threshold Risk Thesis* or *High Risk Thesis*.[7]

While it may appear more promising than the original Risk Thesis, the High Risk Thesis faces at least two problems of its own. The first problem concerns what we might call cases of "low-risk" Russian roulette.[8] Suppose a bullet is placed in a single chamber of one out of a *set* of otherwise empty revolvers. Suppose *A* chooses a revolver at random, spins the cylinder, aims at *B*'s head, and pulls the trigger. The larger the set of revolvers, the lower the risk of harm that *A* imposes on *B*. If the set were sufficiently large, the risk could be lower than any positive threshold and could even be lower than the risk imposed by making toast. Suppose, for argument's sake, that we set the threshold at one in five hundred thousand. In this case, if *A* chooses from, say, one hundred thousand revolvers when playing Russian roulette on *B*, his actions would not infringe the right posited by the High Risk Thesis. And yet, *A*'s action still has all the hallmarks of a rights infringement. In spite of the number of revolvers, *A*'s behavior could never be justified on the grounds that *A*, or others, find it enjoyable. In spite of the number of revolvers, it would be permissible to use force to prevent *A*'s action and legitimate for *A* to be punished if he has no strong reason or excuse (I am inclined to think that even extreme force and severe punishment could still be warranted).

The second problem for the High Risk Thesis concerns cases of "distributed risk"—cases in which there is a high risk that some member of a group will be harmed, even though the risk to each individual member is low. Consider the

caused by, *A* making toast. The general point that many day-to-day activities impose risks of harm on others is familiar in discussions of the ethics of risk imposition. See, for instance, Fried, *An Anatomy of Values*, 192–93; Hayenhjelm and Wolff, "The Moral Problem of Risk Impositions"; Holm, "A Right against Risk Imposition and the Problem of Paralysis"; McCarthy, "Rights, Explanation, and Risks"; Railton, "Locke, Stock, and Peril," 207; Song, "Rights against High-Level Risk Impositions"; Thomson, *The Realm of Rights*, ch. 9, sec. 6.

[7] See Holm, "A Right against Risk Imposition and the Problem of Paralysis," 920; McCarthy, "Rights, Explanation, and Risks," 212; Song, "Rights against High-Level Risk Impositions," 767; Thomson, *The Realm of Rights*, 245.

[8] Nozick, *Anarchy, State, and Utopia*, 73; Thomson, "Some Questions about Government Regulation of Behavior," sec. 4.

following example due to McCarthy.[9] Suppose *A* is considering two options for disposing of a large quantity of a toxic chemical. First, he could surreptitiously dump the chemical into a pond that he shares with his neighbor *B*. Second, he could surreptitiously dump the chemical into the river that flows through his property, even though there are a million people who live downstream. The former option involves a high risk—say a one-in-a-thousand chance—that *B* will be exposed to a harmful quantity of the chemical. The latter option involves a *very* high risk that at least one of the people living downstream will be exposed to a harmful quantity of the chemical, even though the risk to each *individual* person is low—say, a one-in-a-million chance.

If a one-in-a-thousand chance is above the threshold, then, according to the High Risk Thesis, *B*'s rights would be infringed if *A* dumped the chemical in the pond. If a one-in-a-million chance is below the threshold, then, as far as the High Risk Thesis is concerned, there is no person whose rights would be infringed if *A* dumped the chemical in the river. All else equal, then, dumping the chemical in the river would be the morally preferable option, as this involves no rights infringements. But this prediction seems incorrect—after all, dumping the chemical in the river involves a much higher *overall* risk of harm. If there are a million people living downstream who are each subjected to a one-in-a-million risk of harm, then, assuming these risks are independent, the risk of at least one person being harmed works out to approximately 64 percent. The High Risk Thesis appears, then, to create a dubious moral preference for cases in which risk is distributed among a group of individuals (the river option) over cases in which risk is imposed upon a single individual (the pond option).[10]

2. Revisiting the Risk Thesis

In response to these problems, McCarthy abandons the High Risk Thesis and advocates a return to the original Risk Thesis, on which *any* risk imposition, no matter how slight, constitutes a rights infringement.[11] As McCarthy observes,

[9] McCarthy, "Rights, Explanation, and Risks," 213–14. The example is adapted from McKerlie, "Rights and Risk," 247–48. The fact that cases of distributed risk pose a potential problem for the High-Risk Thesis is observed by Railton, "Locke, Stock, and Peril," 209–10.

[10] While my focus here is on the problem of spelling out a right against risk imposition, it is worth noting that *any* approach to the ethics of risk imposition, whether or not it assigns a prominent role to such a right, will still need to thread its way through cases of Russian roulette vs. low-risk Russian roulette vs. day-to-day risk imposition and through cases of distributed vs. non-distributed risk. While other approaches are beyond the scope of the discussion here, I will mention some in passing (notes 21 and 32 below).

[11] McCarthy, "Rights, Explanation, and Risks."

the Risk Thesis offers a more satisfactory treatment of the toxic chemical case. The Risk Thesis predicts that dumping the chemical in the river would involve a million rights infringements while dumping it in the pond would involve only one, leading to an immediate reversal of the above verdict; all else equal, it is the river option that would now be reckoned to be morally worse. More generally, the Risk Thesis predicts that a risk imposition cannot be made more morally acceptable by distributing the risk among a number of individuals—on the contrary, this will simply introduce further rights infringements.

The Risk Thesis still faces a basic problem, of course. As discussed above, it predicts that many of our day-to-day actions will infringe others' rights. According to McCarthy, though, this result is only problematic if it leads to the conclusion that many of our day-to-day actions are *morally impermissible*. We can infer the latter from the former if we assume that rights are *absolute* and can never be permissibly infringed—but, according to McCarthy, this view is untenable.[12] Suppose I suddenly fall ill and you possess a large quantity of a drug that I need to save my life. While I have the opportunity to take the required amount from your stockpile, the situation is so urgent that I do not have time to seek your permission or to procure the drug elsewhere. If I take the drug, then I infringe your rights and do so knowingly—after all, I know that the drugs belong to you, and you have a right that others not take them without your permission. Nevertheless, if this is the only way in which I can save my life, it seems that my action is morally permitted.

In McCarthy's view, an action that infringes the rights of another will be morally permissible if the reasons in favor of performing it sufficiently outweigh the burden to the bearer of the right.[13] So even if many of our day-to-day actions infringe the rights of others, as the Risk Thesis implies, these actions may yet be permissible, provided they are backed by sufficiently strong reasons.

[12] McCarthy, "Rights, Explanation, and Risks," sec. 3. Nozick comes close to endorsing an absolutism about rights, arguing that it is impermissible to infringe a person's rights even if one could reduce the total number of rights infringements thereby (Nozick, *Anarchy, State, and Utopia*, 28–30). Even Nozick allows, however, that it may be permissible to infringe another's rights in order to avert "catastrophic moral horror." For discussion and criticism of Nozick's near-absolutism, see Thomson, "Some Ruminations on Rights." The existence of some absolute rights is defended by Gerwith, "Are There Any Absolute Rights?"

[13] See McCarthy, "Rights, Explanation, and Risks," 209–10; see also Thomson, "Some Ruminations on Rights" and *Realm of Rights*, ch. 6. On one terminology (Gerwith, "Are There Any Absolute Rights?"; Thomson, "Some Ruminations on Rights"), a right is *violated* just in case it is impermissibly infringed—infringed without sufficient justification. We might say, then, that my taking the drug from your stockpile would, under the circumstances, constitute an infringement, but not a violation, of your rights. Absolutism, in this terminology, can be expressed by saying that all infringements are violations. For discussion of the violation/infringement distinction, see Oberdiek, "Lost in Moral Space."

But *are* they backed by sufficiently strong reasons? Think again of the toast example. By making toast, *A* imposes a low risk of injury or death upon his neighbors and, according to the Risk Thesis, he thereby infringes their rights. And yet, the reasons in favor of making toast are, by and large, pretty trivial, so in this case, if we are to have the desired result that the activity is morally permissible, then these rights infringements would have to be *more trivial still*. But there is something jarring about the idea that another person's rights could count for so little. This is sometimes referred to as the "cheapening of rights" problem.[14] One could perfectly well reject absolutism about rights—perhaps on the strength of examples like McCarthy's drug case—and still insist that a rights infringement could never be made permissible by something like a *desire for toast*. That is, one could insist that a desire for a piece of toast (rather than a slice of bread) is not the kind of thing that could ever sufficiently outweigh the burden of having a right infringed.

Here is another way to put the worry: while absolutism about rights will take us from the premise that many of our day-to-day actions infringe the rights of others to the conclusion that many of our day-to-day actions are morally impermissible, it is not the *only* way to bridge this gap. One supposition I have been taking for granted so far is that an act that infringes another person's rights can never be justified solely on the grounds that it will bring *pleasure* to oneself or to others.[15] This is clearly much weaker than absolutism about rights—and is perfectly consistent with McCarthy's preferred verdict about the drugs case—but it is *inconsistent* with the idea that many of our day-to-day actions permissibly infringe others' rights. Many of our day-to-day actions (like making toast) have no discernible benefit other than to bring some small pleasure.[16] In any case, I will not pursue this further here—for McCarthy's defense of the Risk Thesis faces another, perhaps even more serious, objection.

[14] See Song, "Rights against High-Level Risk Impositions."

[15] There are a number of different views as to what rights are and what kind of moral significance they carry—but many would agree that rights must be something *more* than just one further ingredient in the balance of considerations that bear upon the moral permissibility of an action. As observed in note 2, many theorists endorse the metaphor of rights as "trumps," which would seem to require, at a minimum, that there be some considerations that can count in favor of an action but that could never outweigh or counterbalance a rights infringement. This, at the very least, illustrates that there is a broad conceptual space between absolutism about rights and the view that rights infringements can, in principle, be justified by any action-favoring considerations whatsoever.

[16] Even those who grant that a rights infringement can be weighed against goods such as pleasure may still deny that it could ever be justified by a small or trifling pleasure. Thomson (*The Realm of Rights*, 153n2) resists the metaphor of rights as trumps but suggests that they might still be considered "high cards."

3. The Role of Intentions?

Although the Risk Thesis appears to give the correct verdict in the toxic chemical case, the low-risk Russian roulette case continues to pose a problem—though somewhat subtler than the problem it poses for the High Risk Thesis. If *A* plays low-risk Russian roulette on *B*, then the Risk Thesis, unlike the High Risk Thesis, will straightforwardly predict that *A* infringes *B*'s rights. However, given that *A* *also* infringes *B*'s rights when he makes toast, we are still in need of some explanation of the blatant moral difference between the two actions. As mentioned above, even if the only benefit of making toast is to bring some small pleasure to *A*, the action is clearly permissible. But it is clearly *impermissible* for *A* to play low-risk Russian roulette on *B*, no matter the pleasure he might derive by doing so. As discussed, it would be permissible for one to use force to prevent *A* from playing low-risk Russian roulette on *B* and, absent a strong reason or excuse, legitimate for *A* to be punished for such an action. Obviously, one cannot legitimately punish *A* for making toast or permissibly use force to prevent him from doing so.

Picking up on a suggestion from Thomson, McCarthy proposes that the moral difference between these two actions lies in *A*'s *intentions*.[17] To play Russian roulette on an innocent person is to *intend* to impose a risk of death—this seems to be the very point of the action. In contrast, imposing a risk of death is not the point of making toast—while one may be *aware* of this risk, it is not intended. The reason it is impermissible for *A* to play low-risk Russian roulette on *B*, according to McCarthy, is that this action involves an intentional imposition of risk and, thus, intentionally infringes *B*'s rights.[18] In contrast, when

[17] See McCarthy "Rights, Explanation, and Risks," 211–12; and Thomson, "Some Questions about Government Regulation of Behavior," sec. 4. As far as I am aware, it was Nozick who first offered a hypothesis about the moral difference between low-risk Russian roulette and an activity like making toast (Nozick's examples are mining, running trains, and driving). According to Nozick, what distinguishes the former action is that it has no value for society and is not a normal and/or important part of people's lives (*Anarchy, State, and Utopia*, 73, 82). Nozick does not elaborate, but the suggestion does not appear promising. Even if making toast were a rare practice, and few people owned toasters, it is dubious that this would make any moral difference to the activity—and it certainly would not make it into the moral equivalent of playing low-risk Russian roulette on innocent people. We could also imagine a situation in which something like low-risk Russian roulette *was* part of an entrenched social practice to which people assigned importance (as in Shirley Jackson's *The Lottery*). Once again, it is doubtful that this would make much, if any, difference to the moral status of the activity.

[18] Holm defends a variant on this ("A Right against Risk Imposition and the Problem of Paralysis," 921–22): what makes it impermissible for *A* to play low-risk Russian roulette on *B* are *A*'s *reasons for action*, which include the fact that the action will impose a risk upon *B*. When *A* makes toast, the fact that this imposes a risk upon *B* is not one of the reasons for which *A* acts. Holm's proposal is equally subject to the objections in the main text.

A makes toast, although *B*'s rights may be infringed, the infringement is not intended (but merely foreseen).

On closer inspection, though, this suggestion comes nowhere close to capturing the moral difference between these two actions. Suppose *A*'s toaster is broken, and a third party offers to make him a slice of toast if only he plays low-risk Russian roulette on *B*. In this example, *A* has no particular wish to impose a risk of death on *B*—his only aim is to procure toast. Nevertheless, if *A* agrees to this, then his actions are hardly better than if he played low-risk Russian roulette on *B* with the express aim of imposing risk. It is still the case that *A* could be punished for such an action, and it is still the case that *B*, or a third party, could permissibly use force to stop him. The wrongness of playing Russian roulette on an innocent person has little to do with one's intentions—a willingness to impose this risk for a trivial payoff is little better than a direct desire to impose it.[19]

As well as imagining a case in which *A* playing low-risk Russian roulette on *B* does not involve an intentional imposition of risk, we could also imagine a case in which *A* making toast *does* involve an intentional imposition of risk. Suppose *A* does not want toast at all and uses his toaster solely to impose some small risk of death upon his neighbor *B*. While *A*'s motives are certainly criticizable and would seem to reflect poorly on his moral character, it is plausible that his *action* is nevertheless a permissible one.[20] And, even if we do insist that *A* has

[19] One might suggest that, although the imposition of risk is not *A*'s ultimate aim in the new example, it is still being used as a *means*, and this is enough for it to count as *intended*—an idea reflected in certain formulations of the doctrine of double effect. See, for instance, Nelkin and Rickless, "Three Cheers for Double Effect," sec. 1; Scanlon, *Moral Dimensions*, 14; Thomson, "Physician-Assisted Suicide," 512–13. I am unsure if this suggestion is correct—while low-risk Russian roulette is obviously being treated as a means by *A*, it is less clear whether the imposition of risk per se is playing this role. Whatever the truth, the motivational structure of this example is intended to mirror that of the original toast case. In both cases, *A*'s only aim is to get toast. In both cases, *A* employs a means to this end—be it using the toaster or playing low-risk Russian roulette—which he foresees will impose a risk of death on *B*. In neither case is *A* motivated by this imposition of risk—he may even regard it as regrettable (though obviously not so much as to make him reconsider). If the risk imposition counts as intended in the new example, it must also count as intended in the toast example—and whatever makes for the moral difference between the cases is not to be found in *A*'s intentions.

[20] Some philosophers insist on a sharp divide between the moral evaluation of an *action* and the moral evaluation of an *agent*, with intentions and reasons bearing upon the latter but not the former. See, for instance, Oberdiek, "Moral Significance of Risking," sec. 3a; Scanlon, *Moral Dimensions*, esp. ch. 1; Thomson, "Physician-Assisted Suicide," secs. 4 and 5. To undermine McCarthy's strategy, it is not necessary that we endorse this general view—it is enough to maintain that one's intentions are not relevant to the moral permissibility of making toast or of playing low-risk Russian roulette. In fact, even *this* is not strictly necessary, so long as we maintain that whatever difference one's intentions make here is not enough to bridge the moral gulf between these actions.

acted impermissibly, the action is hardly the moral equivalent of playing low-risk Russian roulette on *B*. If *B* became aware of *A*'s reasons for putting on his toaster, he may be perturbed by *A*'s apparent maliciousness toward him, but it would not be proportionate for him to use *force* in order to prevent the action. (If *B were* to use force against *A*—knock him out, break his fingers, even just smash his toaster—our sympathies in this story would quickly switch from *B* to *A*.) Similarly, it would seem cruel and vindictive to *punish A* for putting on his toaster. The overwhelming sense is that, while there is clearly something morally amiss about *A*'s state of mind, the action itself is essentially *harmless* and does not warrant any strong response.

The moral difference between *A* playing low-risk Russian roulette on *B* and *A* making toast in the house next to *B*'s cannot be captured in the way that McCarthy proposes. What *would* capture this moral difference is the verdict that the former action involves an infringement of *B*'s rights, while the latter does not. But this, of course, is the very prediction that neither the Risk Thesis nor the High Risk Thesis seems able to deliver. The right posited by the Risk Thesis is infringed by both of these actions, while the right posited by the High Risk Thesis is not infringed by either.

4. Revisiting the High Risk Thesis

I will now outline a way of delivering the desired verdict. The first step is to highlight a tacit assumption that has guided the discussion so far. Return to the case of low-risk Russian roulette: a single bullet is placed into a single chamber of one out of a large set of revolvers before *A* chooses a revolver at random, spins the cylinder, points at *B*, and pulls the trigger. So far, this has been classified as a "low-risk" scenario on the grounds that the *probability* of *B* being killed is very small. And yet, if we were actually watching these events unfold, the thing that would surely shock us—and move us to intervene if we could—is the perceived *riskiness* of what *A* is doing. It would be natural to have something like the following thought: the bullet has to be located in some chamber, and it would be just as *normal* for it to be in any one chamber as any other—including the chamber that slides into alignment with the barrel of *A*'s revolver when he pulls the trigger. If *B* were shot and killed, then, given the nature of the setup, we would not need any special explanation as to how this could have happened. One who is struck by this thought would not be altogether reassured by learning how *many* revolvers were in the initial set. The more revolvers there are, the more places the bullet could end up—but there is still nothing *preventing* it from being in the one chamber that would result in *B*'s death.

Here is another way to put the point: the most *normal* possible worlds in which *A* plays Russian roulette on *B* will include worlds in which the bullet is in each of the available chambers. As a result, some of the most normal possible worlds in which *A* plays Russian roulette on *B* will be worlds in which *B* is killed—this represents one normal outcome of the action. Clearly, the notion of normalcy that is being invoked here is distinct from the idea of statistical frequency—*B*'s death is not an outcome that would *frequently* arise from this action were it repeated over and over. Rather, this outcome is normal in the same sense that it would be normal for, say, "10, 7, 13, 8, 25, 19" to be the winning lottery numbers—some sequence of numbers has to come up, and this sequence would require no more explanation than any other.

When it comes to *A* putting on his toaster, however, the situation seems altogether different. While it is *possible* that this action could cause a fire that leads to the death of his neighbor *B*, there is no sense in which this would count as a *normal* outcome of the action. On the contrary, there would have to be some explanation as to how the fire started (was there an electrical fault in the toaster or in the wiring of *A*'s house?), how it took hold (was there inflammable material around the toaster, was there a gas leak?), how *A* failed to extinguish the fire or raise the alarm (was he asleep, did he leave the house?), and so on. If we were told that *B* had died in a fire as a result of *A* putting on his toaster, our immediate reaction would be to ask *how* this could have possibly happened. If we were told that *B* had died as a result of *A* subjecting him to Russian roulette, our reaction would be quite different—no matter how many revolvers *A* was choosing from. I think that this contrast in our reactions is tracking a genuine difference between these two hypothetical events. Of all the outcomes that could result from *A* putting on his toaster, *B*'s death is a highly abnormal one. In "possible worlds" talk, the most normal worlds in which *A* puts on his toaster are worlds in which *B* suffers no harm as a result, and any worlds in which this action leads to *B*'s death are highly abnormal.

It has been taken for granted in the discussion so far that the risk of a given outcome is determined by its probability—the greater the probability, the greater the risk, and the lower the probability, the lower the risk. It has also been assumed, accordingly, that any risk threshold we use in spelling out the High Risk Thesis must take the form of a probability value (such as one in five hundred thousand). This "probabilistic" conception of risk is entrenched across a range of areas and has been largely assumed, unquestioned, in discussions of the right against risk imposition.[21] But this conception of risk is

[21] And, indeed, throughout the literature on the ethics of risk imposition. According to one well-known family of views, we are morally required to act in a way that, roughly speaking, minimizes the strongest individual complaint against our action. On the "*ex ante*"

not inevitable—and should be seen as another potential moving part in the puzzles we have been considering. The dominance of the probabilistic account of risk has recently been challenged by several authors who have put forward alternatives such as the *modal* account, the *relevant alternatives* account, and the *normic* account—which will be my focus here.[22]

According to the probabilistic account, the risk that a particular outcome would result from a given action depends on how probable it is that the outcome would result from the action.[23] According to the normic account, the risk that a particular outcome would result from a given action depends on how *abnormal* it would be for the outcome to result from the action. As above, the notion of normalcy at work here is linked with the need for explanation—an outcome is abnormal to the extent that it requires special explanation in terms of factors that are additional to the action. On the probabilistic account, when *A* plays Russian roulette on *B*, the risk to *B* depends upon the number of revolvers involved and can, as a result, be made arbitrarily close to zero. On the normic account, things look altogether different; given the setup, no special explanation would be needed if *B* were shot and killed—no matter how many revolvers were involved, this would represent one of the normal outcomes of the action.

Suppose, as hinted above, that possible worlds can be ranked according to their normalcy—the most normal worlds are assigned a rank of zero, the next

version of this view, when an action imposes a risk of harm upon an individual, they have a complaint against it, the strength of which is discounted according to the level of risk involved (see, for instance, Frick, "Contractualism and Social Risk"; Kumar, "Risking and Wronging"; for critical discussion see Horton, "Aggregation, Complaints, and Risk"). But it is assumed, when determining the strength of such a complaint, that the level of risk is to be measured *probabilistically*. See, for instance, Frick, "Contractualism and Social Risk," 188; Horton, "Aggregation, Complaints, and Risk," sec. 2; Kumar, "Risking and Wronging," sec. 4a. This will make the view ill-equipped to handle cases of low-risk Russian roulette vs. making toast—and also leads to a problem with cases of distributed risk. I will not explore this further here.

[22] For the modal account, see Pritchard, "Risk." For the relevant alternatives account, see Gardiner, "Relevance and Risk." For the normic account, see Ebert, Smith, and Durbach, "Varieties of Risk"; Smith, "Decision Theory and De Minimis Risk."

[23] The probability in question could be understood as "objective"—determined perhaps by the frequency with which the outcome would accompany the act or some such. My own view is that the probability is best understood as epistemic or *evidential*—the probability that a given outcome will eventuate, given the evidence that the action has been performed. There are important questions about how much should be included in the relevant description of an action—and different answers may, of course, give rise to different assessments of the risk that the action poses. In the examples I consider here, the relevant descriptions seem relatively clear—but there will undoubtedly be more difficult cases, and, arguably, a more principled approach to this issue would be needed for any complete ethics of risk imposition.

most normal worlds are assigned a rank of one, and so on.[24] Suppose an action could, in principle, result in harm to a given individual. If the most normal worlds in which the action is performed include worlds in which the individual is harmed, then this outcome will have an abnormality of zero, given the action—it will, in short, represent one of the normal outcomes of the action. If the individual does not come to harm in any of the most normal worlds in which the action is performed, then this will not be a normal outcome of the action—and its abnormality may be gauged by the difference in rank between the most normal worlds in which harm results from the action and the most normal worlds in which the action is performed. If the former worlds are one rank more abnormal than the latter, then the abnormality of the individual suffering harm, given the action, will be equal to one. If the former worlds are seven ranks more abnormal than the latter, then the abnormality of the individual suffering harm, given the action, will be equal to seven, and so on.

Among the most normal worlds in which *A* plays Russian roulette on *B* are worlds in which *B* is shot and killed. On the normic account of risk, when *A* plays Russian roulette on *B*, the risk to *B* is *maximal*, as *B*'s death has an abnormality of zero, given *A*'s action. While I have spoken of cases of "low-risk" Russian roulette (and will, for ease, continue to use that term), on a normic interpretation, there is, in effect, *no such thing* as low-risk Russian roulette—the normic risk is maximal, no matter how many revolvers are involved. In contrast, when *A* makes toast, this will count as a low-risk activity in both the probabilistic and normic senses. While there are possible worlds in which *A*'s putting on his toaster results in a fire in which *B* dies, these worlds are highly abnormal.

It is important to emphasize that I am not proposing the normic account of risk as a *competitor* to the probabilistic account. In my view, both probabilistic risk and normic risk represent legitimate ways of precisifying our ordinary risk concept.[25] In fact, this kind of pluralist approach fits well with the example of low-risk Russian roulette—in which our intuitions about the risks involved do, arguably, pull in different directions. On the one hand, it is intuitive that the risk to *B* diminishes as the number of revolvers is increased—that the risk is halved if two revolvers are used instead of one and halved again if four revolvers are used, etc. This is why low-risk Russian roulette may be morally preferable to standard Russian roulette. On the other hand, it is intuitive that, irrespective of the number of revolvers involved, the risk to *B* is greater than that imposed by *A* making toast or engaging in other everyday activities. This is why low-risk

[24] See Smith, *Between Probability and Certainty*, ch. 8, "The Logic of Epistemic Justification," and "The Hardest Paradox for Closure."

[25] Ebert, Smith, and Durbach, "Varieties of Risk," sec. 6.

Russian roulette will never be the moral equivalent of an everyday activity. While the former intuition is captured by the probabilistic account, the latter is captured by the normic account.[26]

For the original Risk Thesis, it makes no difference whether the "risk" in question is interpreted probabilistically or normically. If an action involves *some* probabilistic risk of harm, then there must be a possible world in which harm results from the action, in which case the action will also involve some normic risk of harm and vice versa.[27] When it comes to the High Risk Thesis, however, the two different ways of disambiguating the notion of risk give rise to two distinct theses: the *Probabilistic* High Risk Thesis (which we have been taking for granted so far) and the *Normic* High Risk Thesis.

> *Probabilistic High Risk Thesis*: We have a right that others not subject us to a high probabilistic risk of harm. More precisely, we have a right that others not act in such a way that the probability of our being harmed, as a result of their action, is above a threshold *t*.

> *Normic High Risk Thesis*: We have a right that others not subject us to a high normic risk of harm. More precisely, we have a right that others not act in such a way that the abnormality of our being harmed, as a result of their action, is below a threshold *t*.

Unlike the Risk Thesis and the Probabilistic High Risk Thesis, the Normic High Risk Thesis can separate the act of playing low-risk Russian roulette on an innocent from the act of making toast. While the right posited by the Risk Thesis is infringed by both acts, and the right posited by the Probabilistic High Risk Thesis is infringed by neither, the right posited by the Normic High Risk Thesis, given an appropriate choice of threshold, will be infringed by the first but not the second.

[26] More precisely, the following three claims are inconsistent if "risk" is given the same interpretation in each:

1. The risk involved in low-risk Russian roulette is halved when the number of revolvers is doubled.
2. Low-risk Russian roulette always involves a greater risk than making toast.
3. Making toast involves some positive level of risk.

[27] This assumes that every possible world is assigned some normalcy rank and some nonzero probability. The latter assumption is typically dropped in the case of infinite probability spaces, and without it, the existence of possible worlds in which an action results in harm will be consistent with the probability of harm being zero, conditional upon the action being performed. In this case, the "vice versa" direction of the above will fail—an action could present some normic risk of harm without presenting any probabilistic risk of harm—though it is doubtful that this would make any difference in practice.

Having distinguished between probabilistic and normic risk, one might wonder *why* it is the second kind of risk that should figure in our right against risk imposition. Why should we have a right that others not subject us to a high *normic* risk of harm? On first impressions, this claim might seem rather mysterious. My primary aim here is to argue that, by understanding the right against risk imposition in normic terms, we are able to solve a number of problems that arise for rights-based approaches to the ethics of risk imposition. Questions about the foundation or basis of such a right lie, for the most part, beyond the scope of this paper—but I will conclude this section with one speculative line of thought based on a connection between normic risk and the limits of our responsibility for the consequences of our actions.

Suppose *A*'s decision to make toast really *did* lead to a fire in which his neighbor *B* died. In this case, it is plausible that this outcome would not be wholly attributable to *A*'s action—for it would owe in part to circumstances (be they faulty wiring, a gas leak, etc.) that lie completely outside of *A*'s awareness and control. As a result, *A* would not be considered *fully responsible* for *B*'s death. Similar remarks apply to any action that presents a low normic risk of harm. If there is a low normic risk that an action will cause harm to an individual, then harm could only result through the intervention of independent, interfering factors, which would serve to mitigate the agent's responsibility.

There is no equivalent connection between responsibility and probabilistic risk. Even if an action presents a low probabilistic risk of harm, one may still bear full responsibility for any harm that ensues. If *A* plays low-risk Russian roulette on *B,* then, irrespective of the number of revolvers involved, *A* would be fully responsible in the event that *B* were shot and killed. These brief remarks do not, of course, amount to a full explanation of why the Normic High Risk Thesis should be true—but they do perhaps dispel some of the mystery that might otherwise surround it.[28]

[28] Another point to bear in mind is that the Probabilistic and Normic High Risk Theses are not formally inconsistent. While I have been assuming (as seems standard in the literature) that there is, at most, *one* right against risk imposition, there is no logical barrier to accepting *both* theses. And the Normic High Risk Thesis would surely seem less mysterious if the Probabilistic High Risk Thesis were accepted alongside it. In this case, any imposition of high risk would infringe a person's rights, no matter how the notion of risk is interpreted. This "combined" view might also offer a more satisfactory treatment of certain cases—such as those in which an individual is subjected to a high probabilistic but low normic risk of harm. See Smith, "Decision Theory and De Minimis Risk," sec. 5. I will not explore this further here.

5. Distributed Risk

In section 1, I presented two problems for the High Risk Thesis—one of which concerned cases of low-risk Russian roulette and the other of which concerned cases of distributed risk. In the previous section, I argued that, on a normic interpretation, there is no such thing as a case of low-risk Russian roulette, and the problem dissolves. In this section, I will argue that the same is true for cases of distributed risk—on a normic interpretation, such cases simply *cannot arise*. A case of distributed risk, recall, was defined as one in which there is a high risk that some member of a group will be harmed, even though the risk to each individual member is low. Consider a group of individuals *C*, *D*, *E*, *F*, etc., and suppose there is a high risk that some member of the group will suffer harm. On the normic account, this means that there is a relatively normal possible world in which some member of the group suffers harm. But any world in which some member of the group suffers harm must either be a world in which *C* suffers harm or a world in which *D* suffers harm, etc., in which case either *C* must be at a high risk of harm or *D* must be at a high risk of harm, etc. That is, if there is a high normic risk that some member of the group will be harmed, then there must be some member of the group who is at a high normic risk of harm. If the risk to each member is equal, then they will *all* face a high normic risk of harm.

More formally, if x is a variable ranging over the members of some group, Hx is read "x is harmed" and $\Diamond$ is read "There is a high risk that . . . ," then, on a normic reading, we will have an instance of the "Barcan Formula": $\Diamond\exists xHx \rightarrow \exists x\Diamond Hx$. That is, if there is a high risk that some individual in the group is harmed, then, on the normic reading, there is some individual in the group who is at a high risk of harm. Since the converse clearly holds, the normic account predicts that the two risk attributions are, in fact, equivalent—it makes no difference whether the high-risk operator or the existential quantifier is given wide scope: $\Diamond\exists xHx \leftrightarrow \exists x\Diamond Hx$.[29]

What, then, will the Normic High Risk Thesis predict in putative cases of distributed risk, such as the toxic chemical case? In the toxic chemical case, we are explicitly told the probabilistic risks associated with each possible action—we are told that if *A* dumps the chemical in the pond, there is a one-in-a-thousand chance that *B* will be exposed to a harmful amount, and if *A* dumps the chemical in the river, then, for each of the million people living downstream, there is a one-in-a-million chance that the individual will be exposed to a harmful amount. Obviously, there are no normic risks stipulated—and the details

[29] If Rx is read "x's rights are infringed" then the High Risk Thesis gives us the conditional $\forall x(\Diamond Hx \rightarrow Rx)$, from which we can derive $\exists x\Diamond Hx \rightarrow \exists xRx$. If the risk is interpreted normically, we can infer $\exists xRx$ from $\Diamond\exists xHx$ (via $\exists x\Diamond Hx$). If the risk is interpreted probabilistically, however, then the inference is blocked.

that would be needed to assess these risks are also largely missing in existing descriptions of the case.

Here, perhaps, is one natural way of filling in the required details. If *A* were to dump the chemical in the pond, then, given the quantity and potency of the chemical, the volume of the pond, and the way in which *B* normally uses the pond water, the amount of chemical to which *B* is exposed will vary throughout a certain range. While most of the values in this range would result in no ill effects, the highest values would cause harm to *B*. In this case, if *B* were to suffer harm as a result of *A* dumping the chemical in the pond, then no special explanation would be needed—this would be like subjecting *B* to a kind of low-risk (or medium-risk?) Russian roulette.

Similarly, if *A* were to dump the chemical in the river, then each individual living downstream faces a potential exposure range, given facts about the quantity and potency of the chemical, the volume and flow of the river, and the way in which the river water is normally used. If some of the values in this range are above the harmful level, then, once again, for a given individual to suffer harm as a result of *A* dumping the chemical in the river would not demand special explanation—this would be like subjecting each of these individuals to a kind of low-risk Russian roulette. When the details are filled in like this, all of the normic risks are maximal—if *A* dumps the chemical in the pond, then *B* is at maximal normic risk of harm, and if *A* dumps the chemical in the river, then every individual downstream is at maximal normic risk of harm. As a result, the Normic High Risk Thesis will predict that, all else equal, it would be morally worse to dump the chemical in the river, as this would involve a million rights infringements while dumping the chemical in the pond would involve only one.

If, however, we were to alter the case in such a way that the presence of the chemical in the river would present only a *low* normic risk to each of the people living downstream, then the Normic High Risk Thesis could make a different prediction. What might such a case look like? Suppose a series of measures is in place to prevent the people living downstream from ever coming into contact with the river water—perhaps the land around the river is private and trespassers face penalties or prosecution, perhaps the river is protected by a high fence, etc. None of this would make it *certain* that a given person living downstream will not be harmed if the chemical is dumped in the river, but it would generate the need for special explanation in the event that they are. How did they get past the fence? Why were they willing to trespass? And so on. If the details are filled in in this way, then, for any individual *x* living downstream, it would be abnormal for *x* to be harmed as a result of *A* dumping the chemical in the river. If the degree of abnormality is greater than the threshold posited by the Normic High Risk Thesis, then, as far as this thesis is concerned, dumping the chemical in the river will involve no rights infringements. If the situation with *B* and the pond is unchanged from the description given above, then, all else equal, the Normic High Risk Thesis will predict that it is morally preferable for *A* to dump the chemical in the river.[30]

As we have seen, though, since the normic risk to each individual living downstream is low, so too is the normic risk to the group—dumping the chemical in the river presents a low normic risk of *any* individual being harmed. Not only, then, does dumping the chemical in the pond infringe *B*'s rights, it also involves a higher overall normic risk of harm. If *A* were to dump the chemical in the pond, then *B* could suffer harm in a way that is consistent with conditions being normal—this would represent one of the normal outcomes of the action. If *A* were to dump the chemical in the river, then, under normal conditions, there is no individual who would be harmed. Furthermore, if an individual living downstream *were* harmed as a result of *A* dumping the chemical in the river, then *A* would not be fully responsible for this harm, as it would be due in part to the individual's own actions—trespassing, scaling the fence, etc.

It might still be the case, of course, that dumping the chemical in the river would involve a higher overall *probabilistic* risk of harm. Indeed, there is no reason why the probabilities could not remain as originally stipulated—given the sheer number of people who live downstream, if *A* dumps the chemical in the river, there is a 64 percent chance that at least one of these people will, for some reason, flout the rules, come into contact with the water, and be harmed.[31] In light of this, some would balk at the idea that facts about normic risk could ever make the river option morally preferable to the pond option. Some would insist that, with the probabilities as they are, the river option would *always* be morally worse.

I will not attempt to engage this position here except to say this: the prediction that the river option *may* be morally worse, depending on how the

[30] We could, of course, avoid this result if we were willing to set the abnormality threshold very high. Even if the penalties, fences, etc., would make it highly abnormal for any given individual to be harmed as a result of *A* dumping the chemical into the river, if the abnormality threshold that features in the Normic High Risk Thesis were higher *still*, then this action would nevertheless infringe the rights of those living downstream, and the pond option would remain morally preferable. But the higher we push the abnormality threshold, the more of our ordinary everyday activities will turn out to infringe others' rights, and, in the limit, we would end up mired in the same problems that beset the original Risk Thesis. Whatever one thinks about this particular case, I do not think that threshold raising is viable as a general strategy for dealing with cases of this kind.

[31] To compensate for the probabilistic effect of the fences, penalties, etc., we could imagine that there is a higher probability that any person who comes into contact with the water suffers harm. Alternatively, we could achieve the same effect by increasing the number of people who live downstream.

non-probabilistic details of the case are filled in, is as close as we can come to the above prediction while working exclusively within the framework of individual rights. The only right against risk imposition that will yield the result that the river option is always morally worse than the pond option is the right posited by the original Risk Thesis—and, as argued, this thesis is untenable. Those who wish to maintain that the river option is always morally worse should, I suggest, give up on attempting to derive this result purely from a right against risk imposition. One could still accept the existence of such a right—and still perhaps find explanatory work for it—but would need to argue that, when it comes to comparing these two options, rights infringements are not the decisive factor.[32]

6. Conclusion

I have argued that the most promising rights-based approach to the ethics of risk imposition comes in the form of the Normic High Risk Thesis—the claim that we each have a right that others not impose a high normic risk of harm upon us. Unlike the Risk Thesis, the Normic High Risk Thesis does not make for rampant, trivial rights infringements. Unlike the Probabilistic High Risk Thesis, the Normic High Risk Thesis does not generate a moral preference for cases in which a risk is distributed among the members of a group. Unlike either of these theses, the Normic High Risk Thesis is able to account for the moral difference between the risks imposed by making toast and the risks imposed by "low-risk" Russian roulette.[33]

University of Edinburgh
martin.smith@ed.ac.uk

[32] Those who insist that the river option is always morally worse than the pond option may find it natural to appeal to considerations of *expected utility*—where the expected utility of an action is equal to the probability-weighted average of the utilities of its possible outcomes. If an individual suffering harm as a result of the chemical is assigned a constant, finite disutility, then, with the probabilities as stipulated, the pond option will always have a significantly higher expected utility than the river option. Once we assign expected utilities a moral role, however, one might think that there is no longer any *need* for the Normic High Risk Thesis or for any right against risk imposition—why not let our moral assessment of a risk-imposing activity be determined purely by its expected utility? There are reasons to be dissatisfied with this approach, however. This view will, for instance, generate the wrong predictions about low-risk Russian roulette, which could, given enough revolvers, have a higher expected utility than making toast. For related discussion, see Smith, "Decision Theory and De Minimis Risk," sec. 3. A full evaluation of this view is beyond the scope of this paper.

[33] This paper was presented (online) at the Risk and Recklessness workshop, University College London in April 2021, at the Epistemology and Normality workshop, Dianoia Institute of Philosophy, Australian Catholic University in January 2022, and at the Imposing

REFERENCES

Bergelson, Vera. "Self-Defense and Risks." In *The Ethics of Self-Defense*, edited by Christian Coons and Michael Weber, 135–51. Oxford: Oxford University Press, 2016.

Dworkin, Ronald. *Taking Rights Seriously*. London: Duckworth, 1977.

Ebert, Philip A., Martin Smith, and Ian Durbach. "Varieties of Risk." *Philosophy and Phenomenological Research* 101, no. 2 (September 2020): 432–55.

Finkelstein, Claire. "Is Risk a Harm?" *University of Pennsylvania Law Review* 151, no. 3 (2003): 963–1001.

Frick, Johann. "Contractualism and Social Risk." *Philosophy and Public Affairs* 43, no. 3 (Summer 2015): 175–223.

Fried, Charles. *An Anatomy of Values: Problems of Personal and Social Choice*. Cambridge, MA: Harvard University Press, 1970.

Gardiner, Georgi. "Relevance and Risk: How the Relevant Alternatives Framework Models the Epistemology of Risk." *Synthese* 199, nos. 1–2 (December 2021): 481–511.

Gerwith, Alan. "Are There Any Absolute Rights?" *Philosophical Quarterly* 31, no. 122 (January 1981): 1–16.

Hayenhjelm, Madeleine, and Jonathan Wolff. "The Moral Problem of Risk Impositions: A Survey of the Literature." *European Journal of Philosophy* 20, no. S1 (June 2012): E26–E51.

Holm, Sune. "A Right against Risk-Imposition and the Problem of Paralysis." *Ethical Theory and Moral Practice* 19, no. 4 (August 2016): 917–30.

Horton, Joe. "Aggregation, Complaints, and Risk." *Philosophy and Public Affairs* 45, no. 1 (Winter 2017): 54–81.

Kumar, Rahul. "Risking and Wronging." *Philosophy and Public Affairs* 43, no. 1 (Winter 2015): 27–51.

McCarthy, David. "Rights, Explanation, and Risks." *Ethics* 107, no. 2 (January 1997): 205–25.

McKerlie, Dennis. "Rights and Risk." *Canadian Journal of Philosophy* 16, no. 2 (1986): 239–52.

Nelkin, Dana Kay, and Samuel C. Rickless. "Three Cheers for Double Effect." *Philosophy and Phenomenological Research* 89, no. 1 (July 2014): 125–58.

Nozick, Robert. *Anarchy, State, and Utopia*. New York: Basic Books, 1974.

Oberdiek, John. "Lost in Moral Space: On the Infringing/Violating Distinction

Risk workshop, University of Manchester in September 2022. Thanks to all of those who participated on these occasions. Thanks also to two anonymous referees for this journal. Work on this paper was supported by the Arts and Humanities Research Council (grant no. AH/T002638/1).

and Its Place in the Theory of Rights." *Law and Philosophy* 23, no. 4 (2004): 325–46.

———. "The Moral Significance of Risking." *Legal Theory* 18, no. 3 (2012): 339–56.

Pritchard, Duncan. "Risk." *Metaphilosophy* 46, no. 3 (July 2015): 436–61.

Railton, Peter. "Locke, Stock, and Peril: Natural Property Rights, Pollution, and Risk." In *Facts, Values, and Norms: Essays toward a Morality of Consequence*, 187–225. Cambridge: Cambridge University Press, 2009.

Rowe, Thomas. "Can a Risk of Harm Itself Be a Harm?" *Analysis* 81, no. 4 (October 2021): 694–701.

Scanlon, Thomas. *Moral Dimensions: Permissibility, Meaning, Blame.* Cambridge, MA: Harvard University Press, 2008.

Smith, Martin. *Between Probability and Certainty: What Justifies Belief.* Oxford: Oxford University Press, 2016.

———. "Decision Theory and De Minimis Risk." *Erkenntnis* (forthcoming). Published online ahead of print, December 31, 2022. https://doi.org/10.1007/s10670-022-00624-9.

———. "The Hardest Paradox for Closure." *Erkenntnis* 87, no. 4 (August 2022): 2003–28.

———. "The Logic of Epistemic Justification." *Synthese* 195, no. 9 (September 2018): 3857–75.

Song, Fei. "Rights against High-Level Risks of Harm Impositions." *Ethical Theory and Moral Practice* 22, no. 3 (June 2019): 763–78.

Thomson, Judith Jarvis. "Imposing Risks." In *To Breathe Freely: Risk, Consent, and Air*, edited by Mary Gibson, 124–40. Totowa, NJ: Rowman and Allanheld, 1985.

———. "Physician-Assisted Suicide: Two Moral Arguments." *Ethics* 109, no. 3 (April 1999): 497–518.

———. *The Realm of Rights*. Cambridge, MA: Harvard University Press, 1990.

———. "Some Questions about Government Regulation of Behavior." In *Rights and Regulation: Ethical, Political, and Economic Issues*, edited by Tibor R. Machan and M. Bruce Johnson, 137–56. Cambridge, MA: Ballinger, 1983.

———. "Some Ruminations on Rights." *University of Arizona Law Review* 19, no. 1 (1977): 45–60.

Journal of Ethics and Social Philosophy
vol. 27, no. 3 · may 2024

https://doi.org/10.26556/jesp.v27i3.1969

DOXASTIC PARTIALITY AND THE PUZZLE OF ENTICING RIGHT ACTION

Max Lewis

Friends and family help. In fact, we think that they, in some sense, ought to help us. But does it follow from the fact that they ought to help that there is nothing suspect with us trying to entice them to help? Consider the following case:

> *Moving*: Ronan and Anna are close friends. One evening when they are hanging out, Ronan receives a call from his landlord. His landlord tells him that they need to do an emergency fumigation of the apartment and that Ronan needs to move many of his belongings out of his apartment. Ronan tells Anna that he needs to move a bunch of his belongings immediately. Anna listens but does not immediately say anything. Ronan then utters, "I've done favors for you in the past. Now I ask that you do me a favor and help me move."

There is something "off" about the way that Ronan requests Anna's help. Notice that he tries to entice her to help by pointing out that he has helped her in the past. Notice further that because they are close friends, Anna seems to have a sufficient normative reason to help Ronan move—that is, a normative reason that is just as weighty as the reasons she has to do anything else and thus is weighty enough to justify Anna's helping. We can also grant that Anna knows (or is in a position to know) this. Moreover, the fact that Ronan has done Anna plenty of favors either provides an additional reason for her to help or at least intensifies the strength of her reason to help him. Nonetheless, it seems odd for Ronan to appeal to this reason in order to entice Anna to help.

However, in many cases involving coworkers or acquaintances, there is nothing odd about appealing to the fact one has helped someone in the past in order to entice them to help. Consider the following case:

> *Shift*: Ming and Nina are coworkers. One evening when they are at work, Ming receives a call from his landlord. His landlord tells him that they need to do an emergency fumigation of his apartment tomorrow and that Ming needs to move many of his belongings out of his apartment.

> Ming tells Nina that he needs to move a bunch of his belongings tomorrow but that he is scheduled to work. Nina listens but does not immediately say anything. Ming knows that Nina is not scheduled to work tomorrow. He then utters, "I've covered your shift in the past. Now I ask that you do me a favor and cover my shift tomorrow."

The fact that Ming has helped Nina in the past seems to provide her with a reason to help Ming in this case, or at least it intensifies the strength of the reason she has to help him. But it does not seem odd for him to appeal to the fact that he has helped her in order to entice her to help.

Perhaps the problem with Ronan's attempt to entice his friend to help him move by mentioning that he has done favors for her is that it indicates that he is keeping track of favors, and intimates do not count favors. In particular, when intimates do each other favors or kindnesses, they do so *unconditionally*—that is, not on the condition that their intimate will reciprocate. But this cannot be the core of the problem. Consider the following case:

> *Campaign 1*: Ville and Olivia are a married couple. Olivia wants to run for city council. She knows that she will need help campaigning, and she wants Ville to help manage her campaign. Ville knows that Olivia wants his help, but he has not offered it yet. Olivia says to Ville, "Please help me run my campaign. After all, it's the prudent thing to do. It'll look good on your résumé, and the experience will mean that you can ask for a higher salary in the future."

There is something "off" about the way that Olivia requests Ville's help. She tries to entice him to help her by pointing to a strong prudential reason for Ville to help her. And it is certainly true that the fact that helping her would make him a stronger candidate for certain jobs is a prudential reason for him to help. But appealing to that reason to entice her husband to help seems odd. Also, notice that Olivia says nothing about previous favors and so the oddness of her request and of Ronan's does not seem to have to do with "counting favors" or failing to give unconditional help.

Note further that trying to entice strangers or acquaintances in the way that Olivia does is not odd at all. Consider the following case:

> *Campaign 2*: Tom and Aisha are acquaintances. Tom wants to run for city council. He knows that Aisha, although she is somewhat inexperienced, has the potential to be a great campaign manager. Aisha knows that Tom wants her help, but she has not offered yet. Tom and Aisha have a meeting in which Tom says, "Please help me run my campaign. After all, it's the prudent thing to do. It'll look good on your résumé,

> and the experience will mean that you can ask for a higher salary in the future."

There does not seem to be anything strange about the way that Tom tries to entice Aisha to help him manage his campaign. But notice that he appeals to the same prudential reasons that Olivia appealed to when she requested that Ville help her manage her campaign.

It can also be odd to try to entice our intimates by appealing to moral reasons they have (e.g., telling a friend that she should come to visit me in the hospital because it is her moral duty) and even by giving them new reasons to help (e.g., imagine I am driving to dinner with a friend when my tire bursts, and then I offer him money to help me change it). Even if it is true that my friend has a moral duty to visit me in the hospital, it is seemingly problematic for me to try to entice him by explicitly mentioning this reason. And, even if I could give my friend an additional reason to help me change my tire by paying him money, it still seems troubling for me to try to entice him in this way.

All of these cases suggest the following thesis:

> *Problematic*: Generally, it seems problematic to appeal to certain facts (e.g., previous favors and prudentially relevant facts) in order to entice our intimates to do things that help us even when those facts actually provide our intimates with sufficient reasons to perform those actions.

To be clear, Problematic is a claim about what is *generally* true. So, I am not claiming that it is always odd to appeal to facts about previous favors or prudentially relevant facts in order to entice our intimates to help us. For example, if one is asking a friend to risk their life (e.g., to save one in a fire), then there might be nothing seemingly odd about mentioning that one has risked one's own life for one's friend. In such cases, the duress one is under might excuse what one does, or the high stakes might make it so that one's friend has a weaker normative reason to help, and thus, one's request can either provide an additional reason to help or intensify the strength of the original normative reason.[1]

Nonetheless, the fact that the above-mentioned ways of enticing our intimates to act even sometimes seem problematic is puzzling because all of the requested actions are actions that the requestees have sufficient normative reason to do regardless of the request. Moreover, it may even be the case that the reasons appealed to in Moving and Campaign 1 are decisive reasons for Anna and Ville to help—that is, they are reasons that are weightier than the reasons to do anything else. So why would it seem problematic to appeal to these reasons in order to entice our intimates to help?

[1] I thank an anonymous referee for suggesting this qualification.

The above cases (Moving, Shift, Campaign 1, and Campaign 2) also suggest the following thesis:

> *Asymmetry*: Generally, it seems more problematic to appeal to certain kinds of facts (e.g., about previous favors or prudentially relevant facts) in order to entice our intimates to do things that help us than it is to appeal to these facts in order to entice nonintimates to perform the same actions.

Asymmetry is also a claim about what is *generally* true. So, Asymmetry is compatible with there being some cases in which it is just as seemingly problematic to appeal to certain kinds of facts in order to entice our intimates to do things that help us as it is to appeal to these facts in order to entice nonintimates to perform the same actions. In some cases, it will only seem problematic to entice intimates in this way. Other times, it will still be seemingly problematic to try to entice nonintimates to help one. For example, imagine that I hire an electrician to wire my new home, and he promises to come by on Tuesday to start the job. Now imagine that I call him on Monday not just to confirm with him, but rather I try to entice him by reminding him that I promised to pay him extra. If it is clear that I am trying to entice him because I did not believe him when he promised to start on Tuesday, then there is something odd about my trying to entice him.[2] But it is still even more seemingly problematic if that electrician is my good friend or my sibling.

Asymmetry is also puzzling because, all else being equal, we are usually allowed to ask more from our friends and family than we are of coworkers and acquaintances. So why would it be more seemingly problematic to entice intimates to help than to entice coworkers and acquaintances in the same way?

Let us call the conjunction of Problematic and Asymmetry *the puzzle of enticing right action*. Solving this puzzle is important because it concerns how we treat our nearest and dearest. If we mistreat or wrong them by appealing to certain kinds of reasons to entice them to help us, then it is important to know this. As we will see, solving this puzzle will show us not only what kind of enticements can be problematic but also what beliefs we should have concerning our intimates. Insofar as we care about our intimates, we should care about treating them well and thus avoiding making problematic enticements or having problematic beliefs about them.

Here is the plan. In section 1, I distinguish this puzzle from a similar puzzle recently proposed by Laskowski and Silver.[3] In section 2, I consider whether

[2] I thank an anonymous referee for suggesting this kind of case.

[3] Laskowski and Silver, "Wronging by Requesting."

Laskowski and Silver's proposal for solving their own puzzle works for the puzzling of enticing right action. I argue that it does not. On their view, enticing right action can seem problematic because the enticers are being disrespectful. In particular, they are disrespecting their intimates by expressing a lack of trust that the intimates will do what they know they are morally required to do.

In section 3, I provide my own solution to the puzzle of enticing right action. My explanation of Problematic is that the enticements indicate that the enticer violates a demand of good intimate relationships. In particular, the enticements indicate that the enticer violates a demand for a certain kind of doxastic partiality—that is, they should trust their intimates to follow what their intimates know is a demand of good intimate relationships when it comes to them. It is a demand of good intimate relationships that people be sufficiently motivated to act so as to protect or promote the needs, desires, interests, projects, and well-being of their intimates for their intimates' own sakes. The above enticements strongly indicate that none of the enticers trusts their intimate to be sufficiently motivated to act in these ways, and so it looks like the enticers violate a demand of good intimate relationships. My explanation of Asymmetry is that while we are required to trust our intimates to be motivated in the way mentioned above, we are not required to trust strangers to have such motivations.

In section 4, I clarify my account by making certain background assumptions about normativity and responsibility explicit. In section 5, I further distinguish my account from Laskowski and Silver's by showing how their view is committed to practical reasons for belief, but my view is not. In section 6, I briefly conclude.

1. Clarifying the Puzzle

The puzzle I am interested in concerns different ways that one might *entice* intimates to help one—that is, ways in which one can *rationally persuade* one's intimate to do something for one. One way to do this is to request that they perform the relevant action. On the orthodox picture of requesting, when A requests that B ϕ, A gives B a new reason to ϕ.[4] So, one attempts to rationally persuade someone to do something by giving them a new reason to do it. On a heterodox view of requesting from Laskowski and Silver, A's request that B ϕ does not give B a new reason to ϕ, but rather points to an already-existing reason that B has to ϕ. So, one attempts to rationally persuade someone to do

[4] Raz, *The Morality of Freedom*, 36–37, and *Practical Reason and Norms*, 100–101; Enoch, "Giving Practical Reasons," 1; Owens, *Shaping the Normative Landscape*, 86; Cohoe, "God, Causality, and Petitionary Prayer," 33; Herstein, "Understanding Standing," 3115; and Lewis, "Discretionary Normativity," 2.

something by pointing to a reason that they already have to do it. A related way to entice intimates is to *explicitly* mention facts that provide them with reasons to do what you want them to do. In this paper, I focus on cases in which a person both requests an action and explicitly mentions facts that provide their intimates with sufficient (and perhaps even decisive) reason to perform the action.

Laskowski and Silver raise a similar puzzle in which it seems problematic to make certain requests of intimates.[5] The puzzle they are interested in is related to, but narrower than, the one I am interested in. Showing just how the two puzzles differ will help clarify the puzzle I am interested in.

Consider the following case from Laskowski and Silver:

> *Bar*: Stefan and Eva are old friends of means at their local bar, planning to tie one on as they usually do. Stefan happens to catch the attention of the bartender before Eva, so he orders the first round of drinks. Stefan then says to Eva, "I bought the first round—please buy the next one."[6]

Laskowski and Silver note that there is something "off" about Stefan's request that Eva buy the next round. They note that Stefan has just done something generous for his friend, Eva, and that she knows that she has decisive normative reason to reciprocate.[7] Moreover, Stefan knows that Eva knows that she has decisive normative reason to reciprocate.[8] Importantly, for Laskowski and Silver, it is the request itself that is problematic. We can call the puzzle that Laskowski and Silver are interested in the *puzzle of requesting reciprocity*.

The puzzle of enticing right action is broader than the puzzle of requesting reciprocity in a few key ways. First, the former concerns not only making requests but also explicitly mentioning certain kinds of sufficient reasons for action (e.g., previous favors or prudentially relevant facts). In fact, I focus on explicitly mentioning these facts as opposed to merely making certain requests. Second, the former concerns not only enticing reciprocity but also enticing help that would not constitute reciprocity (e.g., as in Campaign 1). Third, the former concerns actions that one's intimates have either sufficient or decisive reason to perform, but the latter only involves cases in which the intimates have decisive normative reason to perform the relevant action. In fact, the puzzle of requesting reciprocity can be seen as a specific instance of the more general puzzle of enticing right action. After all, requesting is a way of enticing, as I have defined it, and the reciprocity that Laskowski and Silver are interested in

[5] Laskowski and Silver, "Wronging by Requesting."

[6] Laskowski and Silver, "Wronging by Requesting," 49.

[7] Laskowski and Silver, "Wronging by Requesting," 56.

[8] Laskowski and Silver, "Wronging by Requesting," 56.

is reciprocity that is morally right (i.e., permissible or required). So, my explanation of the puzzle of enticing right action can also be seen as an explanation of the puzzle of requesting reciprocity.

2. Requests, Trust, and Disrespect

In this section, I consider a plausible solution to the puzzle of enticing right action from Laskowski and Silver and argue that this solution is ultimately unsatisfactory.

Laskowski and Silver argue that people have a special moral obligation to trust their intimates to do what their intimates know they are morally required to do.[9] For ease of exposition, I will focus exclusively on the case Moving. Laskowski and Silver would likely point out that before Ronan makes his request and tries to entice Anna to help, Anna knows that she has decisive moral reason to help her friend move because he is her friend and it is an emergency. In addition, Ronan has helped her in the past. And Ronan knows that Anna knows that she has sufficient or decisive moral reason to help him move.

Laskowski and Silver's explanation of the wrongness of Ronan's enticing Anna to help is that it *expresses* a disrespectful belief—that is, the belief that the intimate will not do what they know they have decisive moral reason to do. The belief is especially disrespectful because it constitutes a failure to trust intimates to do what they know they morally ought to do. But, as they argue, we have a *pro tanto* moral obligation to trust intimates to do what they know they should.[10] The above enticements are expressions of disrespectful beliefs and are therefore disrespectful. Thus, according to this explanation, the enticers wrong their intimates by expressing a disrespectful belief.

More formally, the argument goes:

1. The enticers believe that their respective intimates will not do what the intimates know they are morally required to do.
2. Believing that an intimate will fail to do what they know they are morally required to do constitutes failing to trust an intimate to do what they know they morally ought to do.
3. Therefore, the enticers fail to trust their intimates to do what they know they morally ought to do.
4. Failing to trust an intimate to do what they know they morally ought to do constitutes disrespecting them.
5. Therefore, the enticers disrespect their intimates (by not trusting them).

[9] Laskowski and Silver, "Wronging by Requesting," 57–58.

[10] Laskowski and Silver, "Wronging by Requesting," 59.

6. In making their respective enticements, the enticers express their belief that their intimates will not do what they are morally required to do.
7. If one expresses a disrespectful belief about a person, then one disrespects that person.
8. Therefore, in making their respective enticements, the enticers disrespect their respective intimates.
9. If one disrespects another person, then one wrongs her.
10. Therefore, in making their respective enticements, the enticers wrong their respective intimates.

Thus, Laskowski and Silver's explanation of Problematic is that the enticements are troubling because they are morally wrong, and they are morally wrong because they are disrespectful.

Finally, they can explain Asymmetry by appealing to the fact that we only have a special obligation to trust our intimates to do what they know they morally ought to do. Because we lack this special obligation concerning nonintimates, our making similar requests of them is either not disrespectful or not *as* disrespectful.

The main problem with Laskowski and Silver's view is that premise 4 of the argument is false. That is, I think the following claim is incorrect: failing to trust an intimate to do what they know they morally ought to do constitutes disrespecting them. Without even saying much about intimate relationships, we can point to numerous cases in which, for various reasons, someone does not disrespect their intimate by failing to trust them in this way. For example, consider Leopold and Loeb, Thelma and Louise, or Bonnie and Clyde. Their intimate relationships were forged and expressed through immoral behavior (e.g., burglary, robbing banks, and even murder).[11] We can imagine that they knew what they were doing was wrong, but none of them trusted their partner to do what they each knew was the morally right thing—at least much of the time. In fact, they often trusted each other to do the opposite. But their failing to trust each other to do what they each knew was the morally right thing was not disrespectful at all.

Even setting these extreme examples aside, I think it is false that failing to trust an intimate to do what they know they morally ought to do constitutes disrespecting them. Whether or not we respect or disrespect our intimates depends on multiple factors. One factor is whether we treat our intimates in the way that they *want* or *ask* to be treated. In order to be a good friend, partner, spouse, sibling, child, parent, etc., we must listen to and respect the desires of

[11] For more on such cases, see Nehamas, "The Good of Friendship" and *On Friendship.*

our intimates.[12] Failing to do as our intimates ask is a way of disrespecting them. A friend might ask us to treat them objectively and not sugarcoat things. For example, they often want and request our honest and objective (as possible) opinion on their art, business plans, romantic interests, and other life choices.[13]

Sometimes, they might request that we be objective with regard to the morality of their behavior. For example, consider the following:

> *Addiction*: Ahmed and Jerome are close friends. Ahmed has been battling alcoholism. After months of treating his friends and family poorly, Ahmed reaches out to Jerome for help. He tells Jerome that he knows he has a problem and that he knows he is going to likely start drinking again instead of going home to spend time with his family, and he wants Jerome to help him stay on track. Ahmed tells Jerome that his desire to drink is strongest on his way home from work. One day, Ahmed texts Ahmed and tells him that his shift has ended, and he is headed home. Jerome does not trust Ahmed to go directly home.[14]

Jerome fails to trust Ahmed to do what both of them know Ahmed morally ought to do (i.e., go home). However, Jerome's belief is not disrespectful at all. In fact, it shows that he respects Ahmed's request for help and that he respects Ahmed as the kind of being who can autonomously request help.

Another part of being a good friend is actually caring about and being responsive to what is *objectively* good for them, even if they do not explicitly ask us to. That is, we ought to be sensitive to what contributes to or detracts from their objective well-being.[15] In these cases, we show respect by treating their well-being as being important. For example, even if Ahmed had not asked Jerome to not trust him and thus help him beat his addiction, Jerome might have already noticed Ahmed's addiction and been disposed to not trust him. But this lack of trust is not based on a poor view of Ahmed but rather on Jerome's concern for what is objectively best for his friend. Jerome wants to protect Ahmed from hurting himself, and the best way to do that is to cease trusting Ahmed to do what Ahmed knows is the right thing for him to do. So, it is hard to see how Jerome's lack of trust could be disrespectful. If it is objectively

[12] For example, see Cocking and Kennett, "Friendship and Moral Danger"; Ebels-Duggan, "Against Beneficence"; and Elder, "Why Bad People Can't Be Good Friends."

[13] Keller, "Friendship and Belief," 334; and Arpaly and Brinkerhoff, "Why Epistemic Partiality Is Overrated," 43.

[14] Arpaly and Brinkerhoff appeal to a structurally similar case to argue against the claim that part of being a good friend involves developing a disposition to overrate one's friends ("Why Epistemic Partiality is Overrated," 43).

[15] Elder, "Why Bad People Can't Be Good Friends"; and Brink, "Eudaimonism."

best for his friend that he not trust him as a default, and he ceases trusting his friend out of genuine concern, then it seems like he can lack this default trust and still be a good friend. And so, there can be cases in which being a good intimate does not require trusting one's intimate to do what they know they morally ought to do.

One might worry that these are special cases in which one has evidence from one's intimate that they are likely to fail to do what they know they are morally required to do. In the first case, Ahmed explicitly tells Jerome that he is unlikely to do what he knows he is morally required to do. In the second case, Ahmed shows Jerome (via his behavior) that he is unlikely to do what he knows he is morally required to do. So, perhaps what we are required to do is have a kind of *default* trust in our intimates. That is, perhaps Laskowski and Silver have the following view in mind:

> *Default Moral Trust*: When we have no evidence (from our experience with or testimony from our intimate) for or against the claim that our intimate will fail to do something they know they are morally required to do, we ought to believe that they will do what they know they are morally required to do.

The examples of Jerome and Ahmed and Bonnie and Clyde are not counterexamples to this view because in both cases each intimate has evidence *from their intimates* that the intimate will not do what they know they are morally required to do.

However, I do not think this default trust view is quite right either. This is because there will be cases in which your intimate's doing what they know to be the right thing will involve harming *you* or another one of their intimates. In such cases, I seriously doubt that you are always required to trust them to do the morally right thing. Consider the following case:

> *Cheating*: Imagine that you have cheated on a test, and your best friend knows about it. Moreover, you did not have a good reason or excuse. You simply did not want to study and decided to look at another student's answers. Imagine further that your school has an honor code that both you and your friend have promised to follow. Moreover, you have both promised to report anyone who has violated the honor code.

It looks like the right thing to do for your friend is to report you for cheating. It also looks like she knows that that is the right thing for her to do. But would it be disrespectful to fail to trust your friend to turn you in for cheating? I doubt it. In fact, if you believed that she would turn you in, *that* would be disrespectful. This is because it is plausible that you should trust your friends (and other intimates) to protect you or be loyal to you, even if it sometimes involves doing the wrong thing. As the saying goes, "A friend will help you move, but a good friend will help you move a body."[16] The fact that you should trust your friend to *not* turn you in is explained by the fact that being a good intimate requires one to trust one's intimates to do what they know they are required by demands of good intimate relationships to do.[17] This is because, as I just noted, protecting an intimate's well-being or being loyal to them is a demand of good intimate relationships.

3. Trust and Demands of Good Intimate Relationships

I think Laskowski and Silver are correct that the requesters fail to trust their intimates in a way that they are required to trust them. However, I do not think that the problem with this failure of trust is a *moral* problem. Rather, I think the source of the problem is that Ronan and Olivia fail to meet a demand of good intimate relationships. A demand of good intimate relationships is a rule the violation of which constitutes failing to be a *good* friend, parent, spouse, sibling, etc.[18] These demands are internal to intimate relationships in the way that there are demands of good chess playing, good novel writing, good hunting, etc.—that is, rules, the violation of which constitutes failing to do these things well.

More specifically, when I claim that something is a demand of good intimate relationships, I mean that one can justifiably be held accountable in particular ways if one violates one of these demands without an adequate excuse. That is, if one violates one of these demands (without an adequate excuse), then certain reactions are fitting. For example, if *A* violates a demand of good intimate relationships concerning her intimate *B*, it is fitting for *B* to (i) "take it

[16] For more on the possible conflicts between the demands of good friendship and morality, see Cocking and Kennett, "Friendship and Moral Danger"; and Koltonski, "A Good Friend Will Help You Move a Body."

[17] Following Aristotle in the *Nichomachean Ethics*, some have argued that true friendship is only possible between virtuous (or at least somewhat virtuous) people and that friendship involves mutual development of virtue (e.g., Sherman,"Aristotle on Friendship"; and Thomas, "Friendship"). This might seem to call into question whether the people in my example even count as true friends. However, I think the Aristotelean view presents an overly moralistic and unrealistic picture of friendship because it would exclude the relationships of Bonnie and Clyde, Thelma and Louise, and Leopold and Loeb as genuine friendships. For arguments that friendship does not require this kind of moral apprenticeship and can involve conflicts with morality, see Cocking and Kennett, "Friendship and Moral Danger"; Nehamas, "The Good of Friendship" and *On Friendship*; and Koltonski, "A Good Friend Will Help You Move a Body." I thank an anonymous referee for suggesting that I make explicit that my picture of friendship might conflict with this classic view.

[18] Stroud, "Epistemic Partiality," 502–3; and Keller, *The Limits of Loyalty*, 25–26.

personally"—for example, by feeling hurt, let down, or disappointed; and (ii) seek an explanation, excuse, or apology. In addition, it is fitting for *A* to (i) feel regret or guilt and (ii) be motivated to offer an explanation or apology, or to make it up to *B*.

We can distinguish demands of *good* intimate relationships from *constitutive* demands of intimate relationships. Examples of the latter kind of violation include cheating, backstabbing, and other kinds of large betrayals.[19] These kinds of violations threaten the intimate relationship itself. If *A* violates a constitutive demand of intimate relationships concerning *B*, then it is fitting for *B* to (i) "take it personally" by feeling rejected or hurt, (ii) feel resentment or contempt toward *A*, (iii) demand an explanation or apology, and (iv) be motivated to weaken or end her relationship with *B*. In addition, it is fitting for *A* to (i) feel regret or guilt and (ii) be motivated to offer an explanation or apology or to make it up to *B*.

3.1. *Intimate Relationships and Trust*

Good intimate relationships require trust. In particular, they require trusting one's intimates to follow what they *know* (or are in a position to know) are demands of good intimate relationships. In other words, being a good intimate requires trusting our intimates to be good intimates to us—at least when they know (or are in a position to know) what being a good intimate consists in. When I claim that being a good intimate requires trusting our intimates to be good intimates *to us*, I mean that we are only required to trust them to be a good intimate to *us,* and so we are not required to trust them to be good intimates to their *other* intimates.

When I say that being a good intimate requires one to *trust* one's intimate to follow what they know are demands of good intimate relationships, I mean the following: one should believe or be inclined to believe that one's intimate is following or will follow what they know are demands of good intimate relationships even when (a) one has no evidence for or against the proposition that one's intimate is following or will follow these demands, (b) one's evidence for and against this proposition is equally weighty, and (c) one has decent (but nonconclusive) evidence that they have failed or will fail to satisfy these demands. However, one is not required to believe or be inclined to believe that they are following or will follow these demands when (a) they tell one that

[19] For a similar distinction, see Shoemaker, "Attributability, Answerability, and Accountability," 621–22.

they are failing or will fail or (b) one has incontrovertible evidence that they are failing or will fail.[20]

The idea that one must trust one's intimates to follow what they know are demands of good intimate relationships meshes well with the current literature on doxastic (or epistemic) partiality. The broad idea here is that we are required by demands of good friendship to have doxastic states concerning our friends and their behavior that we are not required (by any normative domain) to have toward colleagues, associates, or strangers. For example, we are sometimes required by demands of good friendship to believe in ways that contravene epistemic demands (e.g., of apportioning one's beliefs or credences to one's evidence) when it comes to beliefs about our friends.[21] However, when it comes to colleagues, associates, or strangers, we ought only to believe in accordance with the relevant epistemic demands.[22]

Stroud gives the example of being told by a reliable testifier that one's friend has mistreated someone by sleeping with them and then knowingly not returning any of their phone calls.[23] She asks how a good friend ought to respond to this testimony. Roughly, she thinks that a good friend ought to stick up for their friends not only in their words but in their beliefs as well. Sticking up for one's friends in this way involves exerting more energy than one would exert for a stranger (a) to question and scrutinize damning evidence—for example, by thinking about ways in which the testifier might be untrustworthy, and (b) to look for less damning interpretations of one's evidence. Moreover, it involves giving more credence to these less damning interpretations. One will also try to fit the evidence into a pattern of behavior that is less damning for one's friend. Or, if one cannot do that, one will see some less than stellar attribute of one's friend as a less important part of that person such that one's overall impression of them is not damaged.[24]

While Stroud and other doxastic partialists have focused on sticking up for our friends when it comes to beliefs about their morality, I am focused solely on sticking up for our intimates when it comes to beliefs about whether they are being good intimates to us. Relatedly, I am arguing that when it comes to one's

[20] For example, see Baker, "Trust and Rationality," 3; Morton, "Partisanship," 177; Keller, "Friendship and Belief," 332–33; Stroud, "Epistemic Partiality in Friendship," 504–6; and Hazlett, *A Luxury of the Understanding*, 93–95.

[21] Keller, "Friendship and Belief"; Stroud, "Epistemic Partiality in Friendship"; and Hazlett, *A Luxury of the Understanding*.

[22] Baker, "Trust and Rationality"; Morton, "Partisanship"; Keller, "Friendship and Belief"; Stroud, "Epistemic Partiality and Friendship"; and Hazlett, *A Luxury of the Understanding*.

[23] Stroud, "Epistemic Partiality and Friendship," 504.

[24] Stroud, "Epistemic Partiality in Friendship," 504–9.

intimates only, one is required (by demands of good intimate relationships) to trust one's intimates to follow what they know to be demands of good intimate relationships. However, one is not required to trust colleagues, associates, or strangers to treat one in the way a good intimate is required to. Moreover, one is not even required to trust colleagues, associates, or strangers to follow what they know to be demands of good intimate relationships when it comes to *their* intimates.

3.2. Intimate Relationships and Motivation

What is seemingly problematic about the enticements in the above cases is that the enticers fail to trust their intimates to follow what they know their intimates know (or are in a position to know) is a demand of good intimate relationships. What demand of good intimate relationships is this? It is the demand that one ought to be especially motivated to act so as to protect or promote the desires, interests, needs, projects, and well-being of one's intimates for that intimate's own sake.[25] For the sake of brevity, I will hereafter just speak of protecting or promoting an intimate's "well-being" for their own sake. Under normal circumstances, the fact that an action would help promote or protect one's intimate's well-being should be sufficient for motivating one to perform that action. For example, if my friend needs help moving, then I should be motivated to help them, and I should be *more* motivated to help them move than I am to help a colleague, associate, or stranger move. And I should be motivated to help my friend move *for her sake* and not because it will benefit me—that is, I should not be motivated by prudential considerations.

Moreover, as Cocking and Kennett argue, one's motivation should not, at least sometimes, be "filtered" through one's own evaluative standard (e.g., one's own conception of morality or rationality, one's subjective tastes or attitudes).[26] Elizabeth Bennet in Jane Austen's *Pride and Prejudice* makes a similar claim:

> A regard for [my friend] would often make one yield readily to a request, without waiting for arguments to reason one into it.… In general and ordinary cases of friendship, where one is desired by the other to change a resolution of no very great moment, should you think ill of the person for complying with the desire, without waiting to be argued into it?[27]

[25] Stocker defends the view that people should perform certain actions (i.e., those concerning their friends) *out of* friendship ("The Schizophrenia of Modern Ethical Theories" and "Values and Purposes").

[26] Cocking and Kennett, "Friendship and Moral Danger," 285.

[27] Austen, *Pride and Prejudice*, 54–55.

In such cases, the mere fact that my intimate wants or has requested that I do something for them should be sufficient *by itself* to motivate me to do it.

Why think that it is a demand of intimate relationships that one be especially motivated to protect or promote the well-being of an intimate for that intimate's own sake? The answer is that having this disposition is partly constitutive of noninstrumentally valuing a person (i.e., valuing them for their own sake), and noninstrumentally valuing one's intimate is partly constitutive of being a good intimate.[28] That is, you cannot be a good intimate to someone unless you noninstrumentally value them.[29]

It is constitutive of valuing *X* that we are especially motivated to act so as to protect, preserve, or promote *X*. If I value my membership in some club or group, I will be especially motivated to act so as to ensure that I continue to be a member there—for example, by following the norms or rules of that club or group.[30] If I value my vintage car, I will be especially motivated to act so as to ensure that it does not get scratched, dented, or stolen.[31] Likewise, if I value a person, I will be especially motivated to act so as to protect or promote their well-being for that person's own sake. When a person *noninstrumentally* values something or someone, then one is motivated to act in these ways for the object's or person's own sake. And, given that noninstrumentally valuing a person is partly constitutive of being a good intimate to them, it follows that it is partly constitutive of being a good intimate that one is especially motivated to act so as to protect, conserve, and promote the well-being of one's intimates for that intimate's own sake (i.e., because it is *them*).

We can also look to cases to see that failing to be especially motivated to protect or promote the well-being of an intimate for their own sake makes one criticizable. Consider the following case from Stocker:

> Suppose you are in a hospital, recovering from a long illness. You are very bored and restless and at loose ends when Smith comes in once again. You are now convinced more than ever that he is a fine fellow and a real friend—taking so much time to cheer you up, traveling all the way across town, and so on. You are so effusive with your praise and thanks that he protests that he always tries to do what he thinks is his duty, what

[28] Scheffler, *Equality and Tradition*, esp. chs. 2 and 3.

[29] Aristotle, *Nichomachean Ethics*, 1156b7–11, 1159a11; Montaigne, *Essays*; Badhwar, "Friends as Ends in Themselves"; Blum, *Moral Perception*, 25; Frankfurt, *The Reasons of Love*, 42; Tiberius, *Well-Being as Value Fulfillment*, ch. 5. Of course, this is compatible with you also instrumentally valuing them.

[30] Scheffler, *Equality and Tradition*, ch 3.

[31] Lord, "Justifying Partiality," extends this account to valuing objects.

> he thinks will be best. You at first think he is engaging in a polite form of self-deprecation, relieving the moral burden. But the more you two speak, the more clear it becomes that he was telling the literal truth: that it is not essentially because of you that he came to see you, not because you are friends, but because he thought it his duty, perhaps as a fellow Christian or Communist or whatever, or simply because he knows of no one more in need of cheering up and no one easier to cheer up.[32]

Your friend is failing to be a good friend, because he failed to be sufficiently motivated to promote your well-being for *your own sake*. He failed to visit you *because it was you*. So, this case is further reason to think that good intimates are especially motivated to protect or promote the desires, interests, well-being, and so on of their intimate for that intimate's own sake.

3.3. *Failure to Trust*

Why should we think that the enticers fail to trust their intimates in Moving and Campaign 1? Before answering this question, it is important to keep the following in mind: Anna knows that helping Ronan move will protect or promote his well-being (and Ronan knows that Anna knows this). Ville knows that helping Olivia with her campaign will protect or promote her well-being (and Olivia knows that Ville knows this). Given that Ronan and Olivia have this knowledge about their respective intimates, we can ask: What doxastic attitude did Ronan and Olivia have concerning whether their intimate would be sufficiently motivated to help?

It might have been that they each believed that their intimate would be sufficiently motivated. But then their enticements would have been irrational. After all, if they each knew or believed that their respective intimate would be sufficiently motivated to protect or promote their well-being, then it would have made little sense to make the enticements. After all, if Ronan and Olivia believed that their respective intimate was sufficiently motivated to protect or promote their well-being, then they would have believed (or been disposed to believe) that their friend would help. After all, enticements have the aim of persuading the addressees to perform certain actions. Thus, under normal circumstances, enticements are only used if the enticer is agnostic or skeptical that the addressee will act in a certain way. On this interpretation of Moving and Campaign 1, the enticers act irrationally in making their enticements. The fact that they act irrationally on this interpretation, I think, gives us reason to

[32] Stocker, "The Schizophrenia of Modern Ethical Theories."

rule this interpretation out as a good interpretation of what is happening in these cases. After all, it does not seem that the enticers are being irrational.[33]

This leaves us with either interpreting the enticers as being agnostic or disbelieving that their intimates will be sufficiently motivated to protect or promote their well-being. In either case, the enticers are violating the aforementioned demand that one trust one's intimates to be sufficiently motivated to protect or promote the well-being of their intimates. Thus, on either interpretation, the enticers violate a demand of good intimate relationships.

Why, then, are their *enticements* troubling? They are troubling because they *indicate* or *provide evidence* that the enticers do not trust their intimates. This is distinct from Laskowski and Silver's view. Recall that they think the enticements express a disrespectful belief, which, in turn, entails that the enticers actually have this belief. This is because when some action expresses a belief, that belief nondeviantly causes the action. My view is that the enticements only provide strong evidence that the enticer has a certain belief and, therefore, that the enticer does not trust their intimate. So, it is compatible with my view that the enticer *does* trust their intimate to be a good intimate but nonetheless provides their intimate with strong evidence that they do not trust them.

In addition, the fact that the enticer makes certain kinds of enticements indicates that their intimates act in a certain way—that is, provides evidence that they failed to trust their intimates by *disbelieving* that they would act in accordance with the demands of their relationships. This is because, as Cocking and Kennett and Austen's character, Elizabeth Bennet, point out, people should be especially beholden or inclined to act in accordance with the well-being (i.e., interests and desires) of their intimates, and so merely informing or reminding one's intimates of what one's well-being consists in (e.g., informing or reminding them of one's interests or desires) is normally enough to get one's intimates to act in those ways. So, the fact that the enticers know that their intimates know what would protect or promote their well-being, and the enticers explicitly mention facts that have nothing to do with their well-being in order to entice their intimates, strongly indicates that they believe that their intimate will not be sufficiently motivated by their well-being.

If the enticers were merely agnostic about whether their intimates would be properly motivated by their well-being, they would have merely asked them if they were going to help. After all, if they were truly agonistic, they would be just

[33] I am assuming that the enticers are neurotypical. However, notice that even if *A* had a compulsion to try to entice their friend, *B*, to do something that *A* already believed *B* would do, it would be natural for *A* to apologize to *B* in advance. And it would also be natural for *A* to ask *B* for forgiveness in advance for her persistent requesting even though *A* trusted *B*. I thank an anonymous referee for this point.

as inclined to believe as to disbelieve that their intimates would be motivated by their well-being. Only if the enticers believed that their intimates were not going to be motivated by their well-being would it make sense for them to list reasons that are unrelated to their well-being.

So, my explanation of Problematic is this: the enticements in Moving and Campaign 1 seem problematic because they strongly indicate that the enticers are violating a demand of good intimate relationships—that is, the demand that they trust their intimates to be especially motivated to act so as to protect or promote their well-being for their own sake—at least when the intimates know (or are in a position to know) that this is a demand of good intimate relationships. My explanation of Asymmetry is this: while we are required by demands of good intimate relationships to trust our intimates to be especially motivated to protect or promote our well-being for our own sakes, we are not required (morally or otherwise) to trust nonintimates to have this motivation.

The explanation of the puzzle of enticing right action also explains why good intimates do not do each other favors *only on the condition* that the favor will be returned. The idea is simple: if we do favors only on the condition that our friends will return the favor, then we fail to be sufficiently motivated by their desires, needs, interests, projects, and well-being. That is, if their desires, needs, interests, projects, and well-being are truly *sufficient*, then, in many cases, one should not need any other considerations in order to be motivated to help. Of course, sometimes one needs more in the way of motivation because one's actions, while helping intimates, will come at some cost to one. But the point is that, *in general*, conditional giving or helping will violate the aforementioned demand of good intimate relationships. Thus, the demand of good intimate relationships that I appeal to provides a rationale for the common idea that friends and family should not count favors or help only on the condition that the favors will be returned.

Finally, the demand that an intimate be especially motivated to act so as to protect or promote their well-being for their own sake is derived from the more fundamental demand that an intimate noninstrumentally value their intimates to a high degree. However, noninstrumentally valuing one's intimate to a high degree involves a lot more than just being especially motivated to protect or promote that intimate's well-being for their own sake. Thus, it is helpful to talk about the derivative demand that one be especially motivated to protect or promote an intimate's well-being for their own sake in order to indicate precisely *how* an intimate might fail to be a good intimate.[34]

[34] I thank an anonymous referee for prompting me to be clearer about the relationship between noninstrumentally valuing an intimate and the demand to be especially motivated to protect or promote our intimates' desires, interests, well-being, etc.

4. Clarifying My Account

In this section, I clarify my account by making certain background assumptions explicit. First, my position does not require taking a side in the debate over whether our special duties to our intimates ultimately derive from moral duties.[35] Reductionists think that all the duties we have to intimates reduce to moral duties, while nonreductionists deny this.[36] Nonreductionists can admit that we do have moral duties toward our intimates in virtue of the features they share with other persons. So, for example, we all have moral duties to not kill, torture, or otherwise harm our intimates, and this is derived from the same moral duties that we have to refrain from treating nonintimates in these ways. However, nonreductionists insist that we have additional duties to our intimates in virtue of our special relationships with them, and these duties do not reduce to moral duties. I am inclined toward nonreductionism, but my explanation of Problematic and Asymmetry does require me to take a stand on this issue.[37]

Second, my position assumes that the normative landscape is not "flat." That is, I am assuming that there are different kinds of requirements, demands, and reasons (e.g., moral, prudential, epistemic). However, it might be the case that talk of "moral reasons" or "prudential reasons" is just talk. Fundamentally, there might just be flavorless requirements and flavorless reasons on which these requirements depend. If this is true, then, it might seem that my view is not very different from Laskowski and Silver's view.[38] For example, one might think that the difference would essentially be a disagreement about the source of the requirement to trust one's friends.

Dialectically speaking, my assumption that there are different kinds of requirements seems perfectly above board. Not only do Laskowski and Silver seem to make the same assumption, but this seems to be the orthodox view of

[35] I intend this point to apply not only to duties, but also to obligations, demands, reasons, and so on.

[36] Reductionists include Frankfurt ("On Caring"); McNaughton and Rawling ("Deontology"); and Hurka ("Love and Reasons").

[37] For defenses of nonreductionism, see Wallace, "Duties of Love"; and Brogaard, "Practical Identity and Duties of Love."

[38] Laskowski and Silver sometimes talk as if they think that normativity is flat in the aforementioned sense (e.g., "Wronging by Requesting," 57 and 60). However, they make it clear that one has a *moral* obligation to trust one's intimates and to not disrespect people ("Wronging by Requesting," 57, 58, 59). Moreover, the content of that trust is that one's intimates will do what they know is *morally* right or that they have the *moral* character to do what they know is morally right ("Wronging by Requesting," 58).

normativity.[39] In addition, I think something important is lost by flattening the normative landscape in the aforementioned sense. The reason is that different kinds of requirements seem to license different accountability practices.[40] As I argued above, violating a demand of good friendship (without an adequate excuse) makes certain responses fitting. For example, it is fitting for one's friend to feel hurt or disappointed and to seek an explanation or apology and for one to feel guilt or regret and to be motivated to provide an explanation or apology. However, I do not think violating a moral requirement (without an adequate excuse) makes the same responses fitting. For example, it is fitting for people who are wronged to feel resentment and contempt and to demand an explanation or apology; and it is fitting for the wrongdoer to feel guilt or remorse. So, if there is this tight connection between requirements of different kinds and different accountability practices, it is essential to distinguish different flavors of normativity—or at least different flavors of requirements.

Third, my view does not require some form of doxastic voluntarism—that is, the claim that one's beliefs are under one's direct voluntary control and so one can change one's beliefs at will. Nor does Laskowski and Silver's view. Recall that Laskowski and Silver claim that one has a moral obligation to not believe that one's intimates will fail to do what they know they morally ought to do. Plausibly, they also think that we should not suspend judgment about the matter either because that would be disrespectful to them. That is, it would be disrespectful to an intimate to be unsettled about whether they will do what they know they morally ought to do. This suggests that they think that there is a moral obligation to have a certain belief. On the other hand, my view is that one is required by demands of good intimate relationships to have certain beliefs about one's intimates (or the disposition to have these beliefs). For example, one should believe that one's intimate will satisfy the demands of good intimate relationships.

Given that both of our views put requirements on belief, one might suspect that both of our explanations assume doxastic voluntarism. Doxastic voluntarism is quite controversial, and so it would count against both views if they assumed it. Fortunately, however, neither of our views requires a commitment to doxastic voluntarism. Rather, we both just need a view on which a person can be, in some sense, normatively responsible for their beliefs.[41] Laskowski and

[39] Laskowski and Silver, "Wronging by Requesting," 50n3, 51, 54, 57, and 59.

[40] Darwall agrees because he thinks that moral requirements are conceptually tied to particular accountability practices ("Taking Account of Character," 20; "What Are Moral Reasons?" 5.)

[41] I will continue to talk about "moral" responsibility, but what I have in mind is the kind of responsibility needed for appropriately assessing our cognitive and noncognitive attitudes.

Silver need a view on which a person can be *morally* responsible for their beliefs, and I need a view on which a person can be held accountable for violating a demand of good intimate relationship. I will use the expression "normatively responsible" to indicate the kind of responsibility that is required for both Laskowski and Silver's and my view—which I will assume is roughly the same kind of responsibility. To say that someone is normatively responsible for ϕ (e.g., an action, belief, noncognitive attitude, character trait) is to claim that she is the fitting target of normative appraisal for ϕ.[42]

Fortunately, there are many views of *moral* responsibility that do not require the kind of voluntary control that we have over our actions, and we can just adopt any one of these views as a view of *normative* responsibility for beliefs.[43] So, if any of these (or related) views are correct, then neither Laskowski and Silver's nor my own view require doxastic voluntarism.

5. Distinguishing the Accounts

In this section, I further distinguish my account from Laskowski and Silver's.[44] First, recall that they argued that the enticements in Moving and Campaign 1 are disrespectful and are therefore morally wrong because they violate the *pro tanto* moral obligation to trust intimates to do what they know they have decisive reason to do. My account agrees with them that a core part of the problem has to do with failing to trust one's intimates. However, the content of the trust is importantly different. While Laskowski and Silver think that one must trust one's intimates to do what they know they have decisive *moral* reason to do, I am arguing that we must trust them to do what they know is required of them in order to be a *good intimate*. And, as we saw above, I think that demands of good intimate relationships need not be reduced to moral duties. So, the source of the requirement of trust is different. Second, my account holds that the content of the trust we are required to have is primarily about whether our intimate

[42] However, this does not tell us what kind of normative appraisal is fitting (e.g., blame, praise, indifference).

[43] For an overview of the kind of views of moral responsibility that do not require direct voluntary control, see Smith, "Responsibility for Attitudes." Such views include volitional views (e.g., Fischer and Tognazzini, "The Truth about Tracing"); endorsement views (e.g., Frankfurt, "Freedom of the Will and the Concept of a Person," "Identification and Externality," "Identification and Wholeheartedness"; and Locke and Frankfurt, "Three Concepts of Free Action"); rational relations views (e.g., Smith, "Responsibility for Attitudes"); and hybrid views (McKenna, "Putting the Lie on the Control Condition for Moral Responsibility").

[44] I thank an anonymous referee for pushing me to be clearer about how my account is related but different from Laskowski and Silver's.

will be especially motivated by facts about our well-being *qua* facts about our well-being and not *qua* morally relevant facts. That is, we are not required to trust our intimates to see facts about our well-being as *moral* reasons to help us but as *relationship-based* reasons to help us.

Finally, Laskowski and Silver's view assumes the possibility of doxastic wronging. A doxastic wronging occurs when one person wrongs another in virtue of having a belief with a certain content and not because of any of the consequences of their holding that belief.[45] Thus, on this view, one's beliefs can wrong other people in the sense in which one's actions can. Because of their commitment to doxastic wronging, Laskowski and Silver are committed to the view that there are practical reasons for belief—that is, that one can form or sustain a belief that *p* on the basis of a practical (i.e., moral) reason. In this section, I will explain why their view entails that there are moral reasons for belief, but mine does not. This difference matters not only for distinguishing the two views but also for revealing that their view, but not mine, entails a controversial thesis.

According to Laskowski and Silver, *A* has a *pro tanto* moral obligation to not believe that *A*'s intimate, *B*, will fail to do what *B* knows *B* is morally required to do. This is because it would be especially disrespectful to *B*, given *A* and *B*'s intimate relationship. It would also seem to be disrespectful for *A* to suspend judgment about whether *B* would do what *B* knows *B* is morally required to do. So, it seems like *A* has a moral obligation to not suspend judgment on the matter either. Therefore, it seems like *A* has a moral obligation to believe that *B* will do what *B* knows she is morally required to do. But, to have a moral obligation to have a certain belief concerning *B* is just to have a decisive moral reason to have a certain belief concerning *B*. So, *A* has a decisive moral reason to have a certain belief concerning *B*. So, there are moral reasons to have beliefs.[46] Moral reasons are a kind of practical reason. So, there are practical reasons for belief. However, it is quite controversial whether there are practical reasons for belief.[47]

My view, however, is not committed to practical reasons for belief. As I indicated above, all I mean in claiming that something is a demand of good intimate relationships is that if one violates one of these demands (without an adequate excuse), then certain reactions are fitting. For example, it is fitting

[45] Basu and Schroeder, "Doxastic Wronging," 181.

[46] Basu and Schroeder also admit that if there is doxastic wronging, then there must be moral reasons for doxastic states ("Doxastic Wronging," 190–94).

[47] For arguments against the possibility of practical reasons for belief, see Shah, "How Truth Governs Belief" and "A New Argument for Evidentialism"; Hieronymi, "The Wrong Kind of Reason" and "Controlling Attitudes"; and Schmidt, "On Believing Indirectly."

for one's intimate to feel hurt or disappointed and to seek an explanation or apology, and it is fitting for one to feel guilt or remorse.

However, this view is compatible with saying that one should try to do things to make oneself less likely to violate the demands of good intimate relationships. For example, one has a practical reason to try to become more trusting of one's intimates. But that is different from saying that one has a practical reason to hold certain doxastic states. So, in claiming that it is a demand of good intimate relationships that one trust one's intimate to be a good intimate on some occasion, I am not claiming that one has a practical reason to believe that they will be a good intimate on that occasion. Rather, one has a practical reason to try to get oneself to have this belief—if one does not already have it. If one has an adequate excuse for not having the belief—for example, one tried, but psychologically could not get oneself to have the belief—then one cannot be held accountable in the above-mentioned ways.

Why should we try to make ourselves more inclined to trust an intimate to be a good intimate to us? While I will not commit to any particular answer, there are a few plausible candidates. First, it might be that we should try to make ourselves more inclined to trust them because *that is what being a good intimate requires*. Second, it might be that we should try to make ourselves more inclined to trust them because they are our friend or our parent or our spouse, etc. That is, the answer might be that it is just part of having an intimate relationship with someone that we should try to make ourselves more trusting in the relevant way.

These explanations might seem insufficiently informative or deep. However, as I indicated above, I will not defend a particular view about the fundamental source of these special demands of good intimate relationships.[48] It might be that there is a suite of special demands of good intimate relationships that do not reduce to a single, fundamental demand. This would be an analog of Rossian pluralism about moral duties.[49] Alternatively, there might be one fundamental demand of good intimate relationships from which all other demands derive. This would be an analog of the monism about moral demands found in most normative ethical theories (e.g., Kantianism, consequentialism, and contractualism).

[48] For a view on the fundamental source of these duties or demands, see Brogaard, "Practical Identities and Duties of Love."

[49] Ross, *The Right and the Good*.

6. Conclusion

My explanation of Problematic is that the enticements in Moving and Campaign 1 strongly indicate that the requesters violate a demand of good intimate relationships. In particular, the enticements indicate that the enticers fail to trust their intimate to satisfy what the intimates know is a demand of good intimate relationships—that is, to be sufficiently motivated to protect or promote the desires, needs, interests, projects, and well-being of one's intimate for that intimate's own sake. My explanation of Asymmetry is that, regardless of whether we are morally required to trust nonintimates to do what they know is morally or prudentially right, the demand of intimate relationships that I mentioned (e.g., to be sufficiently motivated to protect or promote our well-being for our own sakes) only applies to our intimates.[50]

Yale University
max.lewis@gmail.com

References

Aristotle. *Nichomachean Ethics*. Translated and edited by Roger Crisp. New York: Cambridge University Press, 2004.

Arpaly, Nomy, and Anna Brinkerhoff. "Why Epistemic Partiality Is Overrated." *Philosophical Topics* 46, no. 1 (Spring 2018): 37–51.

Austen, Jane. *Pride and Prejudice*. Cambridge: Cambridge University Press, 2006.

Badhwar, Neera Kapur. "Friends as Ends in Themselves." *Philosophy and Phenomenological Research* 48, no. 1 (1987): 1–23.

Baker, Judith. "Trust and Rationality." *Pacific Philosophical Quarterly* 68, no. 1 (March 1987): 1–13.

Basu, Rima, and Mark Schroeder. "Doxastic Wronging." In *Pragmatic Encroachment in Epistemology*, edited by Brian Kim and Matthew McGrath, 181–205. New York: Routledge, 2019.

Blum, Lawrence A. *Moral Perception and Particularity*. New York: Cambridge University Press, 1994.

Brink, David O. "Eudaimonism, Love and Friendship, and Political

[50] I wish to thank Antti Kauppinen, Nick Laskowski, Yunjie Zhang, two anonymous referees for this journal, and an audience at the University of Helsinki for their feedback on the ideas in this paper. Special thanks to Nick Laskowski for providing me with written comments.

Community." *Social Philosophy and Policy* 16, no. 1 (Winter 1999): 252–89.

Brogaard, Berit. "Practical Identity and Duties of Love." *Disputatio* 13, no. 60 (May 2021): 27–50.

Chisholm, Roderick M., and Ernest Sosa. "Intrinsic Preferability and the Problem of Supererogation." *Synthese* 16, nos. 3–4 (December 1966): 321–31.

Cocking, Dean, and Jeanette Kennett. "Friendship and Moral Danger." *Journal of Philosophy* 97, no. 5 (May 2000): 278–96.

———. "Friendship and the Self." *Ethics* 108, no. 3 (April 1998): 502–27.

Cohoe, Caleb Murray. "God, Causality, and Petitionary Prayer." *Faith and Philosophy* 31, no. 1 (2014): 24–45.

Darwall, Stephen. "Taking Account of Character and Being an Accountable Person." In *Oxford Studies in Normative Ethics,* vol 6, edited by Mark Timmons, 12–36. New York: Oxford University Press, 2016.

———. "What Are Moral Reasons?" *Amherst Lecture in Philosophy* 12 (2017): 1–24.

Driver, Julia. "The Suberogatory." *Australasian Journal of Philosophy* 70, no. 3 (1992): 286–95.

Ebels-Duggan, Kyla. "Against Beneficence: A Normative Account of Love." *Ethics* 119, no. 1 (October 2008): 142–70.

Elder, Alexis. "Why Bad People Can't Be Good Friends." *Ratio* 27, no. 1 (March 2014): 84–99.

Enoch, David. "Giving Practical Reasons." *Philosophers' Imprint* 11, no. 4 (2011): 1–22.

Fischer, John Martin, and Neal A. Tognazzini. "The Truth about Tracing." *Noûs* 43, no. 3 (September 2009): 531–56.

Frankfurt, Harry G. "Freedom of the Will and the Concept of a Person." *Journal of Philosophy* 68, no. 1 (January 1971): 5–20.

———. "Identification and Externality." In *The Identities of Persons,* edited by Amélie Oksenberg Rorty, 239–52. Los Angeles: University of California Press, 1976.

———. "Identification and Wholeheartedness." In *Responsibility, Character, and the Emotions: New Essays in Moral Psychology,* edited by Ferdinand David Schoeman, 27–45. New York: Cambridge University Press, 1987.

———. "On Caring." In *Necessity, Volition, and Love,* 155–80. New York: Cambridge University Press, 1998.

———. *The Reasons of Love*. Princeton: Princeton University Press, 2006.

Hazlett, Allan. *A Luxury of the Understanding: On the Value of True Belief.* Oxford: Oxford University Press, 2013.

Herstein, Ori J. "Understanding Standing: Permission to Deflect Reasons." *Philosophical Studies* 174, no. 12 (December 2017): 3109–32.

Hieronymi, Pamela. "Controlling Attitudes." *Pacific Philosophical Quarterly* 87, no. 1 (March 2006): 45–74.

———. "The Wrong Kind of Reason." *Journal of Philosophy* 102, no. 9 (September 2005): 437–57.

Hurka, Thomas. "Love and Reasons: The Many Relationships." In *Love, Reason and Morality*, edited by Katrien Schaubroeck and Esther Kroeker, 163–80. New York: Routledge, 2016.

Keller, Simon. *The Limits of Loyalty*. New York: Cambridge University Press, 2007.

Koltonski, Daniel. "A Good Friend Will Help You Move a Body: Friendship and the Problem of Moral Disagreement." *Philosophical Review* 125, no. 4 (October 2016): 473–507.

Laskowski, N. G., and Kenneth Silver. "Wronging by Requesting." In *Oxford Studies in Normative Ethics*, vol. 11, edited by Mark Timmons, 49–69. New York: Oxford University Press, 2021.

Lewis, James H. P. "The Discretionary Normativity of Requests." *Philosophers' Imprint* 18, no. 20 (2018): 1–16.

Locke, Don, and Harry Frankfurt. "Three Concepts of Free Action." *Aristotelian Society Supplementary Volume* 49, no. 1 (July 1975): 95–125.

Lord, Errol. " Justifying Partiality." *Ethical Theory* and *Moral Practice* 19, no. 3 (June 2016): 569–90.

McKenna, Michael. "Putting the Lie on the Control Condition for Moral Responsibility." *Philosophical Studies* 139, no. 1 (2008): 29–37.

McNaughton, David, and Piers Rawling. "Deontology." In *The Oxford Handbook of Ethical Theory*, edited by David Copp, 424–58. New York: Oxford University Press, 2006.

Montaigne, Michel de. *Essays*. Translated by J. M. Cohen. New York: Penguin Books, 1958.

Morton, Adam. "Partisanship." In *Perspectives on Self-Deception*, edited by Brian P. McLaughlin and Amelie O. Rorty, 170–82. Los Angeles: University of California Press, 1988.

Nehamas, Alexander. "The Good of Friendship." *Proceedings of the Aristotelian Society* 110, no. 3 (October 2010): 267–94.

———. *On Friendship*. New York: Basic Books, 2016.

Owens, David. *Shaping the Normative Landscape*. Oxford: Oxford University Press, 2012.

Raz, Joseph. *The Morality of Freedom*. New York: Oxford University Press, 1986.

———. *Practical Reason and Norms*. Princeton: Princeton University Press, 1990.

Ross, William David. *The Right and the Good*. Oxford: Oxford University Press, 1930.

Scheffler, Samuel. *Equality and Tradition: Questions of Value in Moral and Political Theory*. Oxford: Oxford University Press, 2010.

Schmidt, Sebastian. "On Believing Indirectly for Practical Reasons." *Philosophical Studies* 179, no. 6 (June 2022): 1795–819.

Shah, Nishi. "How Truth Governs Belief." *Philosophical Review* 112, no. 4 (October 2003): 447–82.

———. "A New Argument for Evidentialism." *Philosophical Quarterly* 56, no. 225 (October 2006): 481–98.

Sherman, Nancy. "Aristotle on Friendship and the Shared Life." *Philosophy* and *Phenomenological Research* 47, no. 4 (June 1987): 589–613.

Shoemaker, David. "Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility." *Ethics* 121, no. 3 (April 2011): 602–32.

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

Stocker, Michael. "The Schizophrenia of Modern Ethical Theories." *Journal of Philosophy* 73, no. 14 (August 1976): 453–66.

———. "Values and Purposes: The Limits of Teleology and the Ends of Friendship." *Journal of Philosophy* 78, no. 12 (December 1981): 747–65.

Stroud, Sarah. "Epistemic Partiality in Friendship." *Ethics* 116, no. 3 (April 2006): 498–524.

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

Thomas, Laurence. "Friendship." *Synthese* 72, no. 2 (August 1987): 217–36.

Vargas, Manuel. "The Trouble with Tracing." *Midwest Studies in Philosophy* 29, no. 1 (September 2005): 269–91.

Wallace, R. Jay. "Duties of Love." *Aristotelian Society Supplementary Volume* 86, no. 1 (June 2012): 175–98.

Journal of Ethics and Social Philosophy
VOL. 27, NO. 3 · MAY 2024

https://doi.org/10.26556/jesp.v27i3.3346

ELIZABETH ANSCOMBE ON MURDER

Joshua Stuchlik

The topic of murder was among Elizabeth Anscombe's central preoccupations. She garnered international attention in 1956 for protesting Oxford's decision to award an honorary degree to former US President Harry Truman, and the ground of her opposition was that in authorizing the use of atomic weapons at Hiroshima and Nagasaki, Truman had authorized the mass murder of Japanese civilians.[1] After receiving a chair of philosophy at Cambridge, she taught seminars and lectured on the topic of killing human beings for four years in the early 1970s.[2] Concern with the theme of murder is also evident in many of Anscombe's writings. Her indictment of Oxford moral philosophy in "Modern Moral Philosophy" is based on the preparedness of these philosophers to approve of murder and other intrinsically bad types of action. The prohibition on murder appears in her essays on the ethics of war, euthanasia, the principle of double effect, and political philosophy.[3] Finally, Anscombe's seminal work in action theory, *Intention*, investigates a topic that is integral to her account of murder and is based on lectures that may have been provoked by critics of her protest of Truman's degree.[4]

In addition to her published writings, scholars now have access to the archive of Anscombe's unpublished papers, which are held by the Collegium Institute at the University of Pennsylvania. Anscombe takes up the topic of murder in a number of these papers. Some appear to be lecture notes and essay drafts, and others consist in handwritten notes. My goal in this paper is to reconstruct Anscombe's theory of murder by integrating material from the archive with her published writings. Of course, we should be cautious about the canonical status of her unpublished drafts and notes. But a holistic analysis promises to add context, depth, and richness to her published work.

[1] Anscombe, *Mr. Truman's Degree*, 64.

[2] Gormally, "On Killing Human Beings," 133.

[3] See Anscombe on the ethics of war (*Justice of the Present War Examined*, *Mr. Truman's Degree*, and "War and Murder"), on euthanasia ("Murder and the Morality of Euthanasia"), on the principle of double effect ("Action, Intention, and 'Double Effect'"), and on political philosophy ("On the Source of the Authority of the State.")

[4] Geach, "Introduction," xiii–xiv.

The archival documents reveal that Anscombe was working toward a systematic theory of murder—as she puts it in one paper, "an enquiry into what constitutes murder, and what should be our attitude towards it."[5] This quotation indicates the two major aims of her project. The first is the explanatory aim of providing an account of murder and the conditions under which an agent's conduct constitutes murder. The second aim is ethical and involves providing an account of the normative basis of the prohibition on murder. My focus will be on the first of these aims, and I will only briefly address the second in the conclusion. In section 1, I discuss the context of Anscombe's writings on murder. I also explain why Anscombe was concerned to deny the semantic thesis that "murder" means "unjustified or impermissible killing," and I detail three challenges that an account of murder that rejects the semantic thesis must surmount. Sections 2 to 4 reconstruct Anscombe's theory in a way that enables an answer to each of these challenges.

1. Context and Challenges

John Berkman persuasively argues that Anscombe's work in moral philosophy leading up to and including "Modern Moral Philosophy" was driven by her conviction that there is an absolute prohibition on murder.[6] Anscombe's concern with the theme of murder is already present in a pamphlet she published in 1940 with Norman Daniel, *The Justice of the Present War Examined*. In it, Anscombe and Daniel criticize the British government for threatening to attack the civilian population of Germany from the air if the Germans did it first. They claim that deliberately attacking civilians constitutes murder and that a policy of doing so would render Britain's war unjust.[7] Seventeen years later, in *Mr. Truman's Degree*, Anscombe recounts the series of events that led the Allies to adopt the strategy of making indiscriminate attacks on civilian populations, a strategy that culminated in the atomic bombings.[8]

[5] Box 8, file 291, 1, Collegium Institute Anscombe Archive at the University of Pennsylvania, Kislak Center for Special Collections, Rare Books and Manuscripts (CIAA). I cite archival documents by box number, file number, and page number where applicable. Unfortunately, none of the archival writings I refer to are dated. There is significant overlap in their content with the published essays "Murder and the Morality of Euthanasia" (1982), "Prolegomenon to a Pursuit of the Definition of Murder" (1979), "On the Source of the Authority of the State" (1978), and "Action, Intention, and 'Double Effect'" (1982).

[6] Berkman, "Justice and Murder." Jennifer Frey also argues that the three theses of "Modern Moral Philosophy" are unified by the issue of absolute prohibitions, which include the prohibition on murder ("Revisiting Modern Moral Philosophy").

[7] Anscombe, *Justice of the Present War Examined*, 79.

[8] Anscombe, *Mr. Truman's Degree*, 62–64.

Anscombe saw Oxford's decision to award Truman an honorary degree as evidence that its members were unable to affirm the proposition that murder is always to be condemned. In the penultimate paragraph of *Mr. Truman's Degree*, she claims this failure is reflected in the two major systems of Oxford moral philosophy since the First World War (the systems of Ross and Hare), both of which "contain a repudiation of the idea that any class of actions, such as murder, may be absolutely excluded."[9]

Anscombe continues her attack on Oxford moral philosophy in "Modern Moral Philosophy" (originally published in 1958). There she proposes three theses, the third of which is that "the differences between the well-known English writers on moral philosophy from Sidgwick to the present day are of little importance."[10] The claim would likely have struck its intended targets as bizarre. Sidgwick and Moore were consequentialists, while Prichard and Ross were deontologists; and all these philosophers were metaethical realists, and hence disagreed with noncognitivists such as Ayer and Hare. However, in Anscombe's mind, the differences between these philosophers were eclipsed by their shared rejection of absolute prohibitions. In rejecting these, the well-known English moral philosophers contradict the "Hebrew-Christian ethic," which holds that there are some types of action that are morally impossible—forbidden whatever *consequences* threaten:

> The prohibition of certain things simply in virtue of their description as such-and-such identifiable kinds of action, regardless of any further consequences, is certainly not the whole of the Hebrew-Christian ethic; but it is a noteworthy feature of it; and, if every academic philosopher since Sidgwick has written in such a way as to exclude this ethic, it would argue a certain provinciality of mind not to see this incompatibility as the most important fact about these philosophers, and the differences between them somewhat trifling by comparison.[11]

Among the prohibitions characteristic of the "Hebrew-Christian ethic" is the commandment "Thou shalt not kill." In an unpublished lecture titled "Killing and Murder," Anscombe claims that the commandment is better translated as "Thou shalt do no murder."[12] The commandment is meant to be action guiding: one is supposed to be able to use it to reason to someone (including oneself)

[9] Anscombe, *Mr. Truman's Degree*, 71.

[10] Anscombe, "Modern Moral Philosophy," 26.

[11] Anscombe, "Modern Moral Philosophy," 34.

[12] CIAA 13.511.2; cf. 8.298.W2. Anscombe is on firm interpretive ground here. See Markl, "The Decalogue," 18.

that she ought to abstain from some contemplated action on the ground that to do it would be to commit murder. In other words, the commandment is supposed to be apt for use in arguments of the following form:

1. *This* would be an action of such-and-such a kind.
2. An action of that kind is murder.
3. Therefore, to do *this* would be to commit murder.

Anscombe calls this "a powerful piece of practical reasoning."[13] For the agent considering it may have a will not to be murderous, and if so, she will be moved by it to abstain from the contemplated action.[14]

There is one claim, however, whose truth would undermine this pattern of practical reasoning. That is the claim that murder just means "killing that is wrong or unjustified." Anscombe raises this possibility in "Killing and Murder":

> Nowadays if this question [i.e., the question "Is murder ever permissible?"] is asked, someone will say that the word "murder" simply means impermissible killing. If that should be true, "Thou shalt do no murder" only means "Thou shalt not kill human beings in cases where thou shalt not." Then the commandment, however venerable, will be *no* sort of contribution to an argument that something is damnable because it is murder. For it will always *first* have to be determined that the killing of someone is wrong before it can be determined whether it is murder, and so it cannot effectively be argued that it should not be done because it is murder. It will also cease to be a substantial question whether it can be justifiable to murder.[15]

Let us call the claim that "murder" means "unjustified or impermissible killing" the *semantic thesis*. If the semantic thesis is correct, it will not be possible to argue that one ought not to do a certain action on the ground that it would

[13] CIAA 13.511.5; 8.298.W9.

[14] The argument form is in turn a specification of a more general pattern of practical reasoning:

1. *This* would be an action of such-and-such a kind.
2. An action of that kind is *X*.
3. Therefore, to do *this* would be to do *X*.

Other possible values of *X* include: an act of dishonesty, cruelty, cowardice, treachery, adultery, or theft (CIAA 13.511.5; 8.298.w8).

[15] CIAA 13.511.2; 8.298.w2; italicized words underlined in the ms.; cf. Anscombe, "Prolegomenon," 257. David Albert Jones also calls attention to this passage ("Anscombe on Euthanasia as Murder," 273).

be murder.[16] Furthermore, the truth of the semantic thesis would undermine the debate between Anscombe and the Oxford moral philosophers. There is supposed to be a substantive disagreement between them on the question of whether murder is always forbidden. But if the semantic thesis is true, there will be universal agreement that murder is always forbidden, but only for the uninteresting reason that it is wrong by definition. Anyone who thinks, for instance, that the bombings of Hiroshima and Nagasaki were justified on consequentialist grounds would have good reason simply to deny that the civilians killed were murdered.

Against the semantic thesis, Anscombe believes that G. E. Moore was right in *Principia Ethica* to ask the question, "Is murder ever right?"[17] In raising this question, Moore showed he understood that the question whether murder is always unjustified is a substantive one, not one that can be settled by conceptual analysis. But if the question whether murder is always unjustified is substantive, then it is imperative to provide a philosophical account of murder. Moreover, such an account must overcome three challenges, each of which involves explaining how to accommodate certain features of murder that seem to suggest that wrongness or unjustifiability *is* built into the concept of it. They are as follows:

1. The concept of culpability is built into the concept of murder.

In "The Two Kinds of Error in Action," Anscombe makes this point when she claims that *formality* is essential to murder.[18] In this respect, murder may be contrasted with adultery. Suppose a man has sexual relations with a woman he has every reason to believe he has married, but in fact she is married to someone else, and so by the laws of his society, not to him. He has satisfied the definition of adultery, which is sexual intercourse between a married person and someone who is not his or her spouse. Yet it would be unduly harsh to find him guilty of committing adultery. According to Anscombe, the way to resolve any perplexity here is to say that while the man performed actions that were materially acts of adultery, he did not formally commit adultery.[19] Although he did have sexual relations with a woman who is not his wife, his blameless ignorance means he is not culpable for doing so, and this is what is signaled by saying he did not formally commit adultery. By contrast, if a man pours a

[16] CIAA 13.511.4; 8.298.W6.

[17] CIAA 13.511.10; 8.298.W14. Anscombe likely has in mind sec. 95 of *Principia Ethica*, where Moore attempts to provide a consequentialist proof that murder is generally to be avoided.

[18] Anscombe, "Two Kinds of Error in Action," 5.

[19] Anscombe, "Two Kinds of Error in Action," 4–5.

drink for his wife that he reasonably believes is gin but is in fact petrol, and she dies, he has not committed an act of "material murder." The proof that the man could not have reasonably known the liquid was petrol is a proof that we are not dealing with an act of murder at all, but a tragic accident.

What this sort of case shows is that culpability is built into the concept of murder. But if culpability is built in, does that not show that wrongness or unjustifiability is as well? Anscombe identifies this challenge in an archival document, where she observes that it seems that any justification for doing something that causes a person's death will remove culpability and, in doing so, show that the deed is not to be called "murder."[20]

2. The verdict about whether to call particular killings murderous sometimes requires determining whether they are justified by other considerations such as proportionality and necessity.

Consider collateral damage to noncombatants that occurs when military targets are attacked in war. Many people believe that collateral killings are not always murder. But that does not mean that such killings are always ethically in the clear. If the war is manifestly unjust, or the number of noncombatant deaths is grossly disproportionate to the value of the target, or the attack is wholly unnecessary for achieving the aims of the war, then the noncombatants have been murdered. In these cases, we *do* first decide whether killings are justified, and if they are not, we judge them to be murder.

3. In many familiar debates, whether one is for or against the permissibility of a certain general *type* of killing determines whether one believes that type of killing is murderous.

For instance, pacifists and just war theorists disagree about whether killing enemy combatants is ever justified. Pacifists hold that it is not and so say that all killing in war is murder, whereas just war theorists hold that killing enemy combatants is sometimes justified and so deny it always constitutes murder. Debates about capital punishment, abortion, and euthanasia have a similar shape. This may seem to suggest, in turn, that a decision whether to call a killing murderous awaits upon a prior judgment of whether it is justified. Anscombe flags this problem in relation to capital punishment and assassination, writing:

> Now if capital punishment, or if the assassination of a tyrant, are kinds of action which are possibly justifiable as such, then they are not *as such* murder. This is the fact that most helps to make it look as if the notion

[20] CIAA 8.292.

> of wrongfulness *is* built into the notion of murder: a thing there is hardly any temptation to say in the case of adultery.[21]

In the following three sections I reconstruct Anscombe's theory of murder with a view to showing how it provides a response to each of these challenges.

2. MURDER AND RESPONSIBILITY

Let us begin with the idea that formality is essential to murder. Anscombe says that a modern way of putting this is to say that *responsibility* is built into the concept.[22] Responsibility is at the heart of Anscombe's account of murder, for "murder is killing which involves a special degree and kind of responsibility for death."[23] It is this notion of responsibility that explains why murder is more than simply killing or causing the death of a human being but also requires a mental element.

2.1. The Three Levels of Responsibility

The statement that murder involves a special degree and kind of responsibility for death implies there are different sorts of responsibility. In "Murder and the Morality of Euthanasia" and her archival papers, Anscombe distinguishes between three levels of responsibility. In the unpublished essay "Intention and Responsibility," she calls them (1) causality, (2) accountability, and (3) creditability.[24]

At the first level, to say that *S* is responsible for event *E* is to say that *S* is a cause (or condition) of *E*, and to say that *S* is not responsible is to say that he is not a cause of *E*. Even inanimate objects can be responsible at this level: the wind may be responsible for breaking a vase, and a stroke of lightning may be responsible for a wildfire.

The second level of responsibility is accountability or callability to account. To call someone to account for some action or omission is to request or demand an explanation for it, one that is couched in terms of her reasons for acting, or not acting, as she did. Since only a rational agent can give an account, a necessary condition for accountability for some action at time *t* is that the one being called to account is a rational agent able to exercise her rational capacities at *t*.

[21] CIAA 13.511.6-7; 8.298.W10; italicized words underlined in the ms.; cf. Anscombe, "Prolegomenon," 255.

[22] CIAA 13.511.7; 8.298.W11.

[23] Anscombe, "Murder and the Morality of Euthanasia," 261.

[24] CIAA W5.474.4.

A rational agent is not accountable for just anything she does but only for her voluntary actions and omissions and their effects.[25]

The sphere of the voluntary is wider than that of the intentional. Expected side effects are not intentional: if I foresee that I will get a stomachache from taking my medication, it is no part of my aim or purpose to induce a stomachache, but I do bring it about voluntarily. If I am called to account for giving myself a stomachache, my explanation will not refer to the desirability of having one but to the fact that it was a necessary accompaniment of the medical benefit I was trying to achieve. Indeed, on Anscombe's view, the sphere of the voluntary extends to some things the agent is not aware of doing or bringing about. A person's doing such-and-such is voluntary although she is ignorant that she is doing such-and-such when the ignorance is itself voluntary. Ignorance that *p* is voluntary not only when the agent chooses not to do something she knows would result in her learning that *p* (say, because she would prefer not to know whether *p*), but also when she could and should have found out that p but failed to do so because she was negligent or careless.[26]

The third level of responsibility—creditability—comes into play when the agent is responsible at the second level for some effect that is good or evil. In the case of an evil, the agent will bear this sort of responsibility for the evil when (1) she is accountable for it and (2) she lacks an exonerating account for bringing it about. In that case, the agent is *guilty* of bringing about the evil and is appropriately blamed for it.[27]

2.2. Murder as Guilt for Causing Death

When Anscombe says that murder is killing that involves a special kind of responsibility for death, she means level three responsibility.[28] A murderer causes the death of another human being, he is accountable for that death, and he lacks an exonerating account. If you murder someone, then the evil of his death "lies at your door" and "his blood is on your head."[29] Moreover, since murder is a grave injustice, you have seriously wronged the victim, and the

[25] CIAA W5.474.4.

[26] Anscombe, "Two Kinds of Error in Action," 8–9. The latter is a special case of bringing about something by omission—namely, an omission to find out that *p*. I discuss omissions in section 2.2.

[27] CIAA W5.474.4; Anscombe, "Murder and the Morality of Euthanasia," 262.

[28] When Anscombe says that murder involves a special *degree* of responsibility, she likely has in mind degree of guilt. Manslaughter, or killing with culpable negligence, involves a lower degree of culpability than murder (CIAA 8.290.10). See also Anscombe, "Murder and the Morality of Euthanasia," 264.

[29] CIAA 8.292.

evil of this injustice is also imputable to you.[30] To say that someone is guilty of murder is therefore pleonastic, whereas it is not pleonastic to say that someone is guilty of adultery.[31]

It follows that a philosophical account of murder will involve delineating the factors that can exonerate an agent from bearing guilt for causing another's death. Anscombe notes that there are a variety of exonerating answers:

> One who is callable to account may not be guilty, even though he did cause death, because there is an exonerating answer. The range of such answers is very wide: "He was sleep-walking"; "He stumbled"; "He did not know he was administering poison"; "He did not intend death but something else which was quite legitimate"; "He was acting with legitimate authority"; "He had no duty to prevent death."[32]

We can classify these exonerating responses into five types:

1. The agent's responsibility for death is only level one responsibility.

Answers such as "I was sleep-walking" or "I stumbled" are of this type. Strictly speaking, this is not an exonerating answer, for the need for exoneration presupposes an instance of voluntary agency for which the agent is accountable. But if the agent's responsibility is only level one, then there is no voluntary action for him to account *for*.

2. The agent was ignorant that he was doing something that would cause or risk causing death, where the ignorance is not due to negligence.

"I didn't know I was administering poison" is of this type if I had every reason to believe I was pouring a glass of gin. In this case, there is a voluntary action for which I can be called to account (pouring the drink), but my blameless ignorance that the liquid was poisonous means that I was not voluntarily *poisoning* or *killing* the victim—though I am level one responsible for her death.

3. The agent did not intentionally kill the victim, but the death was a foreseen side effect of some course of action.

[30] CIAA 8.289. Anscombe thinks this fact explains why suicide is not a form of murder (self-murder), for it is, strictly speaking, impossible to wrong oneself (CIAA 8.290.10–12). This reflects an Aristotelian and Thomistic conception of justice as the virtue whose sphere is one's relations with others (see Aristotle, *Nicomachean Ethics* bk. 5, ch. 1; Aquinas, *Summa Theologiae* II-II 58.2).

[31] CIAA 8.296.8.

[32] Anscombe, "Murder and the Morality of Euthanasia," 262.

As we shall see in section 3, on Anscombe's view, this sort of response can *sometimes* exonerate the agent from guilt for causing death.

4. The agent intentionally killed the victim, but the agent was exercising legitimate authority in killing her.

This sort of response involves a title to kill, which I discuss in section 4.

5. The victim's death could have been prevented if the agent had ϕ-ed, but the agent had no duty to ϕ.

The ability of this type of answer to exonerate depends on the way we attribute omissions and their effects. An omission cannot be identified with an absence of motion. Even an omission can be omitted. I might be expected, for example, to omit every third name on a list, and when I come to the eighteenth name I write it down and thereby omit to omit it.[33] A person omits to ϕ not merely when she does not ϕ, but when she does not ϕ in circumstances in which she *could* have done so and it was in some sense *expected* that she would. For instance, a cook spoils the potatoes by omitting to add salt because it is the cook's business to add the salt. It is not the business of the restaurant guests to add salt, so even if it was physically possible for them to add salt they do not *omit* to do so, and their not adding salt is not the cause of the potatoes' being spoiled.[34] Following Aquinas, Anscombe summarizes this by stating that an agent causes an effect by omission when it was both "possible" and "necessary" that he should ϕ and he does not.[35] When it comes to preventing death, the relevant sense of "necessity" is moral necessity. Thus, an agent will only be level three responsible for causing someone's death by omission when he could have prevented her death by adopting some means and he had a duty to adopt those means. Moreover, there is no duty to adopt means that involve wronging some people in order to prevent harm from befalling others. "I should have to commit a great wrong" is a plea of moral *impossibility*.[36]

It is crucial to note one possible justification for causing death that does *not* appear on the list of exonerating responses—namely, a justification that refers to the advantages to be gained or disadvantages to be avoided by killing an innocent person.[37] Modern moral philosophers frequently devise scenarios

[33] CIAA 12.539.1.

[34] Anscombe, "Two Kinds of Error in Action," 9.

[35] Anscombe, "Murder and the Morality of Euthanasia," 273. See Aquinas, *Summa Theologiae* I-II 6.3.

[36] CIAA 12.539.11.

[37] Anscombe, "Murder and the Morality of Euthanasia," 262.

in which an agent can kill one or more innocent people as a means to saving a greater number. If this is admitted as a justification, then, Anscombe insists, it is not a justification that exonerates the agent from the guilt of having committed murder. Rather, it is a justification *for murder*.

Why this should be the case is an important question, and I return to it in section 3. At this juncture, I want to show how this claim enables Anscombe to reply to the first challenge from section 1. Recall that she denies that unjustifiability or impermissibility are part of the concept of murder. However, she also asserts that culpability is built into the concept—which is now understood as level three responsibility or guilt. The problem was that if guilt is built into the concept of murder, then it appears that unjustifiability must be as well; for it seems that any justification will remove guilt and so prove the action is not a case of murder.

Anscombe's response is to deny the premise that any (alleged) justification for doing something that causes someone's death will remove the guilt of committing murder. If Alfred kills innocent Betty in order to save five others from being killed, then Alfred has murdered Betty. If a philosopher thinks that Alfred's killing Betty is justified, then what he thinks is justified is Alfred's incurring the guilt of murder.

I noted above that murder is an injustice that wrongs its victim. I think that Anscombe would also reject the claim that this implies that murder is unjustified as a matter of definition. She would no doubt argue that it is also a substantive question whether a person could ever be justified in wronging others. That seems right. Jeff McMahan claims that an agent can sometimes act with "objective moral justification" and yet inflict harm that wrongs its victim. McMahan thinks this can occur, for instance, when an innocent person's rights are overridden by sufficiently strong consequentialist considerations.[38] McMahan's claim does not seem to be conceptually incoherent. The disagreement between him and someone who thinks it is always wrong to commit an injustice is a substantive one.

3. The Significance of Intention

We saw in section 2.2 that according to Anscombe, the distinction between intentional killing and incidental killing, i.e., killing in which death is a foreseen side effect of the agent's conduct, is significant in the following way: the fact that an agent caused someone's death incidentally can sometimes exonerate from the guilt of murder, whereas intentionally killing an innocent person as

[38] McMahan, *Killing in War*, 173–74.

a means to bringing about good outcomes or avoiding evil ones cannot. Anscombe makes this point in *Mr. Truman's Degree* and adds that killing the innocent as an end in itself also always constitutes murder:

> Choosing to kill the innocent as a means to your ends is always murder. Naturally, killing the innocent as an end in itself is murder too.... I intend my formulation to be taken strictly; each term in it is necessary. For killing the innocent, even if you know as a matter of statistical certainty that the things you do involve it, is not necessarily murder. I mean that if you attack a lot of military targets, such as munitions factories and naval dockyards, as carefully as you can, you will be certain to kill a number of innocent people; but that is not murder. On the other hand, unscrupulousness in considering the possibilities turns it into murder.[39]

The claim that the distinction between intentional and incidental harm has ethical significance is characteristic of the principle of double effect (PDE). The PDE has traditionally been formulated as stating necessary and sufficient conditions for engaging in conduct that has both good and bad effects.[40] In two published essays ("Action, Intention, and 'Double Effect'" and "Murder and the Morality of Euthanasia"), Anscombe proposes a more modest version of the principle, which she calls the "principle of side effects":

> *Principle of Side Effects* (PSE): The prohibition on murder does not cover *all* bringing about of deaths which are not intended.[41]

The PSE presupposes the prohibition on murder and the claim that intentionally killing innocent people is always murder. But while you must not aim at the death of an innocent person, "causing it does not necessarily incur guilt."[42] The principle is modest in two respects. First, it does not cover the bringing about of bad effects in general. It is specifically about death and the question of when causing death constitutes murder. Second, it does not attempt to state necessary and sufficient conditions for the permissibility of causing incidental death. It simply says that incidental killings are not always murder.

[39] Anscombe, *Mr. Truman's Degree*, 66.

[40] See, e.g., Gury, *Compendium Theologiae Moralis*; Mangan, "An Historical Analysis of the Principle of Double Effect"; and Cavanaugh, *Double-Effect Reasoning*.

[41] Anscombe, "Action, Intention, and 'Double Effect,'" 220; cf. "Murder and the Morality of Euthanasia," 274.

[42] Anscombe, "Action, Intention, and 'Double Effect,'" 220, and "Murder and the Morality of Euthanasia," 274.

Why accept the claim that the intentional killing of the innocent is always murder? Anscombe's answer is that this is central to our common understanding of what murder is*:*

> The central concept [of murder] … is that of intentional killing of the innocent. This gives us our sharpest and most full-blown picture of the murderer par excellence. He is willing to kill those who have done him no wrong. His hand is ready to shed innocent blood. Everywhere where such actions on the part of murderous rulers, soldiers, terrorists or other armed men are reported, these phrases occur: "killing innocent people," "compassing the death of innocent bystanders," "slaughtering a crowd of innocent and helpless victims," and so on.[43]

> First: murder is for example deliberately killing innocent people. This is the most widely agreed conception. Powerful men who are known to have done this or had it done, whether *ad terrorem* or exulting in their abilility [*sic*] to prove that others were at their mercy, or delighting in cruelty towards any who might conceivably have less than total subservience in their hearts—such men are unhesitatingly regarded as murderers.… *The* thing that fuses the matter is: They have innocent people killed.[44] "Their feet are swift to shed innocent blood."[45] This is our first picture of the murderous among mankind.[46]

The intentional killing of the innocent thus constitutes the "hard core" of murder.[47] Hard core cases involve the central meaning of "malice" as it occurs in the understanding of murder as killing with "malice aforethought." Here "malice" does not connote spiteful feeling but rather the badness of the agent's intent.[48]

The hard core of murder forms a relatively well-defined area. The main place there is apt to be controversy is the question of who counts as "innocent" in war. On Anscombe's view, someone is innocent in war when he is not *nocentes* (not harming or not offending), and people are *nocentes* when they are engaged in an objectively unjust proceeding, such as an unjust attack.[49]

[43] CIAA 8.291.12–13.

[44] The word I have written as "fuses" is difficult to discern in Anscombe's handwriting.

[45] This appears to be a reference to Isaiah 59:7.

[46] CIAA 9.313.R7–R8; italicized words underlined in the ms.

[47] Anscombe, "Murder and the Morality of Euthanasia," 262, and "Action, Intention, and 'Double Effect,'" 219.

[48] CIAA 8.290.23.

[49] Anscombe, "War and Murder," 53; cf. *Mr. Truman's Degree*, 67; *Justice of the Present War Examined*, 77–78.

The PSE states that the prohibition on murder does not cover all bringing about of deaths that are not intended. The rationale for this is that there are both (a) cases in which killing is not intended and yet clearly *are* cases of murder and (b) cases in which killing is not intended and clearly are *not* cases of murder. Examples of the first sort of case include a situation where a man burns down a house not with the intention of killing anyone inside (perhaps he just wants to collect the insurance) but without caring whether anyone is there, and someone is killed; and the case of Euthyphro's father, who neglected to feed and shelter a field laborer tied up in his custody, and who watched with indifference as the man died of exposure.[50] Cases of incidental killing such as these form a "penumbra" that surrounds the hard core of murder.[51] In them, the agent displays a callous disregard for human life that is equally or even more heinous than some cases of intentional killing. Since murder is distinguished from lesser forms of culpable homicide on the basis of its heinousness, it would be unreasonable not to regard these as cases of murder.[52]

On the other hand, there are also cases in which the agent brings about death as a side effect, which are not cases of murder. A scenario that appears in many of Anscombe's unpublished writings is a variation of the *Smith* case.[53] Smith was a petty thief who had stolen property in his car. When he was stopped for questioning, he sped off, and a police officer jumped on the car. Smith drove a zigzag course, and the officer fell into oncoming traffic and died. However, we can change the details of the case so that Smith is a hero who is driving out of town with a bomb that is about to go off. There is no time to explain what is happening, so he speeds away from a traffic stop, and as before, a police officer jumps onto his car. Since the officer is obscuring Smith's view, he drives a zigzag course to shake him off, foreseeing that there is a risk the officer will be killed by oncoming traffic, and the risk is realized. When the facts of the case become known, no one would bring a charge of murder against Hero Smith. He is accountable for bringing about the death of the police officer, but he has a legitimate exonerating response.[54]

[50] CIAA W5.512.3–4. For the case of Euthyphro's father see Plato, *Euthyphro*, 4c–d.

[51] CIAA W5.512.3–4; Anscombe, "Murder and the Morality of Euthanasia," 274, and "Action, Intention, and 'Double Effect'," 220.

[52] Anscombe, "Murder and the Morality of Euthanasia," 263, and CIAA 9.313.R9. We might say that penumbral cases also involve a sort of malice in the will of the murderous agent, which consists in indifference or disregard for human life.

[53] E.g., CIAA 8.291; 8.292; W3.303; W5.512. See DPP v. Smith [1961] AC 290.

[54] Less dramatic examples are cases of surgeons who perform dangerous surgeries as carefully as they can but lose the patient and some cases of closing doors to contain flooding or fire

In between these clear cases there is a gray zone that consists of borderline cases: the penumbra is fuzzy and its edges are blurred.[55] These cases will often be disputable, and their classification will depend on things such as assessments of risk and the balancing of the goods and evil involved.[56] The resolution of such cases belongs to casuistry, and while casuistry "may allow you to stretch a point on the circumference, it will not permit you to destroy the center."[57] In the case of murder, the "center" is the hard core, which consists of the intentional killing of the innocent.

Anscombe's way of arguing for the PSE differs from the way that some other philosophers argue for the principle of double effect. Consider the way that Philippa Foot motivates the PDE in her essay "The Problem of Abortion and the Doctrine of the Double Effect." Foot proceeds by constructing hypothetical cases and consulting her intuitions about whether the agent's conduct is *morally right* or *morally wrong*, or whether the agent *should* or *should not* act as he does. For instance, she thinks it would be outrageous for a judge to order the execution of an innocent man in order to prevent rioters from killing five hostages, but we would say that a driver of an out-of-control trolley should steer it away from five trapped workmen onto a sidetrack where it will kill one.[58] Double effect is then brought in as a way of explaining these intuitions about right and wrong.

Such a method provides, at best, little support for the PDE. First, it is possible that other principles explain our intuitions about the target cases equally well.[59] Second, there are cases in which many people's intuitions about moral permissibility conflict with the PDE. Anscombe raises the following pair of cases, which serve as an illustration:

> *Explode*: A potholer is stuck in the entrance of a cave with people behind him. The water level is rising, and the people will soon be drowned. They can escape by blowing up the stuck potholer with a stick of dynamite.

(Anscombe, "Murder and the Morality of Euthanasia," 275, and "Action, Intention, and 'Double Effect,'" 220).

[55] Anscombe, "Murder and the Morality of Euthanasia," 274, and "Action, Intention, and 'Double Effect,'" 219.

[56] Anscombe, "Murder and the Morality of Euthanasia," 277.

[57] Anscombe, "Modern Moral Philosophy," 36.

[58] Foot, "The Problem of Abortion and the Doctrine of the Double Effect," 23.

[59] In "The Problem of Abortion and the Doctrine of the Double Effect," Foot ultimately rejects the PDE in favor of a principle that distinguishes between the strictness of negative duties and positive duties (27–29).

> *Rock*: Similar to Explode, but the people can escape by opening up another exit. This will require them to move a large rock, which will roll along a path and crush the head of the potholer.[60]

In Explode the stuck potholer will be killed intentionally as a means to clearing the cave entrance, whereas in Rock his death will be a foreseen side effect of moving the rock. Anscombe predicts that many moral philosophers will "pour scorn" on double effect here, finding the intentional/incidental distinction intuitively "morally non-significant" in these cases.[61]

Anscombe's argument for the PSE is not the same as Foot's method, however. Indeed, Anscombe is critical of this method, which she says leaves us "helplessly swivelling our attention back and forth between a *situation* and the concept 'right.'"[62] What is missing from Foot's method is the presence of "middle terms" whose function is to intervene between a situation and our application of terms like "right" and "wrong." These middle terms are thick ethical concepts such as "courage," "hypocrisy," "temperance," and "truthfulness."[63] Anscombe's argument for the PSE revolves around just such a middle term—namely, murder. While her argument does utilize intuitions about cases, the intuitions are about the applicability of the concept "murder" to a situation. The argument, again, is that the intentional killing of the innocent constitutes the paradigm of murder. Additionally, there are some incidental killings that we would readily agree are cases of murder and others (such as the case of Hero Smith) which no one would classify as murder.

Anscombe's argument for the PSE makes available a number of responses to the Explode/Rock pair. First, the PSE does not imply that the cases are morally different. All it implies is that blowing up the potholer in Explode is forbidden, as it is a hard-core case of murder. Second, given that the penumbra is fuzzy, it is inevitable that there will be borderline cases where it is disputable whether they constitute murder. Rock is plausibly just such a case. Third, once our focus is on murder, there are some grounds for distinguishing Rock and Explode. In particular, people who are willing to move the rock but who would not choose the potholer's death as a means of escaping "shew themselves as people who will absolutely reject any policy of making the death of innocent people a means or end."[64] That stance is far from meaningless, for it shows they are unwilling to

[60] Anscombe, "Murder and the Morality of Euthanasia," 275–76, "Action, Intention, and 'Double Effect,'" 221–24, and CIAA 8.291.25–27.

[61] CIAA 8.291.30.

[62] CIAA 13.511.10 and 8.298.W14; italicized word underlined in the ms.

[63] CIAA 13.511.10 and 8.298.W14; cf. Anscombe, "Modern Moral Philosophy," 33.

[64] Anscombe, "Murder and the Morality of Euthanasia," 276.

engage in activity that is most paradigmatic of murder. Finally, upon reflection, we may conclude that Rock is nonetheless a case of murder. If we do, then it is incumbent on us to find an additional principle that explains why it is.[65]

Let me summarize the results of this section. On Anscombe's account, the hard core of murder consists in the intentional killing of the innocent. This core is surrounded by a penumbra, which includes some but not all cases in which death is not intended. The penumbra is fuzzy, which means that there are borderline cases. When killing is incidental, the question of whether it belongs in the penumbra will be a matter of whether the agent possesses an exonerating response.[66] And a common type of exonerating response will refer to the necessity of the agent's conduct for securing some great good or avoiding some great misfortune, as in the scenario of Hero Smith. However, this sort of exoneration is not available when the killing is in the hard core. These killings always count as murder.

We also have a response to the second challenge from section 1. That challenge observed that there are cases in which we first need to decide whether killings are justified by factors such as proportionality and necessity before determining whether they constitute murder; and it took this as evidence that murder should be defined as unjustified killing. While the premise is true for cases in which killing is not intentional, it does not follow that the business of calling something murder *always* waits upon the question of whether it is justified. In particular, it does not do so when what is in question is intentionally killing innocent human beings.

4. Titles to Kill

A person commits murder when she bears level three responsibility for the death of another, and she will bear this responsibility when she is callable to account for that death and lacks an exonerating response. Anscombe's view is that intentionally killing the innocent always constitutes murder. However, there exist other cases in which killing is intentional but plausibly are not cases

[65] Anscombe proposes the following principle: the "intrinsic certainty of the death of the victim, or its great likelihood from the nature of the case" would make it murderous to move the rock ("Murder and the Morality of Euthanasia," 276; "Action, Intention, and 'Double Effect,'" 225). However, this principle also seems to classify as murder actions that the PDE has traditionally been used to support. Is it not very likely "from the nature of the case" that when military targets are bombed, nearby civilians will be killed by the explosions? I am more amendable to the conclusion that moving the rock would not be murderous.

[66] Or an excuse mitigating the degree of blame, which means the case is better classified as manslaughter (see note 28 above).

of murder. Many people accept that it is not always murderous, for example, for law enforcement officers to fight lawbreakers who resist arrest, even to the point of death. Just war theory holds that combatants fighting in a just war do not necessarily murder enemy combatants when they kill them. More controversially, some people believe that capital punishment is sometimes legitimate. It seems, then, that in these areas there are cases in which the agent has an exonerating response to the charge of murder. In order to explain the validity of such responses, Anscombe introduces the notion of a title to kill.

4.1. The Concept of a Title to Kill

A title to kill is an entitlement or authority to kill some person or persons intentionally. An example of a title to kill that Anscombe discusses in multiple archival documents is tyrannicide.[67] Assuming there is such a title, for a killing to be done in the exercise of it, the killing must be intentional under the description "killing a tyrant." It is not a tyrannicide if someone who happens to be a tyrant is killed not *because* he is a tyrant but for some other reason, such as to avenge a jealous passion.[68] This may seem to suggest that the distinguishing mark of a title to kill is that it is done with a public purpose—that is, a purpose that has to do with promoting the common good of a political community.

There is no reason in principle, however, why there could not be titles to do "private" killing. For example, the ancient Romans claimed that as part of the *patria potestas,* a father had a title to kill any of his children. This is not an actual title to kill, but it cannot be ruled out on conceptual grounds.[69]

The core of a title to kill is instead that the essential identification of the act of killing it involves is independent of the advantages that can be expected from it. Rather, the act of killing is identified in terms of the nature of the victim and the relation he stands in to the one holding the title.[70] For instance, in tyrannicide, the key term is "killing a tyrant," and this does not include as part of its meaning a reference to expected goods to be produced or evils avoided. However, even though the type does not include an essential reference to good or bad consequences, it may nonetheless be that a necessary condition for any token of the type to be justified is that it is not expected to lead to certain bad outcomes. A concrete instance of tyrannicide may be wrong, for instance, if it is foreseeable that killing *this* tyrant *here and now* will precipitate a civil war or lead to an even more oppressive regime. Nonetheless, if tyrannicide is a true title to

[67] CIAA 8.292; 8.293; 8.296; 9.301; W5.515.

[68] CIAA 8.293.

[69] CIAA 8.293.

[70] CIAA 8.293.

kill, then when someone exercises this title the *principal point* of an exonerating response to a charge of murder will be that the person he killed was a tyrant.[71]

Cases that involve titles to kill can be contrasted with one in which it is being debated whether it would be justified for a person to kill her aged uncle, where his death sooner rather than later would avert financial woes. In this case, we are not debating the question, "Can avunculicide be justified?" For the man's being the agent's uncle is not the principal thing that would supposedly justify the proposed killing. That is rather that killing the man will contribute to averting financial misfortune.[72]

4.2. Titles to Kill and Civil Authority

The example of the *patria potestas* shows that not every alleged title to kill is an actual title. For any purported title to kill, there must be some rationale that establishes its validity. Anscombe notes that most of the commonly assumed titles to kill are titles that derive from public authority, or as she calls it, civil authority.[73] The killing done by soldiers, the killing of domestic lawbreakers who resist arrest, and the execution of criminals are all killings done by a commission from civil authority. An investigation into the nature of murder must, therefore, include an inquiry into whether these are true titles.

Anscombe defends the proposition that civil authority is a source of titles to kill non-innocent persons in "On the Source of the Authority of the State" and several archival documents. What distinguishes civil authority from large-scale voluntary cooperative associations is that civil authority demands obedience, and its demand is backed by threats of coercive force. What grounds the entitlement of civil authority to use violence is that doing so is necessary for it to perform its task, where this task is, in turn, necessary for human good—namely, the promotion of a peaceful normality where people can live together "in multitudes."[74]

Anscombe distinguishes two functions of government in securing the condition of civic peace. The first is to protect people from unjust attacks on their lives and persons. The second is to prohibit violent private revenge on the part of people who have been wronged by their fellow citizens. The latter promotes peace by forestalling the killing of the innocent that would inevitably occur if each person were allowed to be judge in his own case and use violence to right perceived wrongs against him. These two functions mean that government

[71] CIAA 8.296.15.

[72] CIAA 8.293.

[73] CIAA 8.292.

[74] Anscombe, "On the Source of the Authority of the State," 135–37, and CIAA 8.291.32–33.

prohibits all private right to use force except in immediate self-defense.[75] The task of protecting innocent people from unjust attack, both from internal disturbers of the peace and from external enemies, is what gives rise to institutions such as police, courts, and the military. And it is the use of violence in the administration of justice that distinguishes civil authority from rule by a highly organized and smoothly functioning Mafia.[76]

The form that civil authority ordinarily assumes is the institution of government. Anscombe argues that this cannot be the sole form that it takes, however. The problem, which she highlights in an unpublished paper on assassination, is that if it were, then an invading power could extinguish all right to resistance simply by destroying a nation's government, for that would turn into murder any killing done by resistance fighters.[77] In extraordinary cases where there is a lack of legitimate *de facto* civil authority, due either to usurpation by foreign enemies or to internal ones who overthrow the government, there must be such a thing as the *self-assumption* of civil authority. What this means is that people fighting for a just cause in a rebellion can kill a tyrant or usurper and those fighting on his or her behalf by constituting themselves as soldiers on behalf of such civil authority as there *ought* to be.[78] The title of such soldiers to kill thus rests on authority that is normative and forward-looking.

The focus of this section has been titles to kill deriving from civil authority. But there is also a question of whether there are any private titles to kill—that is, titles that do not derive from authorization or commission from civil authority. It is widely assumed that there is at least one such title—namely, the title of a private person to kill intentionally an unjust attacker in defense of himself or others. However, in both her published writings and archival papers, Anscombe rejects the notion that private individuals possess a title to kill in self- or other-defense. Following Aquinas's treatment of self-defense, she claims that the right to private defense is not a title *to kill* but only a title to use such violence as is necessary to stop an immediate attack.[79] Many jurisdictions do allow self-defense as a justifying defense to murder when the attacker is killed intentionally (assuming that conditions of imminence, proportionality, and necessity are met), but Anscombe claims that in conscience, one's justification

[75] CIAA W5.515.3.

[76] Anscombe, "On the Source of the Authority of the State," 136.

[77] CIAA W5.515.5.

[78] CIAA W5.515.7; 8.292.

[79] Anscombe, "War and Murder," 53, *Mr. Truman's Degree*, 68, and CIAA 8.292. For Aquinas's account of the ethics of self-defense, see Aquinas, *Summa Theologiae* II-II 64.7.

for killing an attacker should be that his death was not intended but was a side effect of adopting the means to stopping his attack.[80]

Even more contentious, however, is the question of whether there exist other private titles to kill. Debates about the morality of abortion and euthanasia, for example, are in part debates about such titles. There is no space for analysis of these debates here. It is a virtue of Anscombe's account of murder that it allows us to pinpoint why these practices are controversial: they involve the intentional killing of innocent human beings, and this appears to put them within the hard core of murder. Because of this, proponents of these practices often proceed by arguing that they have features that distinguish them ethically from other cases of intentional killing. For example, familiar arguments contend that a fetus is not a person and so does not possess the same moral protections as more mature human beings. And a characteristic part of arguments for euthanasia is the claim that in cases of interminable suffering, death is not an evil. Without entering these debates, I note that Anscombe rejects these arguments.[81]

The upshot is that Anscombe holds that all actual titles to kill derive from civil authority, and these are titles to kill only persons who are not innocent.[82] In functioning as exonerating responses, these titles help set the boundaries of what constitutes murder. If an agent intentionally kills a non-innocent person without possessing a title to kill, he will not have an exonerating response to a charge of murder. Cases of this sort do not fall within the hard core of murder, which is the intentional killing of the innocent, but neither do they fall within the penumbra, which is composed of cases where the killing is not intentional. They form a distinct area, which we might conceive as an outlying region that surrounds the core.

The notion of a title to kill is the key to resolving the third challenge from section 1. That challenge observed that in debates about the morality of killing in war, capital punishment, and other areas, the judgment whether certain types of killing are justified precedes the question whether they are murder.

[80] Anscombe, "War and Murder," 54. Anscombe also asserts that a person existing in a state of nature is not a private individual; rather, the public/private distinction is not applicable in this context ("War and Murder," 54). In one place she allows that in a state of nature, individuals may intentionally kill unjust attackers in self-defense (CIAA W5.515.3).

[81] For Anscombe's rejection of the claim that a fetus is not a person, see "Murder and the Morality of Euthanasia," 267–68; for her objection to euthanasia, see "Murder and the Morality of Euthanasia," 269. For an analysis and evaluation of Anscombe's argument that euthanasia constitutes a form of murder, see Jones, "Anscombe on Euthanasia as Murder."

[82] With the possible exception of killing unjust attackers in a state of nature, which would not involve an exercise of civil authority. See note 80 above.

Anscombe contends that what is up for debate in these cases is whether a certain type of killing is the basis of a title to kill. If we think a title to kill exists, then we will think it can form the principal part of an exonerating response to a charge of murder. Therefore, when it comes to titles to kill, the question of justification does precede the judgment of whether the corresponding type of killing is (always) murder. Nonetheless, since the types of killing that figure in titles to kill are not defined in terms of their expected consequences, it remains the case that the possible justification "killing a person as a means to producing a good outcome or avoiding a bad one" does not constitute a title to kill. Hence, Anscombe maintains that if someone attempts to justify killing someone simply on the ground that doing so is a means to producing good consequences, then what she is attempting to justify is murder.

5. Conclusion

Anscombe's interest in the topic of murder was motivated by her belief that murder is absolutely prohibited and, hence, always to be condemned. But logically prior to the question of whether murder is always forbidden is the question of what constitutes murder. In this paper, I have integrated archival materials with Anscombe's published writings to reconstruct her answer to this question. Anscombe was keen to deny the semantic thesis that "murder" means "unjustified or impermissible killing," which, if true, would trivialize the debate between her and the mid-twentieth-century Oxford moral philosophers. Indeed, she goes so far as to call the semantic thesis a "thought-stopping device."[83] The sense in which this is so becomes apparent when we examine her complex account of murder, which involves *inter alia* discussions of responsibility and voluntary agency, the distinctions between actions and omissions and intention and foresight, and the concept of a title to kill and the basis of civil authority.

Anscombe contended that the "Hebrew-Christian ethic," which upholds an absolute prohibition on murder, acknowledges the inherent worth or dignity of human beings in a way that the systems of Oxford moral philosophy do not.[84] In one unpublished paper, she identifies two ways in which this is so.[85] First, the ethic holds that having done nothing to deserve it, a human being is never to be unjustly done away with for the sake of others. Except for justice's sake,

[83] CIAA 13.511.6; 8.298.W12.

[84] Anscombe, "Murder and the Morality of Euthanasia," 267–68. For analysis of Anscombe's remarks about the nature and basis of human dignity see Müller, "The Spiritual Nature of Man"; and Lott, "The Knowledge of Human Dignity."

[85] CIAA 8.289.

no one is to be deliberately killed. Second, it also testifies to the worth of the human being *qua* acting subject: "Others cannot be defiled by his abstaining from evil doing; they may suffer or die because of it, but they cannot be defiled, because each man can be defiled inwardly only by what he does himself."[86] The argument is brief, but I think the central idea can be elaborated as follows. Moral systems that reject an absolute prohibition on murder will sometimes approve, and perhaps even require, doing things that count as murder, such as killing innocent people as a means to avoiding sufficiently bad outcomes. These theories will therefore approve, and even require, people to incur the guilt of murder. But if a person incurs this guilt, then he degrades or defiles himself: he has made himself into a murderer, and his life is tainted by having incurred that guilt. By contrast, the Hebrew-Christian ethic refuses to approve of a person's defiling himself in this way. By including an absolute prohibition on murder, it therefore acknowledges the value of human beings both insofar as they are patients whose dignity is always to be respected and insofar as they are agents who are thereby protected from having to defile themselves.[87]

University of St. Thomas
joshua.stuchlik@gmail.com

REFERENCES

Anscombe, G. E. M. "Action, Intention, and 'Double Effect.'" In *Human Life, Action and Ethics*, 207–26.

———. *Ethics, Religion, and Politics*. Vol. 3 of *The Collected Philosophical Papers of G. E. M. Anscombe*. Oxford: Basil Blackwell, 1981.

———. *Human Life, Action and Ethics: Essays by G. E. M. Anscombe*, edited by Mary Geach and Luke Gormally. Exeter: Imprint Academic, 2005.

———. *The Justice of the Present War Examined*. In *Ethics, Religion, and Politics*, 72–81.

———. "Modern Moral Philosophy." In *Ethics, Religion, and Politics*, 26–42.

———. *Mr. Truman's Degree*. In *Ethics, Religion, and Politics*, 62–71.

[86] CIAA 8.289.

[87] My work on this paper was supported by a fellowship at the Stockdale Center for Ethical Leadership at the United States Naval Academy. Thanks to Daniel Cheely at the Collegium Institute and Sarah Heim at the Kislak Center for Special Collections, Rare Books and Manuscripts for their assistance in accessing the Anscombe Archive. Thanks is also due to audiences at the University of Pennsylvania and the Naval Academy, where I presented material from a draft of this paper.

———. "Murder and the Morality of Euthanasia." In *Human Life, Action and Ethics*, 261–78.

———. "On the Source of the Authority of the State." In *Ethics, Religion, and Politics*, 130–55.

———. "Prolegomenon to the Pursuit of the Definition of Murder: The Illegal and the Unlawful." In *Human Life, Action and Ethics*, 253–60.

———. "The Two Kinds of Error in Action." In *Ethics, Religion, and Politics*, 3–9.

———. "War and Murder." In *Ethics, Religion, and Politics*, 51–61.

Aquinas, Thomas. *Summa Theologica*. Translated by the Fathers of English Dominican Province. 5 vols. New York: Benziger Brothers, 1947.

Aristotle. *Nicomachean Ethics*. Translated by W. D. Ross. In *The Basic Works of Aristotle*, edited by Richard McKeon, 935–1126. New York: Random House, 1941.

Berkman, John. "Justice and Murder: The Backstory to Anscombe's 'Modern Moral Philosophy.'" In Teichmann, *The Oxford Handbook of Elizabeth Anscombe*, 225–70.

Cavanaugh, T. A. *Double-Effect Reasoning: Doing Good and Avoiding Evil*. Oxford: Oxford University Press, 2006.

Foot, Philippa. "The Problem of Abortion and the Doctrine of the Double Effect." In *Virtues and Vices and Other Essays in Moral Philosophy*, 19–32. Oxford: Oxford University Press, 2002.

Frey, Jennifer A. "Revisiting Modern Moral Philosophy." *Royal Institute of Philosophy Supplements* 87 (July 2020): 61–83.

Geach, Mary. "Introduction." In Anscombe, *Human Life, Action and Ethics*, xiii–xxi.

Gormally, Luke. "On Killing Human Beings." In Gormally, Jones, and Teichmann, *The Moral Philosophy of Elizabeth Anscombe*, 133–53.

Gormally, Luke, David Albert Jones, and Roger Teichmann, eds. *The Moral Philosophy of Elizabeth Anscombe*. Exeter: Imprint Academic, 2016.

Gury, Jean-Pierre. *Compendium Theologiae Moralis*. Regensburg: Georgii Josephi Manz, 1874.

Jones, David Albert. "Anscombe on Euthanasia as Murder." In Teichmann, *The Oxford Handbook of Elizabeth Anscombe*, 271–91.

Lott, Micah. "The Knowledge of Human Dignity." In Teichmann, *The Oxford Handbook of Elizabeth Anscombe*, 292–307.

Mangan, Joseph. "An Historical Analysis of the Principle of Double Effect." *Theological Studies* 10, no. 1 (February 1949): 41–61.

Markl, Dominik. "The Decalogue: An Icon of Ethical Discourse." In *The Cambridge Companion to the Hebrew Bible and Ethics*, edited by C. L. Crouch, 9–22. Cambridge: Cambridge University Press, 2021.

McMahan, Jeff. *Killing in War*. Oxford: Oxford University Press, 2009.

Moore, G. E. *Principia Ethica*. London: Cambridge University Press, 1903.

Müller, Anselm Winfried. "The Spiritual Nature of Man." In Gormally, Jones, and Teichmann, *The Moral Philosophy of Elizabeth Anscombe*, 1–32.

Plato. *Euthyphro*. Translated by Lane Cooper. In *The Collected Dialogues of Plato*, edited by Edith Hamilton and Huntington Cairns, 169–85. Princeton: Princeton University Press, 1999.

Teichmann, Roger, ed. *The Oxford Handbook of Elizabeth Anscombe*. Oxford: Oxford University Press, 2022.