JOURNAL *of* ETHICS & SOCIAL PHILOSOPHY

VOLUME XXIII · NUMBER 3
January 2023

ARTICLES

SYMPOSIUM

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

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

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

Journal of Ethics and Social Philosophy
VOL. 23, NO. 3 · JANUARY 2023
https://doi.org/10.26556/jesp.v23i3.1541

THE INHERENT TOLERANCE OF THE DEMOCRATIC POLITICAL PROCESS

Emanuela Ceva and Rossella De Bernardi

TOLERATION is one of the most debated ideals across liberal political theories of democracy. While such prominent liberal theorists as John Rawls celebrate the fundamental role of toleration in the design of well-ordered liberal democracies, critiques of the value of toleration date back to Immanuel Kant's denunciation of this notion as the "arrogant" posture of the powerful, granting the powerless concessions at their discretion.[1] Ultimately, on whether toleration should be abandoned or rescued among liberal democratic core commitments, the jury is still out. This article advances a novel, qualified defense of toleration as a central ideal of a liberal democratic interactive political morality.

To be sure, defenses of toleration as an ideal for contemporary liberal democracies have been numerous in the last couple of decades. Many such defenses follow a twofold strategy. At its essence the strategy consists in the departure from the traditional characterization of toleration. This characterization is indicative of interpersonal relations of forbearance distinguished by an element of disapproval among the participants in those relations. This departure comes in two steps. The basic step is a removal of the emphasis on forbearance. This step presents a normative account of toleration as a general practice of noninterference proper of neutralist political arrangements aimed to protect individual freedom.[2] The most recent among such defenses make a further step by offering a conceptual overhaul of toleration. For example, such defenses redescribe toleration as a positive form of recognition or indifference.[3] They thus reconceptualize toleration, reinterpreting the reference to disapproval. This twofold strategy is the main critical target of this article.

The twofold strategy is partly motivated by an attempt to resist some concerns about the complex relationship of toleration with multiple features of

1 Rawls, *Political Liberalism*, 43; Kant, *An Answer to the Question*, 12.

2 Jones, "Making Sense of Political Toleration"; Balint, *Respecting Toleration*.

3 See, respectively, Galeotti, *Toleration as Recognition*; Balint, *Respecting Toleration*.

contemporary liberal democracies. One concern is that the liberal democratic commitment to protecting individual freedom and respecting pluralism makes toleration redundant.[4] Another concern is about the possible inconsistency of the logic of toleration with that of many other ideals generally thought to sustain liberal democracies. These ideals include neutrality, equality of political power and civic status, and the democratic credentials of the legitimation of state action.[5] Central to these concerns is the thought that the kind of forbearance demanded by toleration is already secured by other fundamental liberal commitments, which also preempts the disapproval implied by the logic of toleration as a ground for political action.

We share the aim of defending the political relevance of toleration that has prompted many recent commentators to adopt the twofold strategy. However, we critically engage with the strategy as we make two main claims. First, the twofold strategy focuses on the realization of toleration in the *political arrangements* (for example, public decisions) produced through political processes. Therefore, it offers a normative account of toleration that underestimates an important "interactive" dimension of what it means for liberal democracies to realize toleration as a property inherent to its constitutive *political processes* (for instance, of decision-making). Second, this interactive dimension of toleration can be defended as central to liberal democratic political morality without requiring the conceptual overhaul of toleration that the twofold strategy proposes.

Our discussion progresses as follows. In section 1, we articulate the twofold strategy, drawing on some prominent views of toleration as an ideal of liberal democratic political morality. We then devote section 2 to discussing how the strategy is too hasty in setting aside the forbearance interpretation of toleration. This hastiness is problematic to the extent that it underplays some important particularities that characterize relations of toleration in the circumstances of deep political disagreement typical of contemporary liberal democracies. Moreover, we show how the twofold strategy relies on a partial view of toleration. This view presents toleration only as an ideal of political morality causally enacted in the freedom-protecting outcomes of political processes. We argue that this partial view fails to do justice to the distinctively relational structure of toleration. We show how to overcome this limitation by focusing also on the properties inherent to the forms of interaction that democratic political processes constitute. In section 3, we vindicate the importance of understanding

4 Heyd, "Is Toleration a Political Virtue?"

5 Balint, *Respecting Toleration*, 32–35; Forst, *Toleration in Conflict*, 518–20; Jones, "Toleration and Neutrality," 97–110; Meckled-Garcia, "Toleration and Neutrality"; Brown, *Regulating Aversion*; Newey, "Is Democratic Toleration a Rubber Duck?"

toleration also as an ideal of interactive political morality. This ideal captures one important aspect of the liberal democratic commitment to establishing a respectful form of interaction between citizens as political agents in circumstances of deep political disagreement. In section 4, we expound the nuanced normative evaluations of the tolerance of a liberal democracy that our account makes possible. In section 5, we conclude by summing up how our argument responds to the concerns of redundancy and inconsistency about the realization of toleration in liberal democracies.

Before we engage in this discussion, take note of two clarifications concerning the contours of our proposal. First, our critical argument remains within the boundaries of the neutralist interpretation of liberalism. In this context, the point of neutrality is to protect individual agency within an institutional framework whose justification does not presuppose the (moral or epistemic) superiority of any particular controversial conception of the good. Thus, we view the democratic polity from the perspective of a justificatory interpretation of liberalism, broadly construed.[6] Second, we discuss the role of toleration within the framework of what it takes to realize some fundamental normative commitments of a liberal democratic political morality in circumstances of deep political disagreement. To borrow Jeremy Waldron's terminology, such a disagreement is one of the main "circumstances of politics," in which the demands of toleration acquire—as the article will show—particular importance.[7] These circumstances of disagreement are actual and, thus, broader and deeper than those indicated by Rawls as "reasonable disagreement."[8]

1. The Twofold Strategy to Defend Toleration as an Ideal of Liberal Democratic Political Morality

A current illustration of the twofold strategy in defense of toleration comes from the joint consideration of Peter Jones's and Peter Balint's prominent discussions. They show how toleration is a significant (nonredundant) idea that belongs to the "furniture" of (and, therefore, is not inconsistent with) a liberal democratic political morality. To this end, Jones and Balint offer an

6 For a general account, see Waldron, "Theoretical Foundations of Liberalism." This characterization covers various understandings of the liberal justificatory project, whether, for example, consensus (Quong, *Liberalism without Perfection*; Rawls, *Political Liberalism*) or convergence driven (Gaus, *The Order of Public Reason*), being compatible with different more or less substantial interpretations of public reason.

7 Waldron, *Law and Disagreement*, 105. See also Newey, "Metaphysics Postponed."

8 Rawls, *Political Liberalism*; Quong, *Liberalism without Perfection*.

interpretation of toleration as a property of freedom-protecting political arrangements.[9] Jones defends a normative account of toleration whose central feature is state protection from intolerance. Balint builds on Jones's work to offer a conceptual overhaul of toleration, which develops a liberal "permissive" view. The latter is of particular interest because it tracks a largely held commonsense understanding of toleration as an instance of indifference. Their works instantiate the twofold strategy because they interpret and defend toleration by departing from the traditional understanding of this idea as indicative of interpersonal relations of forbearance in the face of disapproval.

The traditional or "orthodox" view of toleration falls within the coordinates of three main components: *A* deliberately refrains from acting (non-hindrance component) on their negative judgment of *B*'s beliefs or practices (objection component) despite their being in the (actual or counterfactual) position of doing so (power component).[10] Jones and Balint ask how this orthodox view of toleration may rightfully inform the political arrangements of a democracy grounded in a justificatory neutralist interpretation of the liberal political project while avoiding tensions with the ideals central to that project. Notably, by departing from the orthodox view, they address the concerns that characterizing liberal democracies as "tolerant" risks inconsistency or, at best, redundancy.[11]

The inconsistency and redundancy concerns about toleration stem from the consideration that, once neutral political institutions are in place and citizens' basic rights are protected, the three components of toleration may lose force. It is a defining feature of neutral liberal institutions that certain spheres of individual action—including, for example, religion and matters of conscience—are protected from state interference (within limits standardly associated with some understanding of the harm principle and needs of action coordination). More generally, in a neutralist liberal democracy, those who hold public office simply lack the prerogative personally to decide to use their power (power component) to interfere with individuals' spheres of personal freedom (non-hindrance component) based on their individual negative judgment (objection component) of citizens' life plans or ideas.[12] From this viewpoint, toleration's protective function of individuals' life plans and ideas seems

9 Because this discussion focuses on toleration as an ideal of political *morality*, we leave aside so-called *modus vivendi* theories, which ground tolerant practice in political *prudence*. For a discussion, see, for example, Gray, "Pluralism and Toleration in Contemporary Liberal Philosophy."

10 See Forst, *Toleration in Conflict*, 17–26; Balint, *Respecting Toleration*, 5, 28.

11 For an overview, see Ceva, "Toleration."

12 Forst, *Toleration in Conflict*; Meckled-Garcia, "Toleration and Neutrality?"; Newey, "Is Democratic Toleration a Rubber Duck?"

otherwise catered for in view of more positive ideals—thus making appeals to toleration unnecessary if less than undesirable.[13]

Jones and Balint address such concerns by rethinking toleration, from a conceptual and normative point of view. From the normative point of view, Jones grants that the most politically salient feature of a tolerant polity is its capacity to produce political arrangements that protect people's individual freedom from unjustified external interference. But he adds that this feature cannot be understood by looking at interpersonal relations of self-restraint, especially when these relations involve public officials (*qua* tolerators).[14] For Jones, if toleration were to be conceived as a model for discretionary uses of entrusted political power, reference to this ideal would be clearly inconsistent with liberal democratic political morality and its grounding commitment to neutrality. Differently, Jones argues that the distinctive mark of a tolerant state lies in its being capable of securing people's protection from each other's personal intolerance in society, by enforcing the protection of citizens' rights.[15] Since—according to Jones—"to suffer intolerance is to suffer a loss of freedom," the distinguishing feature of a tolerant state is its freedom-protecting capacity.[16] In this sense, Jones sees toleration not so much as an ideal that characterizes relations of forbearance (between public officials and citizens, or among citizens). Rather, he sees it as a property of certain institutional political arrangements protective of individual negative basic rights.[17]

Balint shares Jones's general strategy and takes it a step further. To carve out some political space for toleration, he proposes an overhaul of the concept that expands the orthodox view, and is (allegedly) more aligned with current common language descriptions of public institutions as "tolerant." Namely, Balint thinks that the non-hindrance and power components, but not the objection component, are necessary to define toleration. According to Balint's "permissive" interpretation, we have a maximally tolerant polity when people are maximally free to "live their lives as they see fit," regardless (not only in

13 Jones, "Making Sense of Political Toleration," 385–86. Note that we do not press, here, on whether Jones and Balint in fact succeed in rejecting the redundancy challenge. Our interest in their views is mainly illustrative of the twofold strategy.

14 Jones, "Making Sense of Political Toleration," 389.

15 In Jones's words, "rules and institutions can be adjudged tolerant because and insofar as … they secure an order of things in which people can live their lives as they see fit, unprevented by disapproving others who might otherwise impede them" ("Making Sense of Political Toleration," 387).

16 Jones, "Making Sense of Political Toleration," 398.

17 Jones, "Legalising Toleration," 266.

spite) of others' objection to their commitments.[18] In this permissive sense, a tolerant polity is primarily characterized by indifference.

While Jones's and Balint's theories differ in ways we cannot further expound upon, they overlap in a way that makes them relevant to our critical discussion. For both, to assess whether a state is tolerant, one must look at the properties of the political arrangements (for example, the content of collective decisions or state policies) that the political process generates, and see to what degree such arrangements protect personal negative freedom. We acknowledge that Jones's and Balint's freedom-based characterization might capture one sensible aspect of the function of toleration within the liberal democratic political project. However, in what follows, we argue that this characterization fails to do justice to the full story of how and why toleration matters as an ideal of political morality in liberal democracies.

2. END STATES, INTERACTIONS, AND THE RELATIONAL STRUCTURE OF TOLERATION

Bluntly put, the structure of toleration is relational at its essence. The orthodox idea of forbearance tolerance illustrates this feature by connoting a relation between an *A* who forsakes their (actual or counterfactual) power to interfere negatively with an objected *B*. The twofold strategy of reinterpretation of toleration sketched in the earlier section denies that toleration characteristically indicates interpersonal relations of forbearance distinguished by an element of disapproval between political agents. As seen, the strategy reinterprets the core of toleration as consisting in a commitment to protecting individual freedoms from unjustified external interference. A conceptual overhaul of toleration follows, involving the removal of the objection component from the definition.[19] Thus reinterpreted, relations of toleration would occur anytime A_1 does not interfere with B_1, irrespective of whether A_1 disapproves of B_1 or is either indifferent to or appreciative of B_1.[20]

We suggest that this rescue strategy of toleration is not fully successful because it rests on a reductive set of assumptions about the core features of the liberal democratic political project and of toleration within it. To be sure, the claim that the commitment to protecting individual negative freedom is a basic aspect of the liberal democratic political project is sensible; so is the view of toleration as a property of political arrangements that contribute to

18 Balint, *Respecting Toleration*, 28–32.

19 Balint, *Respecting Toleration*, 13.

20 Balint, *Respecting Toleration*, 5.

realizing this aspect of the project. The focus on freedom is indeed one aspect of this project, but hardly its whole point. The reinterpretation of toleration that the twofold strategy offers is too hasty because it is implicitly informed by a partial picture of the normative grounds of a liberal democracy. In this picture, the core business of a liberal democracy is fully identified with (1) protecting citizens' individual negative freedom by (2) securing political arrangements that protect citizens from (unjustified) external interference. We find both components of this identification unwarranted.

Following a well-established strand of justificatory liberalism, one should not forget that the basic set of political ideals for a liberal democracy—also and prominently—includes such other ideals as respect.[21] Borrowing from Stephen Darwall's typology of moral attitudes, the political realization of respect is best understood in the terms of "recognition respect."[22] To respect someone in this sense means to reckon with their moral status when we set the terms of our relation to them.[23] Fundamentally, in a standard liberal version, the ground of this moral status is someone's capacity for agency—a bundle of capacities including that to author, choose, and pursue a worthwhile life plan.[24] To respect someone in this sense means to recognize them as persons, as an authority not only on their own life, but also on the life of other persons; any person is called to see any other as a constraint on what they may or may not do when any one person is involved. Recognition respect thus characterizes interpersonal relations of reciprocity.

Interestingly for our discussion, the recognition of this status can be *claimed* by any agent against any other. It is not a mere tribute that agents receive.[25] By entering respectful relations, agents bestow upon each other a special kind of authority that enables them to demand appropriate treatment as persons. Such treatment is commonly taken to require the recognition that persons may not be subjected to arbitrary coercion; they are, rather, entitled to a justification for how we treat them.[26] This idea captures the core of many prominent justificatory accounts of the normative grounds of liberal democracies, and

21 See, for instance, prominent proposals in Larmore, "Political Liberalism"; Waldron, "Theoretical Foundations of Liberalism."

22 Darwall, "Two Kinds of Respect" and *The Second-Person Standpoint*.

23 Darwall, "Two Kinds of Respect."

24 See, for example, Rawls's characterization of the moral agent as possessing the moral powers of a sense of justice and forming, pursuing, and revising a conception of the good (*Political Liberalism*).

25 Darwall, *The Second-Person Standpoint*, ch. 3.

26 See, for example, Bird, "Mutual Respect and Neutral Justification"; Forst, *The Right to Justification*.

encompasses but goes beyond the commitment to protecting individual negative freedoms.[27] Even more importantly for our present purposes, persons' moral agency, which demands mutual respect, is often presented as a liberal normative ground of the authority of the democratic political process. This normative liberal characterization of democracy is prominently present, for example, in many noninstrumental accounts of democracy's value, which insist on the democratic process being rightly responsive to people's status as equally authoritative makers of collectively binding decisions.[28]

Once we recall the centrality of this commitment to recognition respect within the liberal democratic political project, it is easier to grasp the reductivity of the twofold strategy. This strategy is fit for rescuing toleration only in a very narrow sense: it valorizes toleration only insofar as it *causally contributes* to the realization of *one* aspect of the liberal democratic project, the protection of individual negative freedom, by securing political arrangements that shelter citizens from unjustified external interference (from the state and their fellows). From this perspective, toleration is an ideal that belongs to an end-state political morality.

To focus on end-state political morality means to analyze and assess political processes by looking at the features of the political arrangements (or end states) those processes produce. From this standpoint, one looks at whether political processes lead to certain morally worthwhile distributions of goods, resources, opportunities, or powers among citizens. As an ideal of end-state political morality, toleration is the property of political arrangements (or end states) that contribute to maximal distributions of individual freedoms, by protecting citizens from unjustified external interference with their life plans. As such, toleration is paradigmatically realized when constitutional provisions or legislative decisions lead to permissive outcomes whereby citizens' freedoms—for example, to spread their ideas, associate with like-minded fellows, or abide by their religious commitments—are protected from unjustified third parties' restrictive interventions.

However, once recalled how the commitment to protecting negative freedom is only one aspect of the liberal democratic political project, we can start to question the sole adoption of this end-state perspective to theorize about the place of toleration within that project. This questioning is important to grasp the whole difference it makes for citizens, in the circumstances of politics, to have their dealings regulated within the boundaries of liberal democratic

27 See, for example, Waldron's account of how the liberal public order is defined by its being "justified to any last individual" ("Theoretical Foundations of Liberalism," 128).

28 Christiano, *The Constitution of Equality*; Kolodny, "Rule over None II."

political processes. The way in which the twofold strategy analyzes and assesses those processes underestimates the complexity of what establishing liberal democratic institutions means and requires in circumstances of deep political disagreement and the role that toleration may have in that context. To appreciate this complexity, we suggest, the discussion of the components of a liberal democratic political morality must also integrate an interactive aspect.[29]

To focus on the interactive aspect of political morality means to analyze and assess the political processes of a liberal democracy by looking also at how political agents interact with each other within the boundaries of those processes. This focus allows for a discussion of the difference this form of interaction makes to people's political standing and consideration within the process (apart from any end state to which the process may lead). Notably, the adoption of this further (not alternative!) perspective brings to the fore the inherent qualities of the forms of interaction inaugurated between citizens as participants in democratic political processes. This kind of appreciation is important because these processes constitute forms of political interaction that may realize in the circumstances of politics such morally worthwhile forms of treatment between citizens as recognition respect.

Surely, people interact with each other in various capacities (as friends, lovers, co-workers), and various ideals could be relied upon to analyze and assess each form of interaction (compassion, affection, reliability). Some such forms of interaction are often considered of significant political import too.[30] All this granted, the interactions between people as political agents who participate in structured political processes can nevertheless retain their specificity. To understand what difference the establishment of the political processes that compose a liberal democracy makes to the standing and consideration of citizens as political agents, we also need to look at what happens while people interact as the occupants of a role, the political role of a democratic citizen, within those processes. In a democracy, such processes include decision-making and deliberative bodies at various levels (for example, national or municipal), of various kinds (for example, electoral or consultative), and with various competences (for example, basic legislation or small-scale policy issues such as urban planning).

29 See, Ceva, *Interactive Justice*, ch. 1. This distinction generalizes and systematizes the divide between distributivist and relational approaches to social equality; see, among others, Anderson, "What Is the Point of Equality?"; Scheffler, "What Is Egalitarianism?"

30 For a classic reference questioning the boundaries between the "personal" and the "political," see Hanisch, *The Personal Is Political*. See also Okin, "Gender, the Public and the Private."

Now, recall the centrality of recognition respect to the liberal democratic political project. This reminder flags a crucial aspect of the analysis and assessment of political processes: their capacity to establish a form of interaction characterized by the respectful reciprocal treatment among citizens. Take one of the most fundamental political processes in a liberal democracy, the democratic decision-making process. By their participation in that process, people bestow upon each other the political standing as mutual authorities that pose morally binding constraints on deciding what each of them may or may not do. Differently put, the democratic decision-making process enacts inherently respectful procedurally regulated relations between the participants in the process. As discussed in the remainder of the article, the realization of this political form of recognition respect is the core of the interactive political morality that sustains liberal democracies. Crucially for our main argument, this consideration offers the context to appreciate the political significance of toleration as an ideal that realizes this form of respect in circumstances of deep political disagreement. In these circumstances, one may not expect that a respectful form of political interaction is regularly—or even often—grounded in either appreciation or indifference. Disapproval is likely to be the norm, and therefore the kind of forbearance secured by the orthodox view of toleration seems to regain the stage. We develop this thought in the next section.

3. TOLERATION IN THE DEMOCRATIC DECISION-MAKING PROCESS AS AN IDEAL OF INTERACTIVE POLITICAL MORALITY

When we revisit from the vantage point of interactive political morality the two relations of toleration we introduced at the beginning of the previous section, a striking difference emerges between them. In the orthodox account of toleration, *A*'s evaluative attitude toward *B* is telling of a type of relation that is not fully reducible to one of mere noninterference, as is the relation between A_1 and B_1 (in which we saw that A_1 may be indifferent or even appreciative of B_1). The two relations are qualitatively different because the former is one yielding to a special kind of noninterference as an expression of forbearance in the face of *A*'s disapproval of *B*. The distinction between end-state and interactive political morality enables us to appreciate how this difference is meaningful.

As discussed earlier, to follow the twofold strategy means to characterize relations of toleration only from the point of view of end-state political morality. These relations, in a liberal democracy, are relations of noninterference (between the state and citizens and among citizens) enacted in the freedom-protecting political arrangements to which the democratic political process must be capable of leading. Such arrangements include, for example, state policies that leave

citizens free to express their opinions or hold marches to manifest their dissent with some majority decision; such policies can plausibly be considered one important aspect of realizing toleration as a core ideal of liberal democracies. However, from the perspective of end-state political morality alone, it makes no difference to *B* whether such policies allow them to live their life as they sees fit because (1) *A* is indifferent to—or, in fact, even appreciative of—*B* (and for that reason *A* does not interfere with *B*) or because (2) *A* disapproves of *B*, yet *A* takes *B* as a constraint on the ways in which *A* may act with respect to *B* (or they may act jointly), and for that reason *A* does not interfere with *B*. Still, to differentiate between the two cases is important in the circumstances of politics. Think, for instance, of such divisive issues as political disputes over the presence of religious symbols in public places, or about the vaccination campaign against COVID-19, with their relative accusations of "bigotry versus laicism" and "obscurantism versus scientism" between the parties. Insofar as collectively binding decisions must be made in such circumstances of deep political disagreement, there is an important space for an ideal capable of giving normative guidance to realize a respectful form of political interaction, *despite* the parties' disapproval. This ideal intuitively calls for a form of political forbearance that the orthodox view of toleration seems distinctively suitable to sustain.

Bluntly put, in the circumstances of politics, the process of collective decision-making requires and entails the establishment of a form of political interaction articulated through relations of forbearance between the participants. By the very fact of submitting to the liberal democratic process the decision of how (many areas of) their lives ought to be governed, the participants in the process *ipso facto* forsake their (actual or counterfactual) power to adjudicate the matter from their own individual perspective as well as the readiness to coerce others into conforming to their own will. Citizens as collective decision-makers are thus enabled—and implicitly required—to recognize each other as mutual authorities concerning the collective decisions by which they should abide. As participants in the process, citizens develop reasons (other than their own evaluative judgments) that should count in establishing their reciprocal treatment. These are practical reasons of forbearance that guide the participants' interaction, as the participants recognize their reciprocal authority as deliberative partners—their negative evaluative judgments notwithstanding.

Differently put, by engaging with each other as participants in the same collective decision-making process, democratic citizens recognize their mutual authority. By that recognition, they refrain from imposing what they may or may not collectively do from their first-personal perspective alone (as a form of coercion or authoritarianism). In so doing, democratic citizens treat each other with recognition respect in the context of decision-making processes

because they treat each other as constraints on what they may do, individually and jointly. The recognition of mutual authority between the participants in the democratic decision-making process realizes a respectful form of interaction despite the (possible or likely) persistence of the participants' disapproval of some of their views. This respectful form of interaction is particular of relations of toleration in the political domain, and is irreducible to a general form of noninterference. Noninterference is not particular of toleration in the same way; it is in fact a feature that relations of toleration share with many other noncoercive power relations in liberal democracies, which may in fact rest on appreciation or indifference.[31]

Let us pause to illustrate concretely how processes may enact a tolerant form of interaction that realizes recognition respect in circumstances of deep disagreement. An illuminating illustration comes from the Public Conversations Project, a US-based organization for the design and facilitation of conversations on divisive issues such as abortion, sexual orientation, and religion.[32] In particular, from 1995 on, leaders of both sides of the abortion debate have met regularly to discuss the issue. Participants in the conversations were quite varied, including people with more or less extreme "pro-life" (e.g., representatives of Women Affirming Life) and "pro-choice" (e.g., representatives of the Planned Parenthood League) positions.

While it is reported that all parties were initially suspicious because of their reciprocal grounds of objection, their antagonistic interaction did change. The change occurred with the aid of two facilitators, by virtue of a procedure that established each participant with the same authority to demand a certain kind of treatment of the other participants and a duty to reciprocate. So, for example, the participants were asked to refrain from using offensive terms (e.g., pro-lifers were asked not to draw any association between pro-choice positions and murder) or stereotypes (e.g., pro-choicers were asked not to presume their opponents were necessarily religious fanatics), despite their reciprocal disapproval. By their own accounts, the participants terminated their encounters still persuaded of their grounds for objection. However, the research also shows that the participants' way of treating each other had changed and, notably, so

31 Note that our discussion rests on the notion of recognition respect, which is different from that of appraisal respect based on people's being an object of esteem. Such a notion could not be compatible with the logic of forbearance, nor—for sure—could it be realized in democratic political processes (as citizens, clearly, are not placed in relations of mutual esteem and appreciation). On the disambiguation of what notion of respect is compatible with forbearance tolerance, see Carter, "Are Toleration and Respect Compatible?"

32 See Fowler et al., "Talking with the Enemy." We borrow the example from Ceva, *Interactive Justice*, ch. 1.

did the kind of deliberation in which they engaged (and refrained from engaging). We can put forth that the participants' commitment not to silence or insult each other despite their objections indicates their developing a new set of practical reasons—alongside and overriding their individual negative evaluative judgment—to recognize their reciprocal standing in their deliberations, thus forbearing each other. These are visibly reasons of forbearance grounded in the participants' recognition as deliberative partners. The importance of this change can be appreciated in full from the perspective of toleration as an ideal of interactive political morality, which realizes one important aspect of the liberal democratic commitment to establishing a respectful form of human interaction in circumstances of deep disagreement.

The same logic underpins our reading of how democratic decision-making may realize toleration in itself (or is "inherently tolerant"). This process enacts a respectful form of interaction between citizens who forbear each other as political agents in circumstances of deep political disagreement. As discussed, despite their objections, the participants in the process partake in the same authority to decide over each other as concerns the very content of their rights and the contours of their freedoms. In this sense, the tolerant relations of forbearance in the face of disapproval, which we have seen at work in such experimental environments as that of the Public Conversations Project, are institutionalized in democratic decision-making processes. Such processes may be inherently tolerant in the sense that they enact toleration in themselves, in virtue of the forms of interaction they constitute between those who participate in them (not only insofar as they cause tolerant political arrangements).

This particular claim rests on a general view of political processes as more than a set of regulative rules and procedural mechanisms. The processes that compose the public order are institutions in the sense of systems of interrelated rule-governed embodied roles.[33] That such roles are embodied means that the analysis and assessment of political processes may not be reduced to the analysis and assessment of the regulative rules that govern those processes, possibly in virtue of their capacity of leading to certain end states. Such an analysis and assessment must also be cognizant of the *constitutive* rules of the process. These are rules that establish new forms of interaction between the participants in the process and make them possible.[34] These forms of interaction occur through the use of normative powers (rights and duties) that people come to possess only because they occupy a role within a process. The process "institutes" the people

33 Applbaum, *Ethics for Adversaries*; Emmet, *Rule, Roles and Relations*.

34 Searle, *Speech Acts* and *The Construction of Social Reality*; see also Hindriks, "Constitutive Rules, Language, and Ontology"; Ceva, *Interactive Justice*.

who occupy a role within it into a normative status that the role-occupants only have (and upon which they may act) within the boundaries of the process.[35]

This idea elucidates the logic of one of the most fundamental role attributions in the democratic political process: the role of a citizen as a collective decision-maker.[36] People do not normally have the normative power (the right, or the authority) to decide what others may or may not do with their lives. Nor are people normally subjected to the normative power (the duty) to follow other people's determinations of their margins of personal action. Still, as seen, this kind of normative relation is perfectly normal and sensible between democratic citizens when they exercise their normative powers over each other, for example through voting, as parties in the democratic decision-making process. This process is sustained by a special kind of political morality; this political morality is interactive in the sense that it concerns the process-based relations between people in a certain institutional capacity. The mutual authority that the democratic decision-making process bestows upon the participants in the process is thus of a special kind: it is an authority that people may only exercise jointly and over each other within an institutional context.[37] This authority is an entailment of people's acting on the powers bestowed upon them by the constitutive rules of the democratic decision-making process. This mutuality differentiates the authority of the democratic decision-making process from the authority each person has over herself (which such other regimes as anarchies realize too) and from the kind of authority some people may unilaterally have over others (such as the authority realized in an aristocracy). The main claim we make here is that the value of the democratic decision-making process can be understood as a form of recognition respect, which can be realized in the circumstances of politics because it enacts an inherently tolerant form of interaction characterized by the parties' forbearance.

The last consideration is important to capture one central aspect of our qualified defense of toleration. This aspect can be fleshed out by contrast with Rainer Forst's argument that toleration is realized in democratic deliberation to the extent that citizens may not refer to their controversial ethical views as a ground for objecting to the views of others when making collective decisions.[38] For Forst, any such reference would lead to coercive decisions. As such, such reference is disrespectful as a violation of the moral authority that people have

35 The reasoning structure here is the same as that at work in Rawls's "practice conception" of rules ("Two Concepts of Rules").

36 Ceva and Ottonelli, "Second-Personal Authority and the Practice of Democracy."

37 In this spirit, Ceva and Ottonelli discuss democratic voting as a primitive illustration of the practice of democracy ("Second-Personal Authority and the Practice of Democracy").

38 Forst, *The Right to Justification*, 146.

over themselves.[39] This account does not explain why exactly this expression of respect can uniquely be achieved in virtue of the tolerance that the democratic decision-making process realizes in itself. Our account suggests one such explanation by showing that by establishing a tolerant form of interaction, the democratic decision-making process does more than, and something different from, protecting people's authority *over themselves* from arbitrary coercion. The establishment of this process puts people in relation in such a way that enables them as political agents who recognize their mutual standing as the final political authorities *over each other* in collective decision making, despite their grounds for objection. As seen, in the making of collectively binding decisions, this kind of mutual authority can only be enacted in politics in the tolerant form of interaction, articulated through relations of forbearance, that the democratic decision-making process establishes. Absent this process, the tolerant form of interaction in which this form of recognition respect consists could not possibly happen in the circumstances of politics. Consequently, people could not bestow upon one another the relevant status as mutual political authorities that sustains a liberal democracy. It is by adopting the perspective of interactive political morality that we can appreciate this point.

The realization of toleration in democratic political processes as an ideal of interactive political morality is important even when the outcomes of those process are unsettled, or end up frustrating the claims of some of the parties. As we expound in the next section, the outcomes of a tolerant process may fail toleration as an ideal of end-state political morality. And, surely, some of those frustrations may be unjust. But the realization of toleration as an ideal of interactive political morality is not idle or unimportant even when it stands on its own two feet.

4. The Complex Evaluation of Toleration in Political Processes

One of the features of the defense of toleration we have put forth in this article is its philosophical parsimony. Differently from the reinterpretive efforts undertaken by the proponents of the twofold strategy we reviewed in section 1, our discussion does not require us to rethink the ideals that are commonly thought to sustain the liberal democratic political project and the place of toleration within it.

However, our defense also has implications that make the analysis and assessment of political processes more complex. Indeed, we have encouraged an extended consideration of the liberal democratic political project. This

39 Forst, *The Right to Justification*, 21.

consideration includes the analysis and assessment of political processes also in virtue of the forms of political interaction they constitute, not only the political arrangements they cause. This inclusion calls for a joint analysis and assessment of the tolerance realized in and by democratic political processes from the perspectives of interactive and end-state political morality.

To bring together the evaluative perspectives of interactive and end-state political morality is important but challenging. We think that neither perspective is indeed sufficient, taken on its own, to allow for a complete assessment of democratic political processes through the lenses of toleration. Differently put, the all-things-considered normative evaluation of democratic political processes is a complex exercise that may also be internally conflicting. This is because a harmonious joint realization of toleration as an end-state and interactive ideal can prove at times impossible. However, the adoption of each of these two perspectives offers important insights for a *pro tanto* assessment. Let us explore these claims.

To assess political processes through the lenses of toleration two discrete judgments are relevant as concerns whether those processes (a) realize toleration in themselves or (b) are capable of leading to tolerant political arrangements outside the process. Thus, the first site of toleration is internal to political processes. In this first sense, as discussed in section 3, processes realize toleration in themselves insofar as they constitute relations of forbearance between the participants. Such relations of forbearance are valuable insofar as they enact a respectful form of interaction between citizens as political agents in circumstances of deep political disagreement. The second site of toleration is external to political processes. In this second sense, defended by such champions of the twofold strategy as Balint and Jones, processes realize toleration insofar as they result in a form of political noninterference in society. Such forms of political noninterference are valuable insofar as they are capable of generating political arrangements that protect individual negative freedoms. The capacity to distinguish between these two sites of toleration is analytically salient to offer a nuanced evaluation of important aspects of liberal democracies. Some normative challenges emerge too to the extent that the enactment of toleration in the two sites of political interactions and political end states may at times be mutually supportive but also unsupportive. To wit, because the adoption of each of the two discrete perspectives can only give us a ground for a *pro tanto* evaluation, we should expect circumstances in which difficult trade-offs between the two aspects are necessary.

Think, first, of a parliamentary decision that decriminalizes the possession of cannabis for recreational use. The outcome of the parliamentary decision-making process may be tolerant (in the permissive sense) to the extent

that citizens are thereby free from the state's interference with their possession and use of cannabis. However, the process may realize toleration (or not) in itself depending on whether the participants' interaction was structured in such a way that none of the participants was silenced on the ground of other participants' objections toward their particular views. Sometimes we can tick both boxes, but other times we must make disjunct assessments. There can thus be inherently tolerant, as it were, decision-making processes that lead to non-tolerant *qua* freedom-restricting decisions, such as an egalitarian process that culminates in the prohibition of selling tobacco products. But we can also see tolerant policies promoted through non-tolerant processes; think of a policy that allows women to drive cars, thus enhancing their freedom of movement, which is enacted through a male-dominated decision-making process objecting to women's deliberative capacities (whereas their driving skills are recognized).

Consider another example concerning the enfranchisement of such minority groups as third-country migrants in the European Union. Their inclusion in the collective decision-making process changes the institutionalized interaction between majorities and minorities. What changes is the recognition of people's capacity as political agents, by calling them to recognize each other as equally active parties in the political game of mutual authority established by the democratic decision-making process. This change reflects a transformation of the consideration of the minority members' standing, who, once enfranchised, can be heard as authoritative political agents addressing claims in their own institutionalized voice. What is more, the gaining of such a standing occurs despite the persistence of deep political disagreement. This transformation occurs when the constitutive rules of the process grant all participants an equal voice, typically by the rule "one head, one vote," or by enacting rules of order that grant all participants in deliberative processes of consultation a fair hearing. However, fair hearing per se does not presuppose the prospect of an eventual resolution of disagreements, nor does it entail the requirement that any one minority's voice equally finds representation in the final outcome. The enactment of fair hearing signifies a forbearing interaction, but does not preclude by itself an outcome that frustrates some of the participants' preferences.

Such a consideration tells of the complexity of the normative evaluation of political processes through the lenses of toleration. It suggests how enacting toleration as an ideal of end-state and interactive political morality may be internally conflicting in a way that paves the way to moral dilemmas that imply inevitable moral losses. For instance, consider the attempts to restrict the individual political rights of right-wing extremists that have been pursued, but so far failed, in Germany. Article 18 of the Basic Law for the Federal Republic of Germany allows for the "forfeiture of basic rights" if exercised to "combat

the free democratic basic order."[40] So far, the Constitutional Court has ruled that the individual behavior of right-wing extremists does not pose a sufficient threat to the public order to justify the infringement of citizens' rights of political participation. In our terms, this ruling suggests that no sufficient reasons have been offered to restrict toleration as a form of political interaction that is enacted in those citizens' inclusion in the political process of collective decision-making through the attribution and exercise of their voting rights. This ruling, which enacts interactive tolerance, bears the risk of yielding to a degree of end-state intolerance, should the political views of right-wing extremists gain sufficient political traction to result in a restriction of other citizens' freedoms (for example, by curtailing their civic rights). Such an implication might, in turn, give reasons to revise the decision made on the ground of the court's ruling, thus reducing the interactive tolerance of the process in the future (down to the furthest-reaching implication of denying political representation to extremist positions). In either case, we can see that the joint enactment of toleration in the process and/or its resulting arrangements may sometimes be impossible, and call for difficult trade-offs that imply a measure of moral loss.[41]

How to deal appropriately with the conflicts possibly arising in the joint realization of toleration as both an ideal of interactive and end-state political morality is a matter for another time. Circumstantial (for example, prudential) considerations may speak in favor of prioritizing the realization of one aspect over the other on a case-by-case basis. Think of societies where the process of recovering from past collective trauma is still ongoing so that sacrifices in terms of the ideal of interactive tolerance may ultimately be justifiable for the sake of preserving unstable social peace (and possibly avoiding grave end-state injustices).[42]

Ultimately, the claim that the establishment of political processes that inherently realize toleration may be valuable in its own right does not make for an absolute argument for enacting toleration as an ideal of interactive political morality. Each of the perspectives contributes with *pro tanto* considerations

40 For discussion, see Müller, "Individual Militant Democracy."

41 This position is compatible with multiple strategies of containment of extremist parties or citizens, e.g., refusing campaign contributions from certain lobby groups, or creating a *cordon sanitaire* around extremist movements and parties. On the latter point, see, Rummens and Abts, "Defending Democracy." For a discussion of "informal exclusion" as a powerful instrument of containment that must fall short of "formal exclusion," see Dovi, "In Praise of Exclusion."

42 An example would come from post-genocide Rwanda, where political party bans have targeted associational political rights by banning parties that revive the very ethnic divisions underpinning the past violence. For discussion, see Niesen, "Political Party Bans in Rwanda 1994–2003."

to the assessment of political processes, but all-things-considered judgments might be difficult to attain. Our claim is that there is an important moral value inherent to democratic political processes, whose moral significance may not be entirely reduced to their capacity of leading to certain results. By recognizing the presence of these tensions, our argument does not certainly make the assessment of democratic political processes any less simple or straightforward. But it has the advantage of fleshing out two otherwise confused dimensions of political morality. The advantage of this operation resides in the clarification of the possible kinds of evaluations of political processes, as well as the related possible sources of disagreements or contestation of the features of those processes and their outcomes.

5. CONCLUSION

We have proposed a qualified defense of toleration as an ideal of interactive political morality inherent to democratic political processes. We have also shown how such a defense allows us to appreciate one aspect of the relational structure of toleration, that the orthodox view of toleration as forbearance uniquely captures (but recent views underplay). This aspect concerns the establishment of a respectful form of interaction between citizens as political agents in circumstances of deep political disagreement. We have thus pinpointed an important sense in which appeals to toleration are consistent with the commitment to realizing one of the most fundamental ideals of the liberal democratic political project and retain, therefore, their significance within that project, against any concern of redundancy.

We have seen how the relations of mutual authority established between the participants in such political processes as the democratic decision-making process are relations in which the participants recognize each other as a constraint on their individual and joint actions. The participants partake in the same political authority over each other, and yet may preserve their reasons to object to some of their practices or beliefs. This form of democratic interaction is inherently tolerant in accordance with the liberal orthodoxy. Democratic processes establish a form of tolerant human interaction that could not exist absent those processes and is qualitatively different from relations of domination and coercion, but also mutual appreciation or indifference.

The democratic decision-making process can ultimately be seen as a locus for the realization of an important form of toleration. This feature can make the democratic decision-making process valuable *qua* respectful in its own right—that is, independently of whether the end states thereby generated are themselves tolerant. To be sure, the realization of toleration may be in tension

with that of other normative commitments, as is unsurprisingly the case in such a pluralistic project as that of a liberal democracy. However, we hope we have shown how the enactment of toleration as an ideal of interactive political morality gives substance to one of the defining commitments of a liberal democracy. Such a commitment concerns the realization of recognition respect for persons in the circumstances of deep political disagreement.[43]

University of Geneva
emanuela.ceva@unige.ch

University of Warwick
rossella.de-bernardi@warwick.ac.uk

REFERENCES

Applbaum, Arthur. *Ethics for Adversaries: The Morality of Roles in Public and Professional Life*. Princeton: Princeton University Press, 1999.

Anderson, Elizabeth. "What Is the Point of Equality?" *Ethics* 109, no. 2 (January 1999): 287–337.

Balint, Peter. *Respecting Toleration: Traditional Liberalism and Contemporary Diversity*. Oxford: Oxford University Press, 2017.

Bird, Colin. "Mutual Respect and Neutral Justification." *Ethics* 107, no. 1 (October 1996): 62–96.

Brown, Wendy. *Regulating Aversion: Tolerance in the Age of Identity and Empire*. Princeton: Princeton University Press, 2008.

Carter, Ian. "Are Toleration and Respect Compatible?" *Journal of Applied Philosophy* 30, no. 3 (August 2013): 195–208.

Ceva, Emanuela. *Interactive Justice: A Proceduralist Approach to Value Conflict in Politics*. New York: Routledge, 2016.

———. "Toleration." In *Oxford Bibliographies in Philosophy*, edited by Duncan Pritchard. New York: Oxford University Press, 2013.

Ceva, Emanuela, and Valeria Ottonelli. "Second-Personal Authority and the

43 A predecessor of this paper was presented at York University (Toronto). We are grateful for the feedback received on that occasion and to Peter Balint, Michele Bocchiola, Francesco Chiesa, Andrew J. Cohen (and his students!), Peter Jones, Alasia Nuti, Fabienne Peter, and Federico Zuolo for written comments on earlier drafts. Rossella De Bernardi is grateful to the Morrell Centre for Toleration (University of York) and the Arts and Humanities Research Council (grant AH/L503848/1, through the White Rose College of the Arts and Humanities) for support received while working on this project.

Practice of Democracy." *Constellations* 29, no. 4 (December 2022): 460–74.
Christiano, Thomas. *The Constitution of Equality*. Oxford: Oxford University Press, 2008.
Darwall, Stephen. *The Second-Person Standpoint: Morality, Respect, and Accountability*. Cambridge, MA: Harvard University Press, 2006.
———. "Two Kinds of Respect." *Ethics* 88, no. 1 (October 1977): 36–49.
Dovi, Suzanne. "In Praise of Exclusion." *Journal of Politics* 71, no. 3 (July 2009): 1172–86.
Emmet, Dorothy. *Rule, Roles and Relations*. London: MacMillan, 1966.
Forst, Rainer. *The Right to Justification: Elements of a Constructivist Theory of Justice*. Translated by Jeffrey Flynn. New York: Columbia University Press, 2011.
———. *Toleration in Conflict: Past and Present*. Translated by Ciaran Cronin. Cambridge: Cambridge University Press, 2013.
Fowler, Anne, Nicki Nichols Gamble, Frances X. Hogan, Melissa Kogut, Madeline McCommish, and Barbara Thorp. "Talking with the Enemy." *Boston Globe*, January 28, 2001.
Galeotti, A. Elisabetta. *Toleration as Recognition*. Cambridge: Cambridge University Press, 2002.
Gaus, Gerald. *The Order of Public Reason: A Theory of Freedom and Morality in a Bounded World*. New York: Cambridge University Press, 2010.
Gray, John. "Pluralism and Toleration in Contemporary Liberal Philosophy." *Political Studies* 48, no. 1 (May 2000): 323–33.
Hanisch, Carol. "The Personal Is Political." http://www.carolhanisch.org/CHwritings/PIP.html. Originally published in *Notes from the Second Year: Women's Liberation*, edited by Shulamith Firestone and Anne Koedt. New York: Radical Feminism, 2006.
Heyd, David. "Is Toleration a Political Virtue?" In *Toleration and Its Limits*, edited by Jeremy Waldron and Melissa S. Williams, 171–94. New York: New York University Press, 2008.
Hindriks, Frank. "Constitutive Rules, Language, and Ontology." *Erkenntnis* 71, no. 2 (September 2009): 253–75.
Jones, Peter. "Legalising Toleration: A Reply to Balint." *Res Publica* 18, no. 3 (2012): 265–70.
———. "Making Sense of Political Toleration." *British Journal of Political Science* 37, no. 3, (July 2007): 383–402.
———. "Toleration and Neutrality: Compatible Ideals?" In *Toleration, Neutrality and Democracy*, edited by Dario Castiglione and Catriona McKinnon, 97–110. Boston: Kluwer Academic Publisher, 2003.
Kant, Immanuel. *An Answer to the Question: What Is Enlightenment?* 1784. London: Penguin, 2009.

Kolodny, Niko. "Rule over None II: Social Equality and the Justification of Democracy." *Philosophy and Public Affairs* 42, no. 4 (Fall 2014): 287–333.

Larmore, Charles. "Political Liberalism." *Political Theory* 18, no. 3 (August 1990): 339–60.

Meckled-Garcia, Saladin. "Toleration and Neutrality: Incompatible Ideals?" *Res Publica* 7, no. 3 (October 2001): 293–313.

Müller Jan-Werner. "Individual Militant Democracy." In *Militant Democracy and Its Critics: Populism, Parties, Extremism*, edited by Anthoula Malkopoulou and Alexander S. Kirshner, 14–37. Edinburgh: Edinburgh University Press, 2019.

Newey, Glen. "Is Democratic Toleration a Rubber Duck?" *Res Publica* 7, no. 3 (October 2001): 315–36.

———. "Metaphysics Postponed: Liberalism, Pluralism and Neutrality." *Political Studies* 45, no. 2 (June 1997): 296–311.

Niesen, Peter. "Political Party Bans in Rwanda 1994–2003: Three Narratives of Justification." *Democratization* 17, no. 4 (August 2010): 709–29.

Okin, M. Susan. "Gender, the Public and the Private." In *Feminism and Politics*, edited by Anne Philips, 116–42. Oxford: Oxford University Press, 1999.

Quong, Jonathan. *Liberalism without Perfection*. Oxford: Oxford University Press, 2011.

Rawls, John. *Political Liberalism*. 3rd ed. New York: Columbia University Press, 2005.

———. "Two Concepts of Rules." *Philosophical Review* 6, no. 1 (January 1995): 3–32.

Rummens, Stefan, and Koen Abts. "Defending Democracy: The Concentric Containment of Political Extremism." *Political Studies* 58, no. 4, (September 2010): 649–65.

Scheffler, Samuel. "What Is Egalitarianism?" *Philosophy and Public Affairs* 31, no. 1 (Winter 2003): 5–39.

Searle, John. *The Construction of Social Reality*. London: Penguin, 1995.

———. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press, 1969.

Waldron, Jeremy. *Law and Disagreement*. Oxford: Oxford University Press, 1999.

———. "Theoretical Foundations of Liberalism." *Philosophical Quarterly* 37, no. 147 (April 1987): 127–50.

Journal of Ethics and Social Philosophy
VOL. 23, NO. 3 · JANUARY 2023
https://doi.org/10.26556/jesp.v23i3.1611

SLACK TAKING AND BURDEN DUMPING

FAIR COST SHARING IN DUTIES TO RESCUE

Aaron Finley

Globally, millions of individuals need rescue from disease, natural disaster, poverty, and violence. If everyone in a position to perform rescues did her fair share, no one's share would be large. But when some individuals fail to do their part, how much slack must others take up? Peter Singer, Peter Unger, and others have argued that we have very stringent duties to do more when others do less.[1] Many, including J. L. Cohen, Liam Murphy, and David Miller, have argued in response that principles requiring one to take up slack are objectionably unfair. These principles, they argue, demand too much from conscientious individuals by requiring them to do not only their share but also the shares of those who neglect to do their part. Even worse, the principles seem to let the morally negligent off the hook by making their burdens the responsibility of others.

I agree that contributing more than one's fair share to a rescue effort is unfair but disagree that principles are the source of the unfairness. Instead, by shirking their responsibilities, noncontributors unfairly dump part of the burdens they should have borne onto others. Thus, the conduct of burden dumpers, far from being permissible, constitutes a double wrong—they wrong those they fail to rescue, and they wrong those on whom their burdens fall. On this approach, those who do their part have an obligation to take up at least some slack, and burden dumpers remain responsible for failing to do their part.

Importantly, I do not defend the act-consequentialist position that those who do their part must take up *all* the slack left by others. The view I defend here is consistent with deontological views that posit a duty to perform rescues so long as they are not too costly. So long as my fair share of the burdens under

1 Singer, "Famine, Affluence and Morality"; Unger, *Living High and Letting Die*. Singer defends both a strong and a weak principle of beneficence and argues that both are very demanding. In this article, I set aside questions about the demandingness of our duties. Thus, I will usually mention the strong version of Singer's principle, not because I take it to be most plausible, but because it most sharply highlights the contours of the debate.

full compliance is less than the maximum this duty could require of me, others failing to do their share increases my burdens. This article does not address the quantity of slack that must be taken up, arguing instead that those who leave slack treat slack takers unfairly.

Because my central focus is on fairness rather than demandingness, I aim to describe cases in which our natural duty to rescue is clear. Singer's famous drowning-child example will therefore be central. Sadly, however, children drown in the real world as well. In 2013, a vessel left Libya carrying around five hundred migrants. En route, it caught fire and sank off the coast of Lampedusa, a small Italian island in the Mediterranean. Upwards of three hundred of those on board died. The incident attracted international attention, and Italy used its navy to begin a search and rescue program called Mare Nostrum, credited with rescuing some two hundred thousand people during the year it operated. However, due to the cost of the program, Italy appealed to the rest of the European Union for help. In response, Frontex, the EU's border and coast guard agency, was tasked with replacing Italy's program with a new one—Operation Triton. Triton has received criticism for focusing primarily on border control rather than search and rescue, which has left a serious humanitarian crisis in the Mediterranean as thousands of people die or go missing each year attempting to cross from northern Africa to Europe.

The crisis in the Mediterranean provides a vivid example of the kind of large-scale, ongoing rescue efforts we face. Italy recognized a duty to perform at least some rescues, and other EU member states seemingly recognized an obligation toward Italy to share the burdens of performing those rescues. Italy claimed it would be unfair for it to bear the burden of performing all the rescues alone, and others in the EU apparently agreed. It is this intuitive connection between natural duties and fairness obligations that I develop here. I argue that when duties to rescue require someone to do more than her fair share (the amount she would have to do under full compliance), she is being treated unfairly by the people who fail to do their part. This argument draws on features of the literature on group causation and moral responsibility. In particular, I combine Alvin Goldman's vector theory of causation with David Brink and Dana Nelkin's fair-opportunity theory of responsibility. I argue that noncontributors treat contributors unfairly by failing to do their part when (a) the failure derives from a blameworthy lack of responsiveness to features of a situation (such as drowning children or overly burdened rescuers) that give one moral reasons to act, and (b) the failure imposes burdens by leaving slack that contributors must take up.

I lay the groundwork for addressing questions of responsibility under partial compliance in section 1 by articulating an account of the content of our duty to rescue. In section 2, I address a puzzle related to the following question: On

whom exactly do burdens fall when noncontributors fail to do their part? After elucidating the puzzle, I defend a solution in the context of nondiscretionary duties to rescue.[2] In section 3, I expand on the arguments developed in sections 1 and 2 and show that they cover discretionary duties to rescue. More specifically, I argue that when one culpably fails to do one's part, one is implicated in generating the burdens one's failure, together with the similar failures of others, produces. This means that when one's duty to rescue is discretionary, one treats all contributors unfairly by failing to do one's part. In section 4, I consider some objections, and in section 5, I consider further applications of the theory focusing on voting and climate change.

1. THE NO-BURDEN-DUMPING INTUITION

In his landmark paper "Famine, Affluence and Morality," Peter Singer argues that we are obligated to use our resources to rescue those dying from lack of food, shelter, and medical care up to the point of marginal utility—the point at which further sacrifice would make us worse off than those we are helping.[3] He then considers a series of objections, one of which concerns fairness: if each affluent person contributed to ending this kind of suffering, no one would have to donate more than a few dollars.[4] We are all morally required to contribute, so is it not grossly unfair that I, the conscientious person, must donate to the point of marginal utility simply because others are not doing their part? In response, Singer says it is unfortunate that others are not contributing, but that does not change the fact that we have a duty to rescue as many as we can even if others

2 A nondiscretionary duty is a duty with only one means of fulfillment. If I promise to do *X*, I do not keep my promise unless I do *X*. A discretionary duty is one that I may fulfill as I see fit. If I have a duty to help the badly off, I could work at a local homeless shelter, donate to Oxfam, dig wells, and so on. The contrast between the two types of duties is not deep. A nondiscretionary duty is just a discretionary duty with only one fulfillment option. However, the distinction is useful because nondiscretionary duties are an important subclass of duties and are easier to analyze than discretionary duties.

Some theorists identify non-discretionary duties with Kantian imperfect duties. Murphy briefly discusses this view in *Moral Demands in Nonideal Theory*, 71–72; and Igneski, "Perfect and Imperfect Duties to Aid," analyzes duties to aid in terms of Kantian perfect and imperfect duties. For further discussion of imperfect duties as such, see Baron, "Kantian Ethics and Supererogation"; and Hope, "Kantian Imperfect Duties and Debates over Human Rights."

3 Singer, "Famine, Affluence and Morality," 507. This is the strong version of Singer's argument. The weak version says only that we must give until doing so would force us to sacrifice something morally important. Since both the weak and strong versions are very demanding, one can raise the fairness objection to both.

4 Pogge, "Are We Violating the Human Rights of the World's Poor?"

are not. For Singer, the decisive consideration is that "by giving more than £5 [what I would give under full compliance] I will prevent more suffering."[5]

I will assume that we do have a duty to rescue and that this duty does require us to do more when others do less. However, I set aside the question of how much more we are required to do.[6] What I want to draw out is the intuition that those who do not contribute to the rescue effort wrong not only those they fail to rescue but also those who take up the slack. I will refer to those who do not do their part, and thereby leave more work for others, as *burden dumpers*. The claim I defend is not about how much can be demanded of us but about who or what is at fault when that demand is unfair. Cohen, Murphy, and Miller all argue that the principle making the demand is the source of the unfairness, but I argue that it is not. Rather, the unfairness originates in the people who neglect their duties.

Throughout the discussion we must carefully separate the wrong of neglecting one's duty to rescue from the wrong of burden dumping. Consider a variation of Singer's drowning-child example. I and another person are near a pond in which two children are drowning. The other person and I could easily save one child each. However, I see that if I do nothing, the other person will be able to save both children, though just barely. I decide to do nothing, and the other person saves both children. I will say that the child has a *deontic* complaint against me because she had a duty to her that I failed to fulfill. In general, deontic complaints arise when one fails to fulfill an individual duty to another agent that is not generated by a maldistribution of resources.[7] I will say that the other rescuer has a *fairness* complaint against me because she had to do more

5 Singer, "Famine, Affluence and Morality," 507.

6 For a book-length discussion of demandingness in the context of taking up slack, see Murphy, *Moral Demands in Nonideal Theory*. There (and in Murphy, "The Demands of Beneficence"), he argues that we are never required to do more than we would be if everyone were doing her part, even in drowning-child cases. I do not have the space to take up his arguments here, but for concise and forceful replies, see Horton, "International Aid"; and Horton, "Fairness and Fair Shares." Horton argues that one's objection to doing more than one's fair share becomes increasingly weighty the more slack one must take up. Horton suggests that this unfairness, in addition to the extra costs one bears, weighs against one's duty to take up slack past a certain level of sacrifice. However, I am inclined to agree with Karnein, "Putting Fairness in Its Place," that this kind of unfairness does not weigh against one's duty to take up slack. Instead, it should be counted against noncontributors in determining, for example, what kind of compensation they might owe to those who took up their slack.

7 Ridge, "Fairness and Non-Compliance," suggests that when I fail to perform a rescue and no one takes up my slack, I treat the unrescued person unfairly. This claim is sensible since my failure produces a maldistribution of burdens. Because I failed to bear the burden of performing the rescue, the person in need of rescue must bear the consequences of

than her fair share because of my culpable failure to do my part.[8] This case raises questions that I will briefly address before turning to a puzzle about collective burden dumping.

Imagine that the other rescuer and I are positioned such that it is initially unclear which child I should save. When I perform no rescues, have I wronged both children or neither? Do either of the children have a deontic complaint against me? A detailed discussion is beyond the scope of this paper, so here I suggest that my duty only becomes particular once it is clear which child the other rescuer is going to save. We might think that neither child has a right against me that I save *her*, though, plausibly, each has a right that I "save as many of them as [I] could without unreasonable risk to [myself]."[9] Whatever rights the children might have held against me, it is clear that I had a duty to rescue at least one of them and that I wrong both children by simply ignoring it.

What if the burden I shirk is too heavy for the other person to carry, but she can still shoulder some of it? Suppose there are three children drowning and that I could save two children as easily as the other swimmer could save one. Other things equal, my duty is to save two, while the duty of the other is to save one. However, I save none. Through tremendous exertion, the other rescuer saves two children, but the third child still drowns. Clearly the third child has been wronged, but by whom? Given the language of burden dumping, one might think that I dumped my duty to save my two children onto the other rescuer, so that only she wrongs the third child by failing to rescue her. If nothing else, the ought-implies-can principle entails that the second rescuer is not obligated to save all three children. But we can be more precise about each rescuer's obligations: each must perform as many rescues as she can given (1) her relevant abilities (for example, how strong a swimmer she is), (2) the scope of the need, (3) the total costs she can reasonably be required to bear, and (4) the portion of the need others can be expected to satisfy.[10]

---

remaining unrescued. This observation raises questions about the proper delineation of duties by kind that I do not have space to address here.

8 Perhaps the other rescuer has a right against me that I not impose undue burdens on her. Even so, the complaint is about fairness because it concerns a maldistribution of resources. I forced her to use her resources to perform a rescue when my resources should have been expended.

9 Feinberg, "The Moral and Legal Responsibility of the Bad Samaritan," 61. For Feinberg, the sign that this right exists is our sense of moral indignation at potential rescuers when they do nothing (64). See Agnafors, "On Disjunctive Rights," for a further defense of disjunctive rights; and Wolterstorff, *Justice*, ch. 11, for a general discussion of correlativity between rights and duties.

10 Even Singer's strong principle would endorse condition 3. I might be able to give past the point of marginal utility, but Singer thinks I am not morally required to.

Condition 4 becomes important in cases in which burdens must be fairly distributed. If fairness does not demand that others share my burdens, I cannot expect them to contribute anything and conditions 1–3 determine what I am obligated to do. However, once fairness comes into the picture, we might worry that 4 gives noncontributors a free pass to dump their burdens so long as others are willing to take up the slack. If I am in the presence of several conscientious individuals, I might know that if I do nothing, all the drowning children will be rescued. Given this, 4 seems to let me off the hook. Because others can be expected to do everything, I have no obligation to do anything. What is more, as L. J. Cohen worries, if everyone knows there is a duty to take up slack left by noncontributors, even one who is inclined to contribute "could legitimately infer that, if he failed to do so, those with tenderer consciences than himself would make good the deficiency. So any temptation that he might have to withhold his own contribution would be reinforced by the belief that . . . the ultimate outcome would be the same."[11]

This objection highlights an ambiguity in the notion of expectation employed in condition 4. On the one hand, according to a fair distribution, others can be expected—in the sense of being normatively required—to contribute their initial fair share. On the other, according to their actual attitudes, they can be expected—in the sense of being predicted—to contribute whatever they are willing to contribute, which may be as little as nothing. Both notions of expectation are relevant here, and both generate obligations. According to fairness, one is responsible for one's initial fair share of the burdens even if one contributes nothing. But if some can be expected to contribute less than their fair share (according to their actual attitudes), the rest of us are obligated to take up their slack. One person's unwillingness to do her part affects the scope of the need facing others—condition 2—without changing the portion of the need she is normatively required to address. In this way, those of us who contribute become responsible for the burdens of noncontributors, even though the noncontributors remain responsible for their share of the rescues.[12]

What follows is that in the case in which I, in fairness, ought to save two of the three drowning children but save none while another rescuer does her best and saves two, only I wrong the third child. However, I also wrong the second

11 Cohen, "Who Is Starving Whom?" 73–74.

12 What I say here may not fully address Cohen's worry about temptation. On one level, since condition 4 does not let noncontributors morally off the hook when others take up their slack, no one can be tempted by the possibility of avoiding wrongdoing while also failing to do her part. But if Cohen's point is merely psychological, I have nothing to say one way or another. See Ridge, "Fairness and Non-Compliance," for an argument against Cohen's claim about perverse incentives.

child by failing to fulfill my duty to her. What is more, if the other rescuer saved only one child, both of us would wrong both of the other children. I obviously wrong both since I could have saved both, and the other rescuer wrongs both since she could have saved either. She is guilty of a deontic failing toward them by violating her duty to save as many as she can without unreasonable risk to herself. So, my culpable failure to rescue makes a similar culpable failure possible for the other rescuer. As the number of potential rescuers grows, there is no upper bound to the amount of morally culpable wrongdoing a single problem can produce so long as we are all duty bound to solve it.[13]

2. COLLECTIVE BURDEN DUMPING

At this point, a puzzle might seem to arise. Suppose that six children are drowning; Jones, Smith, and I are the only potential rescuers; and each of us has a nondiscretionary duty to rescue the children. We are all on a par as swimmers, and each of us can save two children easily but cannot save more than three. In this case, each of us ought to save two children—that is a fair distribution of rescue-related burdens. I immediately rescue two children. By the time this is done, I see that Smith and Jones intend to save no children. So, my obligation to take up slack kicks in, and I save a third child, after which it is too late for the other children.

Intuitively, I have a fairness complaint against Smith and Jones for imposing the burden of performing a third rescue. Both fail to contribute to the rescue effort, so both play a role in the extra burdens I bear.[14] But Jones might say, "Smith was unwilling to help, so if I had helped, you and I would have saved three children each. You were already saving three children, so *I* dumped no burdens on you." And Smith could say the same. (Call this case *partial help*.)

Smith and Jones's argument seems sensible because it appeals to an intuitively plausible characterization of what it means to play a role in someone's burdens. According to Jones and Smith, one plays a role in another's burdens only when one's contribution would alone be sufficient to reduce the burdens borne by contributors. The complication in this case is that Smith and Jones impose burdens jointly, not individually. So, is there a defensible sense in which each plays a role in my burdens even though neither, acting alone, could reduce

13 See Karnein, "Putting Fairness in Its Place," for an argument that comes to similar conclusions on this point.

14 I use the admittedly awkward phrase "play a role" to avoid using the word "contribute" to refer to opposite phenomena—contributing to rescue efforts and contributing to burdens by failing to contribute to rescue efforts.

them—a defensible sense in which I can still properly raise a fairness complaint against each?[15]

Alvin Goldman defends a potential answer to this question in his analysis of the obligation to vote. Those who vote or refrain from voting almost never cast or withhold a decisive ballot. Thus, those who do not vote, or who vote for a bad candidate, can run an argument parallel to Jones and Smith's. Each person can say that *her* vote or abstention did not affect the outcome of the election, so *she* should not be held responsible.

Goldman responds to this objection by developing what he calls a vector-system analysis of causal contributions. He explains:

> [A vector is a sum] computed from three kinds of forces: (1) forces that are positive in the direction of movement, (2) forces that are negative in the direction of movement, and (3) forces that are zero in the direction of movement. Finally, when thinking about the causation of a given movement, we think of each positive force as a *contributing factor* in the production of the movement, each negative force as a *counteracting*, or *resisting*, factor in the production of the movement, and each zero force as a *neutral factor* vis-à-vis the production of the movement.[16]

Each person who casts a vote for the winning candidate is a causal contributor to—or, in my terms, plays a role in—that person's victory. Similarly, in the case of Jones and Smith, each plays a role in the extra burdens I bear since the inaction of each is a contributing factor in them.[17] But if Jones helps rescue while Smith does not, Jones's action counts as a force in the direction of distributing burdens fairly. Thus, even though Jones's action does not reduce the burdens I bear, his change in behavior changes the direction of his vector contribution.

15 By "each" I mean each individually, not both of them collectively. For a discussion of similar cases in the context of collective responsibility, see Björnsson, "Collective Responsibility and Collective Obligations without Collective Moral Agents."

16 Goldman, "Why Citizens Should Vote," 210–11, original emphasis.

17 Goldman's vector account is best interpreted as an extension and smoothing out of J. L. Mackie's insufficient but necessary part of an unnecessary but sufficient (INUS) condition for causation in "Causes and Conditions." For instance, an INUS analysis of voting is different for even- and odd-numbered electorates in ways that seem to reflect theoretical machinery rather than the ethics of voting (Goldman, "Why Citizens Should Vote," 206–10). The vector account does not face similar technical complications. Still, if a candidate won in a landslide, why did my vote count as a causal contributor when the outcome we care about is not the scalar "force" of the votes, but the binary of victory and defeat? In my view, something like Mackie's INUS analysis is still needed to answer this question. My vote contributed to the victory because in some subset of votes for the candidate, mine was necessary for her victory. Thus, Goldman's theory is best seen as extending or reformulating Mackie's.

The vector analysis, however, is incomplete as an account of responsibility. Suppose I arrive at the polls intending to vote for the best candidate, but as I put pen to paper, an unforeseeable muscle spasm causes me to vote for the worst candidate, after which a strong gust of wind blows my ballot into the counting machine. My vote is a contributing force in the direction of the bad candidate, but I am clearly not responsible for the contribution. We might similarly wonder about those who lack reliable transportation or who are misinformed about the candidates, anxious in crowded places, forgetful, and so on.[18] What we need is a systematic way to distinguish between those who are responsible for the role they play in dumping burdens and those who are not.[19]

David Brink and Dana Nelkin defend a reasons-responsive account of responsibility on which blameworthiness requires a fair opportunity to avoid wrongdoing.[20] This fair opportunity has three parts: a cognitive component, a volitional component, and a situational component. Briefly, the cognitive component involves the "capacity to make suitable normative discriminations, in particular, to recognize wrongdoing."[21] The volitional component involves "the capacity to regulate one's actions in accordance with this normative knowledge [one's recognition of right and wrong]."[22] Finally, the situational component involves "*external* or *situational* factors . . . [such as] *coercion* and *duress* [which] may lead the agent into wrongdoing in a way that nonetheless provides an excuse, whether full or partial."[23]

In the case of voting, failure to vote (or voting for someone other than the best candidate) is blameworthy when the three conditions listed above are satisfied. Cognitively, this requires, for instance, that information about the candidates' policy stances and qualifications is readily available and intelligible. Volitionally, one must be able to vote according to one's considered convictions rather than peer pressure, a candidate's charisma, or other irrelevancies.

18 See Goldman, "Why Citizens Should Vote," 210, for some discussion of misinformed voters.

19 Because those who vote are not required to vote more when others vote less, failing to vote does not dump burdens. I discuss the relevance of my arguments to voting in sec. 5.

20 See Brink and Nelkin, "Fairness and the Architecture of Responsibility"; and Brink, "Situationism, Responsibility, and Fair Opportunity," especially sec. 4. Modern theories of responsibility fall into two broad categories: reasons-responsive theories and attributionist theories. I employ a reasons-responsive theory of responsibility as one I take to be plausible, though not uncontroversial. For recent defenses of attributionism, see Sher, *Who Knew?*; and Smith, "Control, Responsibility, and Moral Assessment." For an overview of the debate, see Talbert, *Moral Responsibility.*

21 Brink, "Situationism, Responsibility, and Fair Opportunity," 132.

22 Brink, "Situationism, Responsibility, and Fair Opportunity," 132–33.

23 Brink, "Situationism, Responsibility, and Fair Opportunity," 134, original emphasis.

Situationally, voting must not jeopardize one's employment, expose one to undue risks, or be otherwise inaccessible. So long as these conditions are met and so long as one lives in a legitimate democracy, one can be blamed for failing to vote.

A similar analysis can be given for the duty to rescue, though the analysis is complicated by the fact that failing to rescue can cause two distinct wrongs—the deontic wrong to those one is duty bound to rescue and the fairness wrong to others involved in the rescue effort. I will consider the cognitive and volitional components first. Cognitively, duties to rescue are usually easy to understand, and information about rescue efforts is widely distributed and easy to find. Volitionally, fulfilling these duties often requires no more than donating to effective organizations. Similarly, the distributional implications of partial compliance are widely understood. At some point, bearing extra burdens will strain one's volitional capacities, but even if this excuses one from bearing the full weight of one's obligations, one must still work as close as possible to the point of critical volitional stress. Plausibly, most individuals satisfy the cognitive and volitional requirements for responsibility in relation to their duties to rescue and their obligations to distribute the burdens of those rescues fairly. Henceforth, I set aside cognition and volition, and focus on the situational component of responsibility.

Within this reasons-responsive framework, Jones might run the following argument against the claim that she is responsible for dumping burdens. First, she might acknowledge that she is responsible for failing to perform rescues, and that in some mechanical sense, this failure "contributed" a vector force pushing in the direction of burden dumping. Still, she should not be held responsible for those burdens because she did not have a fair opportunity to prevent them. Everyone knew that Smith was not going to do his part, and Jones's contribution alone could not make a difference in the burdens I bear. It is therefore unfair to hold her responsible, even partly, for dumping burdens since she had no opportunity to do otherwise. We might reply by noting that Jones could still have done her part, in which case her vector contribution would have changed and she would no longer count as a burden dumper. But this does not quite capture the spirit of Jones's reply. Her claim is that she could perform rescues, so she had reason to, but she could not lighten my burdens, so she had no reason to. If Jones was faced with no distribution-related reasons, I cannot blame her for taking no distribution-related action. To evaluate this objection, it will be helpful to consider a case in which Jones is unable to act unless Smith acts.

Suppose I have been poisoned and will soon be dead if no antidote is administered. The antidote consists of two ingredients each of which is ineffective if

administered alone. As it happens, Jones and Smith have one ingredient each, and both are present. Unfortunately for me, Smith refuses to give up his ingredient for morally indefensible reasons (he likes the look of its color). Jones, however, rushes to my side to do what she can for me, entreating Smith to do the same. But without Smith, Jones cannot help me, and I die. Intuitively, and I think rightly, Smith is responsible for my death and Jones is not. One possible explanation is that Smith, unlike Jones, had a fair opportunity to make a difference to the outcome. Whatever Jones did, she could not prevent my death. On this line of reasoning, she did not have the relevant situational control, so she cannot be held responsible. If this were right, Jones might argue that the same line of reasoning applies in the partial help case. There too she cannot be held responsible for any "vector contribution" she makes to dumping burdens because she did not have the situational control necessary to prevent extra burdens from falling on me.

Jones's argument that the poison and partial help cases are relevantly similar conflates reasons to change outcomes with reasons to be *willing* to change outcomes. In the poison case, Jones displays concern for my condition and attempts to convince Smith to act. This shows that she is responsive to the available moral reasons. Contrast this with a case in which, for indefensible reasons, neither Smith nor Jones is willing to give up their ingredients. Now, it seems, *both* are responsible for my death even though neither, acting alone, can avert it. Neither displays any willingness to do their part, which is precisely what the situation calls for. If they were appropriately responsive to the available moral reasons, each would show willingness to contribute an antidote ingredient. They would then administer the antidote, and I would be saved. In the partial help case, appropriate responsiveness to the available moral reasons means performing one's share of the rescues. This is something Jones can do even if she cannot reduce the burdens I bear, so she does have the situational control needed to be appropriately reasons responsive. Thus, her failure to perform any rescues marks her as a blameworthy causal contributor in the direction of burden dumping.[24]

24 What if Jones displays willingness to distribute burdens fairly but no independent willingness to rescue the drowning children? Do I still have a fairness complaint against her? I propose that the answer is yes. Jones ought to respond to the full set of moral reasons available to her, and partial responsiveness does not imply partial blameworthiness. Imagine that I love slashing tires, which is both *expensive* and *upsetting* for my victims, and that these are the only relevant moral reasons in the situation. If I were responsive to both reasons, I would not slash tires. But I only care about upsetting people (I am fully responsive to this reason), which alone does not outweigh my enjoyment. It seems to me that when I slash tires, I am blameworthy for upsetting my victims even though I am fully responsive to the moral reasons that their distress gives me to refrain.

Given the preceding arguments, I propose the following characterization of what it means to play a role in dumping burdens: one plays a role in (makes a vector contribution to) unfair burdens borne by contributors when one's failure to be sufficiently reasons responsive in a context of fair opportunity, together with similar failures on the part of others (the number of others may be zero), is sufficient to impose on contributors more than their fair share of the costs to be distributed.[25]

I have argued that when one fails to do one's part in a rescue effort, one treats other rescuers unfairly, at least in nondiscretionary cases. What remains to be seen is whether the arguments I have laid out extend to discretionary duties to rescue. If I only have the resources to contribute to one rescue scheme but there are five equally good schemes to choose from, who is treated unfairly when I do nothing?

3. THE PARTICULARITY PROBLEM

So far, I have focused on rescue scenarios that, by hypothesis, impose a nondiscretionary duty to rescue, which in turn means that any fairness obligations are owed to other rescuers on the scene. I have a nondiscretionary duty to rescue *that child* (or these children), which means that I have a duty to help *these* people perform the rescue. However, some will reject the claim that this duty is nondiscretionary. Singer, for instance, argues that because our duty to rescue does not take distance into account, saving a child right in front of me is morally on a par with saving a child on the other side of the world (other things equal). In the same way that I may choose which drowning children to save when I cannot save them all, I may choose which rescue efforts to participate in when I can only participate in some.

There is controversy over whether the duty to rescue those who are close is more stringent than the duty to rescue those who are far way and whether I have the discretion to contribute to rescue efforts other than the most efficient one.[26] I do not attempt to address these questions here, and I assume for the

25 Ridge, "Fairness and Non-Compliance," offers an alternative solution to collective burden-dumping cases. He argues that the burdens left by noncontributors ought to be shared among rescuers and rescuees alike. Thus, partial help cases will not arise since any additional contribution will reduce my burdens at least marginally. This line of reasoning is mistaken because it incorrectly classifies obligees as obligors. If those to whom obligations are owed are not responsible for bearing a share of those obligations initially, it is unclear why they would become responsible when some obligors fail to contribute.

26 For representative arguments, see Igneski, "Perfect and Imperfect Duties to Aid"; Feinberg, "The Moral and Legal Responsibility of the Bad Samaritan"; Smith, "Control, Responsibility, and Moral Assessment"; and Kamm, "Faminine Ethics."

sake of argument that distance does not matter and that one has at least limited discretion to choose rescue options that are not maximally efficient. That said, questions about discretion arise regardless. I might be equidistant from two drowning children, each of whom has a rescue effort dedicated to her; and, of course, I am confronted with a wide range of organizations to contribute to that carry out rescues all over the world. So, even if there is controversy over the degree to which duties to rescue are discretionary, there is widespread agreement that they allow for at least some discretion.

Duties to rescue being discretionary raises a potential problem for the account of burden dumping I have defended. The claim that by failing to contribute, I dump burdens on other rescuers seems to require a particular rescue effort to which I am bound to contribute—I must have a reason to contribute to *that effort* in particular. If there are no particularizing reasons, then there are no particular burdens I am required to help bear, and thus no answer to the question of who is unfairly burdened when I do nothing. Burden dumping appears incompatible with discretionary duties to rescue. Call this the *particularity problem*.[27]

To flesh out the problem, consider again the case of migrants attempting to cross the Mediterranean, and suppose Italy is doing all it can to rescue vessels in distress. Suppose also that France and Spain have a similar duty to rescue distressed vessels. However, the need is so great that if France or Spain (not both) does all it can, there will be no less for Italy to do. But if both France and Spain helped, each of the three would carry significantly lighter burdens than it would working alone or in conjunction with only one other country. Even so, neither France nor Spain helps, and each rebuts Italy's fairness complaints by saying that it is not imposing burdens on Italy because the other is also unwilling to contribute. This, of course, is just the partial help case. As I argued above, because both France and Spain fail to show proper regard for what is morally important (the migrants' lives), and since these failures are sufficient to impose extra burdens on Italy, each plays a role in dumping the burdens Italy picks up. On these grounds, Italy has a fairness complaint against each country.

27 The particularity problem has parallels in the literature on political obligation. There, the problem applies to theories grounded in the natural duty of justice. Political obligation seems to be owed primarily or exclusively to the institutions that apply to me, while the natural duty of justice seems to allow discretion regarding which institutions I support. For a defense of natural-duty theories, see Waldron, "Special Ties and Natural Duties." For a statement of the particularity problem, see Simmons, "The Natural Duty of Justice"; and for a reply to Waldron, see Simmons, "Natural Duties and the Duty to Obey the Law," 170–79.

At this point, the discretionary nature of duties to rescue leaves open a further possible response for France and Spain. Suppose everyone knows that if Spain contributes to any rescue effort, it will be to one run by Bulgaria, not Italy. Thus, even if both France and Spain contribute their fair share to rescue efforts, Italy's burdens will not be lightened. In this way, France argues that it does not treat Italy unfairly because France's and Spain's failures to respond appropriately to the relevant reasons do not lead them to withhold contributions that would be sufficient to reduce Italy's burdens.[28] Since Spain has the discretionary freedom to choose which rescue effort to contribute to, neither it nor France counts as contributing to Italy's burdens. If this is the end of the story, there are simply more burdens to go around than can be borne. Full compliance with the duty to rescue would require maximum sacrifice from everyone required to make any sacrifice. In that case, France's argument goes through, and neither it nor Spain dump any burdens on Italy. However, assuming that full compliance will not require maximal sacrifices from everyone, Italy's argument can take a further step to match the step taken by France's argument.

So far, I have presented the case as though Spain and France are the only (relevant) actors not doing their part. But this is an artifact of thinking of duties as nondiscretionary. Now that we are thinking of a discretionary duty to rescue, the relevant pool of burdens is all the burdens associated with all the rescues that need to be performed and that require collective action.[29] Given this, the pool of potential contributors includes every agent—natural or artificial—who is bearing less than her fair share of the overall burdens. If we now imagine that no one fails to do her part through a blameworthy failure to be reasons responsive, we will imagine a scenario in which all these agents bear their fair share of the total pool of rescue-related burdens. If we have good reason to think that Italy's burdens would be reduced in *this* situation, then Italy has a fairness complaint not only against France and Spain, but against everyone who is not contributing her fair share to rescue efforts around the world. This is the partial help argument writ large.

So, even though it seems right to say that the duty to rescue allows for significant discretion on the part of those bound by it, those already rescuing almost certainly have legitimate fairness complaints against most noncontributors. Because they fail to be appropriately reasons responsive, they play a role

28 See Feinberg, "The Moral and Legal Responsibility of the Bad Samaritan," 60–64, for a discussion of similar cases in the context of imperfect duties (duties that lack a prescribed time or place of fulfillment; these are precisely discretionary duties in my sense).

29 See Ostrom, "Beyond Markets and States," for a detailed analysis of collective action problems and the contexts in which they often arise.

in (make a vector contribution to) the unfair imposition of burdens by failing to bear their portion of the total pool of rescue-related burdens.

4. Omission and Collective Action

One might worry that the partial help argument has been writ too large. Consider the following case. Italy is rescuing migrants crossing the Mediterranean, Bulgaria is rescuing migrants crossing the land border from Turkey, and both efforts are on a par in all relevant respects. France, however, is rescuing no one. Italy knows that if everyone were doing her fair share of rescues, its burdens would be lighter than they currently are. Unfortunately, Italy can only influence France. Thus, Italy begins making fairness complaints against France, and France, exercising its discretion, begins contributing to Bulgaria's scheme. Nothing has changed for Italy, but since France is now doing its part, Italy no longer has a fairness complaint against it. This seems odd. Italy claims to be treated unfairly by France because France plays a role in Italy's excessively heavy burdens. Yet France successfully satisfies its fairness obligation to Italy without reducing Italy's burdens. One might take this to show that the partial help argument is not ultimately concerned with fairness. If it were, it would argue that Italy's claim against France removes France's discretion so that it must contribute to Italy's rescue effort.

This line of objection can be interpreted as asserting one of two underlying thoughts. First, the objection might be another way of claiming that *X* treats *Y* unfairly by failing to contribute just in case *X*'s contribution alone would be sufficient to reduce *Y*'s burdens. My main argument up to this point has been aimed at rejecting this intuition, so I will set this interpretation aside. Alternatively, one could take the objection as expressing something like the following: if *X* treats *Y* unfairly by not contributing to any rescue effort, then it is also the case that *X* treats *Y* unfairly by contributing to any rescue effort other than *Y*'s. So, in the EU example, since France could lighten Italy's burdens, France treats Italy unfairly when it contributes to Bulgaria's rescue scheme instead of Italy's.

To see where this second suggestion leads, suppose for the sake of argument that France imposes burdens on Italy when it performs no rescues *and* when it performs its fair share of rescues in Bulgaria's rescue effort. Granting this, it might seem to follow automatically that France treats Italy unfairly by contributing to Bulgaria's scheme.[30] But how can this be? Recall that France's duty to

30 If this were right (and the ought-implies-can principle were true), it would be a serious problem for my view. If, for instance, Italy and Bulgaria announced fairness complaints against France at the same time on the same day, France would be forced to treat one of them unfairly (assuming it can only feasibly contribute to one scheme).

rescue is supposed to be discretionary, and it seems clearly right to say that before Italy makes its complaint against France, France is free to contribute to either scheme. So, what changes when Italy makes its claim? Sarah McGrath gives us a potential answer in her theory of causation by omission. She argues that omission "*o* causes [event] *e* iff *o* occurs, *e* occurs, and [commission of the act of which *o* is an omission] C_o is a normal would-be preventer of *e*."[31] A would-be preventer of *e* is something that would prevent *e* if it occurred. A would-be preventer is normal if it is *supposed* to prevent *e* according to some *actual standard*.[32] The thought is that Italy's act of making an unfairness claim against France establishes a standard according to which France is supposed to help Italy and that this standard dissolves France's discretion about which rescue efforts it may contribute to.

This proposal fails for several reasons. For one, it is not enough to simply establish a standard; the standard that is established must be shown to be important. McGrath's notion of a standard is "of very general application," covering "chess moves, dance steps, quiz answers, beliefs, baseball pitches, ways of beating eggs and stitching hemlines."[33] Each involves a standard of correctness that can be used to judge good and bad chess moves, dance steps, and so on. In that sense, all the standards are normative. However, they do not all have moral force. In fact, morality can be conceived as another standard according to which actions can be judged to be appropriate or not. Since, according to the duty to rescue, France has moral discretion to contribute to whatever rescue effort it chooses, the standard established by Italy's complaint will be ineffectual unless it can be shown to have overriding moral significance. Since the mere statement of the complaint does nothing to change the facts of the situation, it is unclear where this significance could come from.

Even if this difficulty could be overcome, problems still arise. Suppose France can only contribute to one scheme and Italy and Bulgaria make simultaneous fairness complaints against it, each demanding that France contribute to their rescue effort. To whose scheme should it contribute? The most natural take on the situation is that France is free to choose which scheme to contribute to. In this case, its discretion persists. The only apparent alternative is to say that even when France entirely fulfills its duty to rescue, by helping Italy for instance, it is still guilty of unfairly dumping burdens on Bulgaria. Surely this

31 McGrath, "Causation by Omission," 142. McGrath offers a more precise formulation of the same principle, but this will do for my purposes here.

32 McGrath, "Causation by Omission," 138.

33 McGrath, "Causation by Omission," 139.

is a principle that *should* be rejected for imposing unfair burdens, though here the unfair burdens are placed on burden dumpers rather than slack takers.[34]

Additionally, it is not clear why Italy's articulation of its fairness complaint should *create* a standard for France. Italy's speech act appears descriptive, not performative. It reports reasons to which France ought to respond; it does not create them. Thus, the standard according to which France treats Italy unfairly unless it contributes to Italy's rescue efforts applies whether Italy makes a declaration or not. But then, since Bulgaria is in the same position as Italy relative to France, it too must have an identical claim to France's contribution. So, France will have just as much reason to contribute to Bulgaria's scheme as to Italy's whether or not Bulgaria or Italy or anyone else makes a fairness complaint against it. France once again finds itself unfairly bound to shoulder more burdens than it can bear.

The initial objection was that something has gone wrong with the partial help argument since Italy's unfairness complaint against France, grounded in its unfairly heavy burdens, does not obligate France to contribute to Italy's rescue effort. Intuitively, we might think that if France treats Italy unfairly, it ought to contribute to Italy's scheme. But this intuition is misguided because its focus is too narrow. France is not the only noncontributor, and Italy's scheme is not the only one around. Still, one might try to vindicate the intuition by arguing that once Italy makes its claim on France, France counts as causing Italy's extra burdens by omission. As we have seen, however, this argument does not look promising.

5. Further Applications

In this paper, I have argued that when we fail to contribute our fair share in a rescue effort and others must take up the slack, we treat those others unfairly. Problems we have a duty to solve and that require collective action to address are subject to distributive norms that generate fairness obligations between rescuers in addition to the natural-duty obligations owed to those in need of rescue. The central objection to which I respond argues that one only dumps

34 One might attempt to run a similar omissions argument by appealing to a Lewisian view on which o causes e iff C_o would have prevented e (see Lewis, "Causation as Influence"). By this standard, every agent in the world whose contribution to Italy's scheme would reduce its burdens, if it so contributed, counts as causing Italy's burdens by omission. But if this is right, we clearly have not landed on a normatively significant sense of omission. Suppose Bulgaria begins its scheme before Italy. The fact that Bulgaria causes by omission Italy's excessive burdens clearly does not mean that Bulgaria treats Italy unfairly or that Bulgaria ought to terminate its own scheme to contribute to Italy's.

burdens when one's contribution alone would be sufficient to lighten the burdens of current contributors. I argue that this claim is mistaken. By failing to do one's part in the absence of excusing conditions, one fails to be appropriately reasons responsive. This failure makes one a blameworthy member of the vector group whose actions or omissions push in the direction of burden dumping. Thus, those who fail to do their part are implicated in the resulting unfair distribution of burdens.

This argument appears problematic in the context of discretionary duties. When I am not obligated to contribute to any particular rescue effort, it is not clear who is treated unfairly when I do less than my fair share. I argue that this worry can be dispelled by broadening the scope of the argument. The argument I develop in response to the partial help case shows that one can play a role in the unfair burdens borne by individuals performing rescues even if one's contributions alone would not reduce their burdens. Thus, no matter how much discretion I have in fulfilling my duty to rescue, I unfairly dump burdens on those who do their part when I fail to do mine.

It is worth considering how the arguments presented here apply in other contexts. Very briefly, I discuss voting and climate change. In the context of burden dumping, voting and rescuing are fundamentally different because one cannot dump one's duty to vote on others. I should not vote twice in an election because someone else did not vote at all. This does not mean, however, that the arguments I have developed are inapplicable.

In its most general terms, the view I defend identifies responsible causal contributors to the outcomes of collective actions or omissions. This was the payoff of combining Goldman's vector theory of causation with Brink and Nelkin's theory of moral responsibility. For any case in which we can identify the reasons to which individuals ought to respond, we can, in principle, identify those who are blameworthy (or praiseworthy) for the outcomes of their actions or omissions. In the case of voting, bad outcomes of elections or referendums can be very destructive even though no burden dumping is involved. The account I have defended allows us to identify those who are blameworthy for pushing toward these negative outcomes even when, for instance, the better candidate wins. Burden dumping can therefore be seen as a special case focusing on situations in which partial compliance affects the distribution of burdens. Many collective action problems are plagued by partial compliance, and in these cases, it is worth understanding how to assign blame and responsibility for unfair distributions.

Climate change is structurally much closer than voting to rescue cases and so raises similar distributive questions. Our responses to climate change, whether in the form of mitigation (preventing future climate change),

adaptation (responding to unavoidable change), or compensation (to those unjustly affected), require collective action and allow burden dumping. Climate change raises additional complex questions about the initial fair distribution of burdens, intergenerational justice, cosmopolitanism versus nationalism, and so on.[35] However we answer these questions, the analysis presented here can help us respond appropriately to actors who fail to do their part.

Some duties are quite stringent, and this stringency can obscure distributive concerns. In rescue cases, for example, complaining that I must do more than my fair share of rescues when those being rescued are in dire need might seem melodramatic. It is worth remembering, however, that burden dumping can impose very heavy burdens, especially when the duties involved are stringent. Additionally, those on whom the burdens fall may be better positioned to hold accountable those who refuse to do their part. This last point is especially relevant in the case of climate change.

There is widespread agreement that individuals acting independently cannot respond adequately to climate change.[36] Individuals, corporations, governments, and supranational organizations must act in concert if we are to minimize the damage of climate change to human wellbeing. But many actors are, and have been, unwilling to do their part; the US government, for instance, has consistently failed to pursue meaningful emissions-reduction policies. [37] While people outside the United States often feel the effects of these failures most strongly, it is US citizens that can act most effectively to change the trajectory of US policy. One strategy for pressuring the government is to voice complaints that the failure of the government (and others) to adequately respond to climate change has imposed unfair burdens on individual members of the population, requiring them to unilaterally reduce their private emissions or attempt to organize their own emissions-reduction schemes. Even the need to voice complaints is an avoidable and unfair burden. If we focus exclusively on the harmful effects of climate change, these grievances will go unnoticed. Not only does this let blameworthy actors partially off the hook, but it also

35 For discussion of who should pay for the costs of mitigation, adaptation, and compensation, see Caney, "Cosmopolitan Justice, Responsibility, and Global Climate Change." For discussions of intergenerational ethics related to climate change, see Gardiner, "A Perfect Moral Storm"; and Gosseries, "Historical Emissions and Free-Riding."

36 Whether individuals acting independently can adequately respond to climate change is separate from the question of whether individuals have a duty to reduce their own emissions when others fail to act. For discussion, see Sinnott-Armstrong, "It's Not My Fault"; Schwenkenbecher, "Is There an Obligation to Reduce One's Individual Carbon Footprint?"; and Hourdequin, "Climate, Collective Action and Endividual Ethical Obligations."

37 See Jamieson, *Reason in a Dark Time*, for an overview of the history of climate change.

robs those seeking change of a potentially important means of pressuring those who neglect their duties.

I have here only scratched the surface of the various ways in which the ethics of slack taking and burden dumping might be applied. My hope is that this discussion will help promote further applications by illuminating not just the structure of our duties to rescue but a more general relationship between natural duties and fairness obligations.

University of California, San Diego
afinley@ucsd.edu

REFERENCES

Agnafors, Marcus. "On Disjunctive Rights." *Southern Journal of Philosophy* 55, no. 2 (June 2017): 141–57.

Baron, Marcia. "Kantian Ethics and Supererogation." *Journal of Philosophy* 84, no. 5 (May 1987): 237–62.

Björnsson, Gunnar. "Collective Responsibility and Collective Obligations without Collective Moral Agents." In *The Routledge Handbook of Collective Responsibility*, edited by Saba Bazargan-Forward and Deborah Tollefsen, 127–41. New York: Routledge, 2020.

Brink, David O. "Situationism, Responsibility, and Fair Opportunity." *Social Philosophy and Policy* 30, nos. 1–2 (January 2013): 121–49.

Brink, David O., and Dana K. Nelkin. "Fairness and the Architecture of Responsibility." In *Oxford Studies in Agency and Responsibility*, vol. 1, edited by David Shoemaker, 284–313. Oxford: Oxford University Press, 2013.

Caney, Simon. "Cosmopolitan Justice, Responsibility, and Global Climate Change." *Leiden Journal of International Law* 18, no. 4 (December 2005): 747–75.

Cohen, L. Jonathan. "Who Is Starving Whom?" *Theoria* 47, no. 2 (August 1981): 65–81.

Feinberg, Joel. "The Moral and Legal Responsibility of the Bad Samaritan." *Criminal Justice Ethics* 3, no. 1 (1984): 56–69.

Gardiner, Stephen M. "A Perfect Moral Storm: Climate Change, Intergenerational Ethics and the Problem of Moral Corruption." *Environmental Values* 15, no. 3 (August 2006): 397–413.

Goldman, Alvin I. "Why Citizens Should Vote: A Causal Responsibility Approach." *Social Philosophy and Policy* 16, no. 2 (Summer 1999): 201–17.

Gosseries, Axel. "Historical Emissions and Free-Riding." *Ethical Perspectives* 11,

no. 1 (April 2004): 36–60.
Hope, Simon. "Kantian Imperfect Duties and Debates over Human Rights." *Journal of Political Philosophy* 22, no. 4 (December 2014): 396–415.
Horton, Keith. "Fairness and Fair Shares." *Utilitas* 23, no. 1 (March 2011): 88–93.
———. "International Aid: The Fair Shares Factor." *Social Theory and Practice* 30, no. 2 (April 2004): 161–74.
Hourdequin, Marion. "Climate, Collective Action and Individual Ethical Obligations." *Environmental Values* 19, no. 4 (November 2010): 443–64.
Igneski, Violetta. "Perfect and Imperfect Duties to Aid." *Social Theory and Practice* 32, no. 3 (July 2006): 439–66.
Jamieson, Dale. *Reason in a Dark Time: Why the Struggle against Climate Change Failed—And What It Means for Our Future.* Oxford: Oxford University Press, 2014.
Kamm, F. M. "Faminine Ethics: The Problem of Distance in Morality and Singer's Ethical Theory." In *Singer and His Critics*, edited by Dale Jamieson, 162–208. Oxford: Blackwell, 1999.
Karnein, Anja. "Putting Fairness in Its Place: Why There Is a Duty to Take Up the Slack." *Journal of Philosophy* 111, no. 11 (November 2014): 593–607.
Lewis, David. "Causation as Influence." *Journal of Philosophy* 97, no. 4 (April 2000): 182–97.
Mackie, John L. "Causes and Conditions." *American Philosophical Quarterly* 2, no. 4 (October 1965): 245–64.
McGrath, Sarah. "Causation by Omission: A Dilemma." *Philosophical Studies* 123, nos. 1–2 (March 2005): 125–48.
Miller, David. "Taking Up the Slack? Responsibility and Justice in Situations of Partial Compliance." In *Justice for Earthlings: Essays in Political Philosophy*, 206–27. Cambridge: Cambridge University Press, 2013.
Murphy, Liam B. "The Demands of Beneficence." *Philosophy and Public Affairs* 22, no. 4 (Autumn 1993): 267–92.
———. *Moral Demands in Nonideal Theory.* New York: Oxford University Press, 2000.
Ostrom, Elinor. "Beyond Markets and States: Polycentric Governance of Complex Economic Systems." *American Economic Review* 100, no. 3 (June 2010): 641–72.
Pogge, Thomas. "Are We Violating the Human Rights of the World's Poor?" *Yale Human Rights and Development Journal* 14, no. 2 (2011): 1–33.
Ridge, Michael. "Fairness and Non-Compliance." In *Partiality and Impartiality: Morality, Special Relationships, and the Wider World*, edited by Brian Feltham and John Cottingham, 194–222. Oxford: Oxford University Press, 2010.
Schwenkenbecher, Anne. "Is There an Obligation to Reduce One's Individual

Carbon Footprint?" *Critical Review of International Social and Political Philosophy* 17, no. 2 (2014): 168–88.

Sher, George. *Who Knew? Responsibility without Awareness.* New Yord: Oxford University Press, 2009.

Simmons, A. John. "Natural Duties and the Duty to Obey the Law." In *Is There a Duty to Obey the Law?* by Christopher Heath Wellman and A. John Simmons, 121–88. Cambridge: Cambridge University Press, 2005.

———. "The Natural Duty of Justice." In *Moral Principles and Political Obligations*, 143–56. Princeton, NJ: Princeton University Press, 1979.

Singer, Peter. "Famine, Affluence and Morality." In *Ethical Theory: An Anthology*, edited by Russ Shafer-Landau, 505–12. Malden, MA: Blackwell, 2007.

Sinnott-Armstrong, Walter. "It's Not *My* Fault: Global Warming and Individual Moral Obligations." In *Perspectives on Climate Change: Science, Economics, Politics, Ethics*, edited by Walter Sinnott-Armstrong and Richard Howarth, 285–307. Bingley: Emerald, 2005.

Smith, Angela M. "Control, Responsibility, and Moral Assessment." *Philosophical Studies* 138, no. 3 (April 2008): 367–92.

Smith, Patricia. *Liberalism and Affirmative Obligation.* Oxford: Oxford University Press, 1998.

Talbert, Matthew. *Moral Responsibility: An Introduction.* Malden, MA: Polity Press, 2016.

Unger, Peter. *Living High and Letting Die: Our Illusions of Innocence.* New York: Oxford University Press, 1996.

Waldron, Jeremy. "Special Ties and Natural Duties." *Philosophy and Public Affairs* 22, no. 1 (Winter 1993): 3–30.

Wolterstorff, Nicholas. *Justice: Rights and Wrongs.* Princeton, NJ: Princeton University Press, 2008.

Journal of Ethics and Social Philosophy
VOL. 23, NO. 3 · JANUARY 2023
https://doi.org/10.26556/jesp.v23i3.1817

IS MORALITY OPEN TO THE FREE WILL SKEPTIC?

Stephen Morris

In contemporary discussions about free will, philosophers from the various camps (compatibilists, libertarians, and skeptics) have come to something approaching a consensus insofar as they tend to agree that free will is best understood as being a necessary condition—or, more specifically, the control condition—for moral responsibility.[1] From this it follows that if human beings lack free will, we must also lack moral responsibility in some sense. The question as to exactly what kind of moral responsibility is precluded by the absence of free will is one I will return to shortly. Given the standard view that free will skepticism implies a kind of skepticism about moral responsibility, one might naturally think that the free will skeptic is committed to denying the truth of moral claims in any form. After all, if a person cannot be held morally responsible for his actions, it seems reasonable to think that he cannot be said either to have acted in a way that was morally wrong (or right), or that he is morally obligated to act/not act in a particular fashion. This was the view of C. A. Campbell, who claimed that if we cannot possess the kind of moral responsibility that requires free will we are thereby forced to give up "the reality of the moral life."[2] And yet in perusing the free will skeptic literature, one finds no shortage of moral language by its proponents, including a variety of moral exhortations and prohibitions. Freewill skeptic Gregg Caruso speaks to the tendency of free will skeptics (henceforth simply *skeptics*) to make moral claims by noting how the view "that moral responsibility is a necessary condition for morality … is directly challenged by most skeptics."[3] Caruso himself

1 Compatibilists hold free will to be compatible with the truth of causal determinism. Incompatibilists believe that free will is not possible if causal determinism is true. Libertarians are incompatibilists who maintain that human beings are capable of exercising free will. Freewill skeptics are incompatibilists who deny the possibility that human beings can exercise free will.

2 Campbell, *On Selfhood and Godhood*, 166–67. Other notable philosophers sharing this view include Wolf, "The Importance of Free Will"; and Van Inwagen, *An Essay on Free Will*.

3 Caruso, "Skepticism about Moral Responsibility."

has made the case that human agents and their institutions are open to moral assessment despite people lacking free will and an important kind of moral responsibility that requires free will. For instance, he frequently employs moral language when discussing the proper basis for criminal punishment. It is in this context that we find him saying things such as, "From the skeptical perspective … [we] need, therefore, to confront the *moral challenge* of balancing the individual liberties with the advancement of the public good."[4] A central point in his overall argument is that while the public health-quarantine model of treating criminals that he favors is morally justifiable, retributivist models of punishment are not. Ted Honderich is another prominent free will skeptic who has argued that moral terms are applicable to human agents and their actions even if no human actions are free. He states that "each of us has a moral standing. There are corollaries having to do with right action, and good men and women."[5] Derk Pereboom, whom I discuss in greater detail later in this essay, is another example of skeptics who reject moral responsibility while holding that human agents and actions can be proper subjects of moral appraisal.

In this essay I consider whether skeptics' assertions of moral claims pertaining to human agents and their activities are consistent with their rejection of free will. In casting doubt on the prospect of constructing a compelling skeptical defense of morality, my project can be seen as following a line of thought that stretches at least as far back to Immanuel Kant, who argued that morality only applies to beings capable of exercising free will.[6] Part of what seemed to be motivating Kant to view free will and morality as inherently connected was the intuition that, unless a person acted freely, it would be unreasonable to hold them morally culpable in such a way that could merit punishment. Hence, we find Kant stating that moral accountability

> could not happen if we did not suppose that whatever arises from one's choice (as every action intentionally performed undoubtedly does) has as its basis a free causality which from early youth expresses its character in its appearances (actions); these actions, on account of the uniformity of conduct, make knowable a natural connection that does not, however, make the vicious constitution of the will necessary but is instead the consequence of the evil and unchangeable principles freely adopted, which make it only more culpable and deserving of punishment.[7]

4 Caruso, "Free Will Skepticism and Criminal Behavior," 33, emphasis added.

5 Honderich, *A Theory of Determinism*, 172.

6 Some of Kant's primary arguments for why morality requires free will can be found in *Groundwork for the Metaphysics of Morals*, 47–54, and *Critique of Practical Reason*, 122–29, 189–94.

7 Kant, *Critique of Practical Reason*, 194.

This connection between free will, morality, and deservingness of punishment for Kant is notable in that, as I discuss in more detail below, it resonates with how ordinary persons (i.e., the folk) tend to conceive of morality. This, in turn, is relevant to the aims of this paper since I will demonstrate how skeptics rely heavily upon folk intuitions to motivate their position. I will argue that one of the main obstacles confronting skeptical accounts of morality is that the folk—echoing the intuitions of Kant—embrace an account of morality that is at odds with that of the skeptic insofar as it justifies certain kinds of punishment that the skeptic believes can never be warranted for beings lacking free will. This being the case, this essay can be viewed as providing a modern defense of Kant's view that morality requires free will.

In order to properly address the issue of whether entities lacking free will are open to moral assessment, it will be necessary to see what some of the prominent skeptics have to say with regard to both what the denial of free will implies, and the kinds of moral claims that they take to be in line with their metaphysical commitments. Getting clearer about the kinds of moral claims that the skeptic deems to be consistent with her position will provide us with a better understanding of how the skeptic's view differs from that of her opponents (compatibilists and libertarians). In doing so, we will get a better understanding of what is at stake in the free will debate. I begin by considering the kinds of moral claims that virtually all skeptics hold to be inapplicable to agents lacking free will. From there I discuss the more contentious point of whether it is ever appropriate to attribute moral obligations to such agents or—what amounts to more or less the same thing—whether they can ever be proper targets of moral "ought" statements. Following an examination of folk moral judgments, I discuss what is possibly the most well-known and detailed skeptical defense of morality in the free will literature—namely, that provided by Pereboom in his book *Living without Free Will*. Although I argue that Pereboom's defense of morality is unsuccessful, it is nonetheless instructive insofar as what he says in the process of building his case for morality (in combination with his criticisms of revisionist defenses of free will) points to ways that skeptics might attempt to defend morality that are best avoided. In particular, Pereboom's discussions of free will and morality highlight why it would be problematic for skeptics to reject free will while embracing a sort of morality that is founded on a purely forward-looking notion of moral responsibility. Through this analysis I aim to show that any skeptical defense of morality is likely to be impeded by the skeptics' reliance upon folk intuitions to motivate their skepticism. As I discuss below, skeptics like Pereboom have insisted on the importance of preserving folk concepts when it comes to navigating the philosophical debates about free will and moral responsibility. In fact, it is this alleged necessity of retaining

folk concepts that has served as the main tactic by which skeptics have tried to thwart the efforts of free will revisionists like Manuel Vargas. And yet many of these same skeptics (e.g., Pereboom) have employed a similar revisionist approach when trying to defend concepts like "moral rightness" and "moral wrongness." I will argue that there does not appear to be any justifiable reason for approving of the revision of terms like "moral rightness" while disallowing any similar revision with regard to terms such as "free will" or "moral responsibility." In light of this, I conclude that the skeptic must choose one of two paths. On the one hand, they can approve of revising key terms in the free will debate in a way that differs from how the folk understand them. In doing so, however, they would appear to undermine the main justification they have offered for why skepticism is preferable to opposing views on free will, such as Vargas-style revisionism. On the other hand, they could hold fast to the necessity of using key terms in the free will debate that do not veer far from the folk conception of them. Opting for this route, however, would seem to commit them to dispensing with moral language altogether.

1. What are free will skeptics committed to?

1.1. The Rejection of Backward-Looking Moral Responsibility, Basic Desert, and Praise and Blame

Given that the issue of free will has been argued over endlessly for over two millennia with seemingly little progress being made in providing a definitive answer to the fundamental question—"Do human beings have free will?"—one would be forgiven for suggesting that any further discussion of the matter would be pointless. While such a sentiment is understandable, it fails to recognize the significant progress that has been made in the past few decades with respect to clarifying both the key concepts in question and the primary points of contention that are driving the debate. Speaking to the former, I mentioned earlier how philosophers have converged on an understanding of free will as being the control condition for moral responsibility. While there are some philosophers who reject the understanding of free will in terms of its connection to moral responsibility (e.g., Bruce Waller), it nonetheless remains the accepted view among the vast majority of the more prominent participants in contemporary debates involving free will.[8] Speaking to the central role that moral

8 See, for example, Pereboom, *Living without Free Will* and *Free Will, Agency, and Meaning in Life*; Strawson, *Freedom and Belief* and "The Impossibility of Moral Responsibility"; O'Connor, *Persons and Causes*; McKenna, "Ultimacy and Sweet Jane"; Nielsen, "The Compatibility of Freedom and Determinism"; Campbell, *On Selfhood and Godhood*;

responsibility has played in driving discussions surrounding free will, Galen Strawson maintains that it "is a matter of historical fact that concern about moral responsibility has been the main motor—indeed the *ratio essendi*—of discussion of the issue of free will."[9]

Given that free will is generally understood as being the control condition for moral responsibility, it is not surprising that virtually every self-identified free will skeptic currently writing on the subject of free will has denied the existence of moral responsibility for human beings in some form or another. This is apparent among proponents of a popular branch of skepticism (the view that Pereboom calls "hard incompatibilism"), according to which free will is most likely impossible whether or not causal determinism is true. Capturing the view of many skeptics, Galen Strawson has stated that "it makes no difference whether determinism is true or not. We cannot be truly or ultimately morally responsible for our actions in either case."[10] Pereboom himself contends that hard incompatibilism lends itself to skepticism toward a kind of deep and ultimate type of moral responsibility. Other skeptics who deny that this sort of moral responsibility is available to human agents include Richard Double, Gregg Caruso, Thomas Nadelhoffer, and Neil Levy.[11]

In order to better understand the skeptic's position, it is necessary to consider the kinds of moral claims that they believe can and cannot legitimately be made given that human agents lack free will. One moral concept that virtually all self-identified skeptics believe is inapplicable to human agents is *basic desert.* In fact, when skeptics claim that a lack of free will precludes human beings from being morally responsible, it is typically the basic desert sense of moral responsibility that they have in mind. Pereboom offers a summary of what the basic desert sense of moral responsibility is in this passage:

> For an agent to be morally responsible for an action is for it to belong to her in such a way that she would deserve blame if she understood that it was morally wrong, and she would deserve credit or perhaps praise if

Clarke, "An Argument for the Impossibility of Moral Responsibility"; Levy, *Hard Luck*; Van Inwagen, *An Essay on Free Will*; Vargas, "Desert, Responsibility, and Justification"; Nahmias, "Response to Misirlisoy and Haggard and to Bjornsson and Pereboom"; Caruso, *Free Will and Consciousness*; and Nadelhoffer, "The Threat of Shrinking Agency and Free Will Disillusionism." In what follows I simply assume that the majority of philosophers investigating free will are correct in holding free will to be the control condition for moral responsibility.

9 Strawson, "The Impossibility of Moral Responsibility," 8.

10 Strawson, "The Impossibility of Moral Responsibility," 5.

11 Double, *The Non-Reality of Free Will*; Caruso, *Free Will and Consciousness*; Nadelhoffer, "The Threat of Shrinking Agency and Free Will Disillusionism"; and Levy, *Hard Luck.*

> she understood that it was morally exemplary. The desert sense at issue here is basic in the sense that the agent, to be morally responsible, would deserve blame or credit just because she has performed the action … and not by virtue of consequentialist considerations.[12]

Speaking to the central role that the rejection of basic desert moral responsibility plays in motivating the skeptic's view, Caruso says:

> What all these skeptical [hard incompatibilist] arguments have in common, and what they share with classic hard determinism, is the belief that what we do, and the way we are, is ultimately the result of factors beyond our control and because of this we are never morally responsible for our actions in the *basic desert* sense—the sense that would make us *truly deserving* of blame or praise.[13]

To get a better idea of what philosophers mean by "desert" or the "basic desert sense" of moral responsibility, it may help to consider what Strawson has said about the kind of moral responsibility that the free will skeptic rejects and that many if not most of us consider ourselves to have. Referring to this type of moral responsibility as "true moral responsibility," he has described it as "responsibility of such a kind that, if we have it, then it *makes sense*, at least, to suppose that it could be just to punish some of us with (eternal) torment in hell and reward others with (eternal) bliss in heaven."[14] In response to this suggestion, some philosophers have argued that the excessively retributivist notions of eternal suffering or eternal bliss at work here cannot accurately capture the more modest desert element seemingly at work in the folk understanding of moral responsibility. While this may be true, the idea of divine retribution in the afterlife (perhaps in a form more limited than the kind of eternal retribution captured by traditional concepts of heaven and hell) seems a plausible way of understanding the folk notion of desert that plays an important role in many people's notions about moral responsibility.[15] *Retributivism* refers roughly to the justification for treatment whereby an individual is either rewarded or punished as payback for the moral rights/wrongs he has committed. Consequen-

12 Pereboom, "Hard Incompatibilism," 86.

13 Caruso, "Free Will Skepticism and Criminal Behavior," 26.

14 Strawson, "The Impossibility of Moral Responsibility," 9.

15 Pereboom also suggests that the basic desert sense of moral responsibility is closely connected with retributive attitudes. As he puts it, of all of the justifications for punishment, retributivism "is the one that most intimately invokes the basic desert sense of moral responsibility, together with the freedom it entails" ("Free Will Skepticism and Criminal Punishment," 52).

tialist considerations do not figure into justifications for treatment from this perspective.[16]

It is worth noting here that the basic desert sense of moral responsibility that is rejected by virtually all skeptics is not the only type of moral responsibility on the table. Whereas this kind of responsibility is completely *backward looking* in that its focus is on the type of responses that are warranted strictly by an agent's past decisions or behaviors, skeptics have been much more willing to embrace a kind of moral responsibility that is considered to be purely *forward looking* in nature, such that certain reactions to agents—such as judgments, rewards, and punishments—can only be justified on consequentialist foundations such as future protection, future reconciliation, or future moral formation.[17] While I have restricted my discussion in this section to how almost all skeptics reject moral responsibility in the backward-looking sense, in section 3 I will consider whether skeptics could succeed in defending a type of morality founded upon a forward-looking sort of moral responsibility.

In the earlier statement in which Caruso discusses how modern skepticism is characterized by its rejection of basic desert moral responsibility, he mentions how this implies that people are never truly deserving of blame or praise. This idea that human agents are never genuinely praiseworthy or blameworthy is another defining feature of skepticism. Speaking to this point, Pereboom says:

> The feature of our ordinary conception of ourselves that would most obviously be undermined if hard incompatibilism were true is our belief that people are typically praiseworthy when they perform morally exemplary actions, and they are typically blameworthy when they perform actions that are morally wrong. To be blameworthy is to deserve blame just because one has chosen to do wrong. Hard incompatibilism rules out one's ever deserving blame just for choosing to act wrongly.[18]

Other prominent skeptics who reject the applicability of attributions of genuine blame or praise to human agents include Nadelhoffer, Strawson, and Levy.[19]

1.2. Moral "Ought" Statements and Moral Obligations

While there is a general consensus among skeptics that neither backward-looking moral responsibility, nor basic desert, nor backward-looking praise or blame

16 For further analysis of the role that basic desert plays in contemporary debates about free will, see McKenna, "Basically Deserved Blame and Its Value."

17 See Pereboom, *Free Will, Agency, and Meaning in Life.*

18 Pereboom, *Living without Free Will*, 139–40.

19 Nadelhoffer, "The Threat of Shrinking Agency and Free Will Disillusionism"; Strawson, *Freedom and Belief* and "The Impossibility of Moral Responsibility"; and Levy, *Hard Luck.*

are attributable to us—I will refer to this consensus as the "core moral denials of free will skepticism"—there is less certainty with regard to whether moral "ought" statements or moral obligations are applicable to human agents. It should be stated from the onset that, with regard to current discussions of free will in the philosophical literature, the concepts of *ought* and *obligation* tend to be inextricably linked. This is to say that participants in these discussions tend to use these concepts interchangeably such that the statement, "Agent *A* *morally ought to/ought not* do action *X*" is generally taken to be equivalent to "Agent *A* is *morally obligated to/obligated not to* do *X*," and vice versa. In light of this, the discussion that follows assumes that where a particular moral "ought" statement is applicable to a specific human agent, it must be the case that the agent is subject to an equivalent moral obligation.

With this in mind, let us consider whether free will skepticism is consistent with moral ought statements of the form, "You *ought to/ought not* do action *A*," where the failure to act in accordance with such a statement is taken to ground either the moral rightness or moral wrongness of one's action (or failure to act). In terms of why one might think that such statements are never warranted given the truth of free will skepticism, it helps to recognize that virtually all skeptics are at least open to the possibility that causal determinism is true. And while most skeptics are hesitant to commit themselves to determinism given what contemporary physics tells us about quantum probabilities, they virtually all recognize that determinism is a live possibility with some even going so far as to claim that determinism is, for all intents and purposes, true. As Al Mele acknowledges, even if we accept that quantum mechanics is correct, this does not "ensure that any human brains themselves operate indeterministically," nor does it rule out that "any indeterminism in the human brain is simply irrelevant to the production of actions."[20] Neuroscientist Sam Harris, himself a free will skeptic who has attempted to defend traditional moral notions, has gone so far as to say that, on the basis of science, "we know that determinism, in every sense relevant to human behavior, is true."[21]

Given that free will skeptics generally agree that causal determinism is a genuine possibility (if not a likelihood with regard to human behavior for all intents and purposes), one might think that these skeptics should also agree that neither moral "ought" statements nor claims of moral obligation can be justified with regard to people. This conclusion gains force once we combine the possibility of determinism with the widely accepted principle asserting that "*ought implies can*" (hereafter OIC). The idea behind this principle is that in order

20 Mele, *Effective Intentions*, 157.

21 Harris, *Free Will*, 16.

for it to be true of any agent *A* that he (morally) ought to do *X*/have done *X*, it must actually be possible for *A* to do/have done *X*. OIC has figured prominently in arguments that incompatibilists have levied against compatibilists. One of the basic incompatibilist intuitions driving the free will debate is that since determinism makes it impossible for anyone to *actually*—as opposed to *hypothetically or counterfactually*—do something other than what they ultimately do, it is a mistake to claim that anyone morally ought to have done something other than what they in fact did given that determinism is true.[22] Given, then, that (a) skeptics accept that determinism is a live possibility, if not most likely true with regard to human behavior (for all intents and purposes), (b) determinism precludes the kind of (actual) ability to do otherwise that skeptics believe is relevant to discussions about free will, and (c) the plausibility of the OIC principle, one might reasonably conclude that the skeptic must, at the very least, remain agnostic about the applicability of "ought" statements and obligations to human agents.[23] Put another way, the foregoing considerations suggest that, from the skeptic's perspective, no statements involving moral "oughts" or obligations can be justified with regard to human agents.[24]

22 Whereas the actual sense of the ability to do otherwise requires that an agent could have acted other than she did given the way she and the world *actually was* when she acted, the hypothetical sense requires only that the agent could have done otherwise if something about the world (e.g., the agent's psychology) had been *different* during the time at which the action in question occurred. Though there is much debate as to whether the actual or hypothetical sense of the ability to do otherwise is the sense relevant to the issue of free will, addressing this controversy is not really necessary insofar as the aims of this paper are concerned since skeptics (and incompatibilists generally) are more or less in agreement that insofar as the ability to do otherwise is required for free will, it is the actual sense of the ability to do otherwise that is needed.

23 While the OIC principle might be considered to pose a challenge to compatibilist as well as to skeptical accounts of morality, there is reason to think that the challenge it poses to skeptical accounts is more serious. This is because while skeptics typically view the *actual* ability to do otherwise as being relevant to free will and moral responsibility, compatibilists typically think that the *hypothetical* ability to do otherwise is what matters. Since determinism only appears to prevent an agent from exercising the former ability, the compatibilist is better positioned than the skeptic to deny that determinism precludes the kind of ability to act otherwise that moral responsibility (or morality more generally) requires. That is, a compatibilist is better situated than the skeptic to argue in a given case that since a determined agent *could have* (hypothetically) done something other than the morally bad act in question, it is reasonable to assert that he, therefore, morally *ought to* have done so.

24 Pereboom suggests that such ought statements might not apply to human agents even if determinism were false. As he puts it, "one might claim that if our choices and actions are partially or truly random events, then we could never do otherwise by the sort of agency required for it to be true that we ought to do otherwise" (*Living without Free Will,* 143).

Locating a compelling objection to the conclusion that the free will skeptic must refrain from asserting that moral "ought" statements are ever justifiably uttered about human agents is difficult to locate in the literature. Since traditional skepticism is founded on both an assumption that determinism might be true (if not altogether true) and that the actual ability to do otherwise, rather than the hypothetical ability to do otherwise, is the sense of "could have done otherwise" that is relevant to debates concerning free will, it seems reasonable that a skeptic interested in defending the applicability of moral "ought" statements or obligations for human agents would find it necessary to challenge the OIC principle.[25] As stated earlier, this principle is widely accepted today, and its strong intuitive appeal likely accounts for its having been accepted by historic luminaries such as Kant and G. E. Moore as well as many prominent contemporary participants in the free will debate, such as Ishtiyaque Haji, who has said that it is reasonable to suppose that "any theory of moral obligation . . . should 'validate' [the OIC principle]."[26]

25 Another option for the skeptic is to attempt to justify the applicability of moral obligations/moral "ought" statements to human agents by basing them on the kind of forward-looking moral responsibility that many skeptics are willing to ascribe to people. The issue of whether such an appeal to forward-looking moral responsibility could contribute to a skeptical defense of morality is addressed in section 3.

26 Haji, *Moral Appraisability*, 53. Haji's forceful defense of OIC and how it provides reasons for doubting the possibility of moral obligations for determined agents can be found in Haji, *Moral Appraisability*, 50–53. It is worth pointing out that while Haji's views in this book share some similarities with my own with regard to the kinds of moral claims that do not appear to be available to the skeptic, there are important differences. With respect to the similarities, in addition to the view that determinism precludes the kind of ability to do otherwise that is required for moral obligations (namely, the actual rather than hypothetical sense), he and I also seem to agree that the skeptic is committed to denying that human actions can be morally right or wrong. In terms of how his and my views differ, since Haji asserts that neither moral goodness/badness nor moral blameworthiness are incompatible with determinism, he leaves open the possibility that the skeptic could justify applying such terms to agents in a deterministic world. Concerning blameworthiness, Haji argues that it only requires that an agent *believe* that what she did was morally impermissible, whether or not this was actually the case (see *Luck's Mischief*). Since a skeptic can consistently agree that an agent could believe what she did was morally impermissible, it follows that Haji would not take issue with a skeptic attributing moral blameworthiness to an agent in the actual world, whether or not determinism holds true.

The fact that Haji would presumably allow the skeptic to ascribe certain moral properties to agents (e.g., moral blameworthiness) but not others (e.g., moral obligations) points to what is probably the most significant difference between our positions—namely, that while my account seeks to ground the meaning of moral terms in folk usage, Haji's does not. As I discuss in detail below, empirical research on folk attitudes about free will and moral responsibility suggests that the key folk moral concepts at issue (including moral blameworthiness) are intimately connected with backward-looking features like basic

These points notwithstanding, there is at least one notable skeptic—namely, Pereboom—who calls the OIC principle into question. While Pereboom acknowledges that the principle is "indeed attractive," he hedges by saying that he is "not sure" whether moral "ought" statements can be true given the truth of determinism.[27] The basis of his uncertainty is that there is a sense in which one might appropriately employ "ought" judgements in order to guide others away from engaging in future behaviors that this individual believes they ought not do. According to Pereboom, such *practically rational* "ought" statements are justified insofar as (a) they effectively encourage others from partaking in problematic behaviors that are prohibited by the "ought" statements, and (b) given that the target of the statement is unaware of what their future actions are (regardless of whether or not determinism is true), it appears *epistemically* open to the recipient of the "ought" statement that he can choose to obey it or not.[28] Of course, to say that a moral "ought" statement may be effective in influencing one's behavior while appearing epistemically justifiable from one's personal (and quite possibly mistaken) standpoint of having an open future is not to say that such a statement is *metaphysically justifiable* (a point that Pereboom acknowledges), which, from the skeptic's point of view at least, is the sense of justifiable that matters. After all, were skeptics to accept Pereboom's epistemic understanding of moral "ought" statements—or any similarly "looser" account of them that did not require the actual, rather than hypothetical, ability to do otherwise in order to be applicable—they would now face a serious problem. This is because moral "ought" statements would now seem to apply to determined agents insofar as their behavior can be influenced and their futures appear epistemically open to them. But if determined agents can have moral obligations, it is unclear why we should deny the compatibilist's claim that they can also be morally responsible and, thus, possess free will. I suspect that Pereboom's recognition of this issue may have played a role in his hesitance to adopt an epistemic understanding of "moral ought" statements.[29] It is telling that while he does not commit to holding that all moral "ought" judgments are false in actuality (regardless of whether they appear epistemically justified from

desert that the skeptic is committed to rejecting. Given that it appears that the skeptic's position depends crucially on adopting the folk usage of these key moral terms, I will argue that the skeptic cannot consistently adopt the more compatibilist-friendly understanding of some moral terms that Haji accepts.

27 Pereboom, *Living without Free Will*, 142, 147.

28 For a more detailed discussion, see Pereboom, *Living without Free Will,* 147–48.

29 I return to a similar worry for skeptics in section 3.

a subjective perspective or whether they are effective in influencing others to act), he does claim to be "somewhat sympathetic" to this view.[30]

Another way that the skeptic might attempt to refute the OIC principle is by citing research indicating that the folk find it counterintuitive. As I will elaborate on below, skeptics often speak to the importance of folk intuitions in philosophical debates about free will, and they frequently attempt to support their position by claiming that it fits best with key folk intuitions. Were it the case, therefore, that the folk generally reject the OIC principle, this could provide the skeptic with the ammunition she needs to make a compelling case for how moral "ought" statements are applicable to human beings even if determinism makes it impossible for them ever to have acted otherwise in the actual sense. So what evidence is there that the folk reject the OIC principle? In two separate studies—one conducted by Wesley Buckwalter and John Turri and the other by Vladimir Chituc, Paul Henne, Walter Sinnott-Armstrong, and Felipe De Brigard—subjects appeared to attribute moral "ought" statements and obligations to agents even when they lacked the actual ability to perform the actions that they were viewed as being obligated to do.[31] Caruso has commented that these empirical findings support the claim that "the OIC principle is a philosopher's invention infected by mistaken assumptions about moral responsibility."[32] A more recent study conducted by Miklos Kurthy and his associates, however, points out how each of these previous studies was flawed insofar as the prompts given to subjects made it difficult to draw any firm conclusions as to whether or not they accept the OIC principle.[33] To provide a more accurate analysis of this issue, Kurthy et al. ran another version of the 2015 Buckwalter and Turri studies using prompts that eliminated the kind of ambiguity in the prompts employed in the original studies. With these improved prompts, Kurthy and his colleagues completely reversed the results generated by the original Buckwalter and Turri experiments, leading them to conclude that "people do make judgments largely compatible with the OIC principle, at least in cases in which the inability is not self-imposed."[34] Insofar as the studies by Kurthy et al. seem to provide more accurate insights into folk attitudes regarding the OIC principle, there is reason to agree with the principle's validity.[35] And if this is true,

30 Pereboom, *Living Without Free Will,* 148.

31 See Buckwalter and Turri, "Inability and Obligation in Moral Judgment"; and Chituc et al., "Blame, Not Ability, Impacts Moral 'Ought' Judgments for Impossible Actions."

32 Caruso, "Skepticism about Moral Responsibility."

33 Kurthy, Lawford-Smith, and Sousa, "Does Ought Imply Can?"

34 Kurthy, Lawford-Smith, and Sousa, "Does Ought Imply Can?" 15.

35 For additional evidence that the folk accept something like the OIC principle, see C. Clark et al., "Free to Punish."

it reinforces the claim that skeptics cannot justifiably attribute moral "ought" statements or obligations *to* human agents. At the very least, the burden is on the skeptic to defend how the use of such terms can be warranted with regard to people who may be subject to deterministic laws. Such a conclusion, however, turns partly on how much weight we should give to folk intuitions when arbitrating disputes about morality.

1.3. The Relevance of Folk Intuitions with Regard to Morality

When it comes to discussions about any philosophical concept that has its basis in folk discourse (e.g., free will, morality, God), a fundamental question that must be addressed concerns the extent to which philosophers' usage of the concept should resemble the folk conception of it (assuming that there is a discernable folk understanding of the concept in question). One popular view among contemporary philosophers is that philosophical discussions of such concepts should resemble how they are used in ordinary language. The main justification behind this view is that were philosophers to use a term like "morality" in a way that was too far afield from the ordinary conception of it, their subsequent discourse would be more likely to muddy the conceptual waters than to provide clarity. Put another way, the worry is that by diverging too far from the folk concept, philosophers would be essentially changing the subject in such a way that their discussions would have very little to do with the original concept that sparked philosophical discussions in the first place.[36]

Freewill skeptics have been especially vocal in making the case that philosophical discussions of concepts that are rooted in folk discourse—such as free will—must not veer too far from the folk concepts that originally gave rise to the philosophical controversies. Hence, you find skeptics such as Pereboom warning against revising a root folk concept "so radically that the concept used is no longer the same."[37] His argument for skepticism is based on the claim that it stands as the most plausible perspective on free will if we go by what the folk mean by terms like "free will" and "moral responsibility." Nadelhoffer is another skeptic who places importance on how the folk understand these terms, and the case he makes for retaining the folk understanding of them in the philosophical debates is compelling. As he puts it:

> I think that what we call things matters. And I also think the terms "free will" and "moral responsibility" carry an awful lot of both metaphysical and historical baggage. Given this web of historical associations [e.g.,

36 For further discussion of the role folk intuitions play in philosophical discourse, see Morris, *Science and the End of Ethics.*

37 Pereboom, "Hard Incompatibilism and Its Rivals," 24.

> religious overtones, Cartesian dualism] I do not think that we should revise the terms "free will" and "moral responsibility." If we do not have the kind of agency and responsibility that people have traditionally thought we had, we invite confusion by continuing to use the old terms to talk about what we actually do have—especially when we could use other terms which are less loaded.[38]

Like Pereboom, Nadelhoffer believes that skepticism is the preferred view in light of how we cannot possess free will and moral responsibility as the folk understand them.[39]

That skeptics (and incompatibilists more generally) should emphasize the importance of folk intuitions makes sense given how incompatibilism is more metaphysically demanding than compatibilism. As I pointed out in a paper co-written with Nadelhoffer, Eddy Nahmias, and Jason Turner, incompatibilism is more demanding than compatibilism since the conditions that it requires for free will—e.g., "at a minimum, indeterministic event-causal processes at the right place in the human agent, and often, additionally, agent causation"—go beyond what the compatibilist requires.[40] The question then becomes why we should agree with the incompatibilist's conception of free will given that it is more complex than its compatibilist counterpart and, all things being equal, a less complex understanding of a philosophical term is generally preferable to a more complex understanding. The most plausible answer that incompatibilists can give is that all things are not equal since their view comes closest to the notion of free will accepted by the folk.

It is not, however, only skeptics and other incompatibilists who believe that philosophers ought to preserve key aspects of folk concepts when discussing free will and other related issues. Nahmias is a compatibilist who, along with myself and the other co-authors, also speaks to the importance of paying heed to folk intuitions. Echoing the views of the aforementioned skeptics, he agrees that "because the free will debate is intimately connected to ordinary intuitions and beliefs via [certain] values and practices, it is important that a philosophical theory of free will accounts for and accords with ordinary people's understanding of the concept and their judgments about relevant cases."[41] Mele, another who falls outside of the incompatibilist camp, makes a similar point by suggesting that where the folk hold a widely shared view of a particular philosophical

38 Nadelhoffer, "The Threat of Shrinking Agency and Free Will Disillusionism," 176–77.

39 That is, in the sense in which these terms are intimately connected to basic desert. Strawson and Caruso are other skeptics who argue along these lines.

40 Nahmias et al., "Is Incompatibilism Intuitive?" 31.

41 Nahmias et al., "Is Incompatibilism Intuitive?" 30.

concept, "such judgments provide evidence about what the concept is," and warns that a philosophical analysis "that is wholly unconstrained by [the folk] concept runs the risk of having nothing more than a philosophical fiction as its subject matter."[42]

Given that philosophical discussions involving concepts based in folk discourse ought to proceed with an understanding of these concepts that is not unduly at odds with how the folk understand them—a view, again, that skeptics are particularly committed to—the question then becomes: How do the folk actually conceive of morality and its constitutive parts, such as morally right and morally wrong actions?[43] While a full-fledged account of how the folk conceive of morality is a subject that I am not equipped to address in this essay, I do wish to point out that the manner in which the folk understand it seems to clearly involve attitudes that can properly be called backward looking. To be more specific, the folk concept of morality appears to be, to a significant extent, grounded on moral responsibility in the basic desert sense discussed earlier. The notion that there is a fundamental desert-based component to people's moral judgments has a long history in philosophy. John Stuart Mill asserted that "[w]e do not call anything wrong, unless we mean to imply that a person ought to be punished in some way or another for doing it."[44] Speaking to how he has the basic desert understanding of punishment in mind, he continues by saying that the "sentiment of justice, in that one of its elements which consists in the desire to punish, is thus, I conceive, the natural feeling of *retaliation or vengeance*, rendered by intellect and sympathy to those injuries that is, to those hurts, which wound us through, or in common with society at large."[45] Echoing the same opinion, Friedrich Nietzsche mentions that "wherever [moral] responsibilities are sought, it is usually the instinct of wanting to judge and punish which is at work."[46] Richard Joyce lends a more contemporary voice to this perspective by pointing out that "when we examine our ordinary concepts of desert and justice, what we seem to find is an idea of the world having a kind of 'moral equilibrium.' When a wrong is done this equilibrium is upset, and the

42 Mele, "Acting Intentionally," 27. Speaking to how he is reluctant to identify himself as a genuine incompatibilist, Mele has stated that he is "officially agnostic about the truth of compatibilism" (*Free Will and Luck*, 78).

43 To be clear, the issue here is not what *specific actions*, etc., the folk generally hold to be moral or immoral, but rather what the folk generally *mean* when they say that a certain action or agent is morally good/bad or morally right/wrong.

44 Mill, *Utilitarianism*, 187.

45 Mill, *Utilitarianism*, 188, emphasis added.

46 Nietzsche, *Twilight of the Idols*, 499.

administration of the appropriate punishment is seen as the procedure that will effect its restitution."[47]

Theoretical support for the claim that our moral judgments are steeped in the kinds of retributivist attitudes that characterize the basic desert sense of moral responsibility comes by way of the prevailing evolutionary account of our moral faculty. According to this view, our moral faculty is an evolutionary adaptation that allowed our ancestors to transcend their selfish proclivities in order to engage in the type of cooperative behaviors that gave their groups a fitness advantage over their less cooperative rivals. In order to ensure that members of the group conformed to its norms, it was necessary, this account goes, that the moral sentiments included the desire to punish wrongdoers and reward those who abided by the group's rules. Charles Darwin himself favored this account of morality's origins, and it has been championed in more recent times by psychologist Jonathan Haidt among others. Recent empirical work has lent strong support for this evolutionary picture of our moral faculty and how retributivist sentiments play a central role in our moral judgments.

In a series of studies, psychologist Cory Clark and her colleagues provide evidence that people's beliefs in free will are motivated by a desire to punish perceived wrongdoers. For instance, subjects were significantly more likely to attribute free will to the actions of an immoral agent than they were to the actions of a neutral agent. Interpreting their results, Clark et al. claim that considering the immoral behavior of others caused subjects to attribute higher levels of free will and moral responsibility to the immoral agents in order to justify their desire to punish them. As they put it:

> Moral responsibility is a construct that permits societies (and individuals) to blame and punish others for their misdeeds. Insofar as free will is a prerequisite for moral responsibility, ascribing free will to criminals or other miscreants provides a crucial justification for punishing them for their actions.[48]

In what follows, Clark and her associates explain how their experiments provide empirical reinforcement to the evolutionary account of the human moral faculty discussed earlier: "The core of our argument is that this subjective experience of free will gains motivational reinforcement by facilitating the assignment of moral responsibility, which in turn supports the crucial task of punishing individuals who act in ways that are detrimental to cohesive group functioning."[49]

47 Joyce, *The Evolution of Morality*, 68.

48 Clark et al., "Free to Punish," 503.

49 Clark et al., "Free to Punish," 509.

Clark et al.'s studies suggest that retributivist motivations are driving people's tendencies to assign free will and moral responsibility to agents viewed as acting immorally and, hence, that it is the basic desert sense of moral responsibility at work in people's minds when they are passing judgment on agents deemed to be immoral. However, their studies do not explicitly aim to uncover whether the desire to punish that is apparently driving subjects' judgments to assign free will to immoral agents is retributivist or consequentialist in nature. Another series of studies by Azam Shariff and his associates set out to determine the nature of these punitive inclinations with more precision.[50] Their results provide robust support for the view that retributivist impulses—as opposed to consequentialist considerations—were primarily responsible for eliciting increased attributions of free will for immoral agents in the experiments conducted by Clark et al. The work of Shariff et al. also provides some of the strongest evidence yet that folk moral judgments invoke the basic desert sense of moral responsibility and that, hence, basic desert constitutes an essential component of folk moral judgments.

Shariff and his colleagues set out to discover whether reducing people's beliefs in free will would make them less retributive about punishment. They assumed that if folk attributions of moral responsibility are dependent upon attributions of free will, then we should find that people tend to be less retributive in their judgments about the kinds of treatment that a particular individual should receive insofar as they believe that the individual in question did not act freely. An implicit assumption these researchers were operating under is that folk conceptions of moral responsibility are closely connected to deservingness of retributivist treatment. Put another way, the researchers assumed that the folk are operating under a conception of basic desert-based, as opposed to consequentialist-based, moral responsibility. The results of their studies suggest that their assumption was correct. In study 1, they found that while stronger beliefs in free will predicted retributivist punishment, they were not predictive of consequentialist punishment. In study 2, the researchers found that the subjects who had their beliefs in free will diminished by reading a scientific argument against free will recommended prison sentences that were roughly half of those that were recommended by subjects in a control group, suggesting that reducing the free will beliefs of the test group made them significantly less retributive than subjects in the control group. The results led Shariff and his colleagues to infer a tight connection between actions perceived as immoral with the natural desire to inflict retributivist punishment upon immoral agents. Hence, they conclude that "Humans respond to transgressions with an urge

50 See Shariff et al., "Free Will and Punishment."

to exact punitive costs on the transgressor."[51] Shariff et al. add further support to the notion that there is an essential basic desert element at work in people's moral judgments by highlighting the important role that blameworthiness appears to play. More specifically, they mention how "the mediational effect of perceived blameworthiness made a strong case for the role of moral responsibility in the effect of diminished free will belief on retribution."[52] In light of the ample empirical evidence suggesting that exposure to perceived immoral acts tends to elicit strong retributivist sentiments from test subjects which, in turn, influences their attributions of free will and moral responsibility, the burden of proof would appear to fall upon those who would deny that folk moral judgments have a retributivist (basic desert) element as a central component. This burden is made even stronger by more recent research by Jim Everett and his colleagues that set out to explain why political conservatives tend to believe in free will more than political liberals.

In order to account for the empirical observation that conservatives have both a higher belief in free will and a greater tendency to attribute it to agents, Everett et al. set out to test their hypothesis that it was conservatives' greater tendency to moralize than liberals—i.e., to give moral weight to a larger set of actions and behaviors than liberals—that accounts for this phenomenon. While the studies that they ran reinforced the claim that conservatives have a greater tendency to both believe in free will and to attribute it to others, they found that it was conservatives' tendency to make more *moral judgments* (particularly judgments about moral wrongness)—mediated by a desire to hold agents blameworthy— that accounted for this difference among liberals and conservatives rather than political views or metaphysical beliefs about human

51 Shariff, et al., "Free Will and Punishment," 1564.

52 Shariff et al., "Free Will and Punishment," 1567. Additional empirical support for the standard evolutionary account of our moral faculty and how the instinct toward retributivist punishment acts as a driving force in our moral judgments can be found in Fehr and Gächter, "Altruistic Punishment in Humans"; and Hamlin et al., "How Infants and Toddlers React to Antisocial Others." Fehr and Gächter's experiments indicate that the desire to punish wrongdoers runs so deeply in the human psyche that people will punish others for transgressions even when doing so comes at a significant cost to themselves. The work of Hamlin and her colleagues demonstrates not only that children as young as twenty-one months administer reward and punishment to others based on their good or bad behavior, but also that babies as young as eight months show an affinity for those who dole out punishment to "bad guys" and an intolerance for those who reward "bad guys." To attribute the behavioral tendencies of young children that Hamlin and her colleagues discovered to moral instruction seems pretty clearly to overestimate the ability of young children to comprehend complex social norms and to manifest such moral lessons into appropriate behaviors. The more likely explanation is that these tendencies are manifestations of innate moral capacities involving retributivist instincts that have been forged by evolutionary pressures.

autonomy. In one of their studies, for instance, Everett and his associates found that when liberals were more motivated than conservatives to hold an agent as having acted morally wrong, liberal subjects tended to assign a greater level of free will to the agent than conservative subjects. Everett et al. draw the following conclusion: "Supporting the idea that differences in moralization underpin the specific free will attributions, we found that when adding perceived moral wrongness . . . to the model, political ideology no longer predicted ascriptions of free will . . . with only reported moral wrongness significantly predicting free will attributions."[53] They go on to say that "people endorse the idea of free will in order to justify their desire to blame others for moral wrongdoing."[54]

Combining the research of Shariff et al. and Everett et al. suggests that folk moral judgments—including those about moral rightness and wrongness—are inseparable from free will, backward-looking moral responsibility, and blame. Shariff and his colleagues found that folk judgments about free will are tied to backward-looking moral responsibility and retributivist, as opposed to consequentialist, punishment. Everett and his associates helped further clarify the relationship between folk moral judgments, free will, and backward-looking moral responsibility by finding that judgments about moral wrongness (regardless of political ideology) were positively correlated to judgments about free will and a desire to blame immoral agents.

It is important to clarify here that my point is not that there are not any forward-looking elements in the folk concept of morality, but merely that it contains fundamental backward-looking elements as well. This will be important to remember since my primary argument against skeptical defenses of morality—such as that provided by Pereboom—is that they illicitly aim to revise the folk understanding of morality by eliminating these essential backward-looking elements. I will argue that it is inconsistent for the skeptic to alter an important folk concept in this way since their primary objection to free will revisionists is that they themselves are improperly altering the folk concept of free will by stripping it of its fundamental incompatibilist elements. Employing a parallel type of reasoning to what I mentioned above in reference to how the folk understand morality, skeptics do not deny that the folk concept of free will is infused with elements that are central to compatibilist notions of free will (such as the ability to deliberate among choices and act upon reasons). Their primary issue with the revisionists is that the watered-down notion of free will that they favor—that is, one without any incompatibilist commitments—is

53 Everett et al., "Political Differences in Free Will Belief Are Associated with Differences in Moralization," 470.

54 Everett et al., "Political differences in Free Will Belief Are Associated with Differences in Moralization," 479.

too different from the folk concept to be relevant to the primary philosophical debates surrounding free will. Likewise, I will argue that the weakened notion of morality that skeptics defend, insofar as it diverges too far from the folk concept of morality by eliminating its basic desert elements, is equally ill suited for helping address key philosophical debates in ethics.

To this point I have discussed how skeptics are in general agreement that the lack of free will commits them to rejecting the backward-looking type of moral responsibility, moral praise/blame, and moral desert with respect to human beings and their actions. Furthermore, I have mentioned why the skeptic seems unable to justify the attribution of any genuine moral "ought" statements or obligations to human agents. But if this is correct, one may reasonably ask what kind of morality is left for the skeptic to defend. To address this question, it will be helpful to examine the account of morality defended by Derk Pereboom since it is perhaps the most detailed account of morality offered by a free will skeptic.

2. Pereboom's Skeptical Account of Morality

In addition to accepting what I am calling the "core moral denials of free will skepticism," Pereboom seems willing to grant the inapplicability of moral "ought" statements (and presumably moral obligations as well) to human agents. Nonetheless, he still believes that people and their actions can properly be subject to moral appraisal. As he puts it, "Even if moral 'ought' judgments are never true, it would still seem that moral judgments such as 'it is morally good for *A* to do *x*' and 'it is morally bad for *A* to do *y*' still can be."[55] The question now arises as to how Pereboom conceives of terms such as "morally good," "morally bad," "morally right," and "morally wrong" given that such terms cannot be cashed out in terms of moral responsibility, desert, praise/blame, or even moral obligations. Put another way, what exactly would it mean for Pereboom to assert that a human agent's action was *morally wrong*?

Insight into the kind of morality that Pereboom believes can be reconciled with free will skepticism can be found in his book, *Living without Free Will*. So what does he say? Though Pereboom rarely offers specifics in terms of the kind of morality he believes can coexist alongside skepticism, it is possible to get at least a rough outline of what he has in mind from some examples and discussions that he provides. The following is one such example:

> Suppose you say to an animal-abuser, "You ought not to abuse that animal," but then you find out that he has a psychological condition

55 Pereboom, *Living without Free Will*, 143.

> (which he could have done nothing to prevent) that makes animal-abusing irresistible for him, so that he cannot help but abuse the animal. From my point of view, there is an appreciable strong pull to admitting that the "ought" judgment was false, but there is relatively little to denying that abusing the animal is morally wrong for him.[56]

As for what Pereboom takes "morally wrong" to mean, he says that "[h]ard incompatibilist moral worth is indeed moral, but it is more similar to the value we might assign to an automobile or a work of art."[57] Along the same lines, he asserts that the moral goodness of a human agent "is more like the aesthetic sort than is often thought because it does not involve blameworthiness or praiseworthiness, but it is no less moral for that."[58]

Pereboom even maintains that skepticism is compatible with non-consequentialist forms of morality. As he puts it, "most of the descriptive and prescriptive content of any ethical system is consistent with hard determinism, and more inclusively, with hard indeterminism."[59] "The reason for this," he says, "is that the metaphysical bases for non-consequentialist positions in general, insofar as they have been developed, do not clearly involve an essential appeal to notions of freedom unavailable to the Hard Indeterminist."[60] Evidently Pereboom believes that one could subscribe to a genuine version of Kantian ethics, for example, so long as he abandons certain aspects of this moral outlook such as free will, moral responsibility, and blameworthiness.

In laying out my reasons for rejecting Pereboom's attempt to salvage morality in the face of skepticism, I begin by mentioning a general point that appears applicable to any defenses of morality by genuine free will skeptics. The point is that there are strong reasons for rejecting any such defense insofar as it breaks in fundamental ways from the folk conception of morality. In section 1.3 I discussed how many philosophers—especially free will skeptics—have stressed the importance of not deviating too far from relevant folk concepts when engaging in philosophical debates. And yet by advocating for a kind of morality that eschews blame, desert, and backward-looking moral responsibility, this is precisely what skeptical defenders of morality are doing. Concepts such as these are so deeply entrenched in our ordinary moral judgments that there is little doubt that a typical representative of the folk would have enormous difficulty resonating with a moral outlook where notions such as moral blame, etc.,

56 Pereboom, *Living without Free Will*, 147.

57 Pereboom, *Living without Free Will*, 153.

58 Pereboom, *Living without Free Will*, 153.

59 Pereboom, *Living without Free Will*, 150.

60 Pereboom, *Living without Free Will*, 150.

have no place. At the very least, therefore, it seems incumbent on any skeptical account of morality to explain why we should accept a notion of morality that is so vastly at odds with ordinary intuitions.

This problem is especially acute for Pereboom, who has argued extensively about the importance of using philosophical terms in a way that closely resembles folk usage. Addressing the efforts of philosophers like Manual Vargas who have argued that making headway in the free will debate requires us to revise key terms like "free will" and "moral responsibility" in a compatibilist way that eliminates some of the more controversial elements that are included in the folk understanding of them, Pereboom poses the following "crucial question":

> [Is] there a defensible compatibilist conception of free will near enough to the folk's to count as a natural extension of it, one that can do enough of the work the folk conception does in adjudicating questions of moral responsibility and punishment, and in governing our attitudes to the actions of those around us?[61]

The essence of his case against compatibilists who seek to revise key terms in the fashion that Vargas and others do is that they end up with an understanding of free will that is too different from that of the folk. Pereboom's stance is that by changing the meanings of these key terms so drastically, philosophers are likely to cause confusion in the eyes of the folk (and quite possibly other philosophers) who will interpret the revisionists to be defending the old folk concepts. He illustrates this point using the following example:

> If people started saying "he's morally responsible for the murder since he did it of his own free will," but did not mean to claim that he in the basic sense deserved blame or punishment, then it could well be misleading to use the old terminology, *since an audience might well be confused about which concept these words stand for*.[62]

Since Pereboom believes that the folk have the basic desert sense of moral responsibility in mind when it comes to free will, he rejects revisionist attempts to extricate this sense of moral responsibility from what it means to exercise free will.

One of the main reasons that Pereboom wants to preserve the folk concepts of free will and moral responsibility in philosophical discourse is that he believes that one of the primary jobs of philosophers is to show us when our ordinary attitudes are mistaken. In particular, he believes that philosophers' discussions of free will are key to showing us that our concepts of blameworthiness

61 Pereboom, "Hard Incompatibilism and Its Rivals," 25.

62 Pereboom, "Hard Incompatibilism and Its Rivals," 26, emphasis added.

and retributivist desert are indefensible and ought to be jettisoned. Since he believes that the folk take the existence of free will to provide the primary philosophical justification for these backward-looking attitudes, Pereboom worries that revising the term "free will" along the lines that Vargas suggests would make it more difficult to convince the folk that they ought to dispense with these attitudes as well as the problematic practices that they give rise to (e.g., retributivist criminal punishment). Of note here is that Pereboom believes the importance of preserving folk concepts in philosophical discussions extends beyond the free will debate. As he puts it, "More generally, when deciding how to revise, we need to retain concepts that facilitate our thinking that some of our attitudes and beliefs are mistaken."[63]

The main point I wish to make here is that the same arguments that Pereboom offers against revising the folk concepts of free will and moral responsibility can be applied with just as much force against revising the folk concept of morality in the manner he suggests given its close connection to notions such as moral responsibility, blame/praise, basic desert, and obligation. Even Pereboom himself admits that the watered-down kind of morality that he believes can coexist with skepticism "differs significantly from the ordinary conception."[64] A major question facing Pereboom, then, is whether he can consistently defend his version of morality given the kinds of arguments he offers against free will in general and Vargas-style compatibilism in particular.[65] From what has preceded, we can discern three distinct, though related, criteria that Pereboom believes an adequate revisionist account of free will must satisfy: (a) the notion of free will that it employs must be near enough to the folk's notion to count as a natural extension of it, one that can do enough of the work the folk conception does in (among other things) governing our attitudes to the actions of those around us; (b) it must not result in the audience being confused about which concepts key words like "free will" and "moral responsibility" stand for; and (c) it must not revise the key term(s) in question so drastically that it would damage the effort of retaining concepts that facilitate our thinking that some of our attitudes and beliefs are mistaken. Pereboom's argument against Vargas's revisionist account of free will is, in essence, that it fails to meet each of these criteria.

The question before us is whether the kind of test that Pereboom proposes for a revisionist account of free will would be passed by his revisionist conception

63 Pereboom, "Response to Kane, Fischer, and Vargas," 203.

64 Pereboom, *Living without Free Will,* 153.

65 While trying to categorize Vargas's view on free will can be tricky, it seems appropriate to consider it a form of compatibilism. Hence, we find Michael McKenna and D. Justin Coates using the term "revisionist compatibilism" in referring to Vargas's position ("Compatibilism").

of morality. In terms of whether his skeptical understanding of the term "morality" comes close enough to the folk concept to play all the key roles it does in holding people accountable, etc., it seems pretty clear that a negative answer is warranted. As I alluded to earlier, once we strip from the folk notion of morality concepts like moral responsibility, praise and blame, basic desert, and having moral obligations, it is not clear what would be left since, as I have discussed, all of these elements seem to be central features of folk moral judgments. While Pereboom maintains that we can still properly use terms like "moral good/bad" and "moral rightness/wrongness" to describe human agents and their actions, these terms have traditionally been so inextricably tied to moral responsibility (including backward-looking moral responsibility), praise/blame, etc., that any usage of terms like "morally wrong" in the way that Pereboom conceives of them would change their usual meaning in ordinary discourse. After all, it seems beyond contention that the folk do not use a term like *moral rightness* as more or less a term of aesthetic appreciation or as the kind of label we might give to "an automobile or a work of art" that pleases us on some level—both of which capture how Pereboom suggests one should interpret this term from a skeptical perspective.[66] Furthermore, since folk attitudes toward others, as well as their treatment of one another, would change dramatically under the revised kind of moral outlook that Pereboom favors (i.e., one that rejects retributivist attitudes and practices), it seems fair to say that his revisionist account of morality is not a natural extension of the folk understanding of morality.

It seems equally clear, if not more so, that by revising morality in the manner he suggests, Pereboom would be promoting the kind of terminological confusion that he warns us so sternly about with regard to free will revisionism. Under the reasonable assumption that a statement such as "*A* was morally wrong to do *X*," when uttered by the folk (and most other philosophers) generally implies, among other things, that *A* is both morally responsible and morally blameworthy for *X*-ing and that *A* ought not have done *X*, there is little doubt that the meaning of this statement is very different from the same sentence being uttered by Pereboom. Given this, we should expect that confusion would often arise among a general audience when hearing Pereboom and other like-minded skeptics using moral terms that are devoid of the implications that most people usually take them to have.

Finally, if Pereboom is worried that revising the term "free will" away from its usual meaning would make it more difficult to facilitate our acknowledgment that some of our attitudes and beliefs (e.g., retributive punishment is

66 See my prior discussion of Everett et al. with regard to how the folk appear to understand moral rightness/wrongness.

justifiable) are mistaken, he should be equally if not more concerned about revising the term "morality" along the lines he suggests. Recall that his worry is that revising the term "free will" in the manner that Vargas suggests would make it more difficult to change the folk's views about basic desert, retributive punishment, and the like. But since the folk understanding of "morality" is just as encumbered (if not more so) with backward-looking attitudes, any effort to preserve the use of this term in our ordinary language would appear to cause the same kinds of inability to confront folk views about basic desert that Pereboom deems both false and problematic.

3. Can Forward-Looking Moral Responsibility Help Salvage a Skeptical Account of Morality?

I mentioned earlier that while it appears that skepticism precludes a kind of backward-looking moral responsibility, few, if any, philosophers have argued that skepticism is incompatible with a forward-looking type of moral responsibility. In light of this, a skeptic might suggest that a defensible account of morality can be constructed once we eschew the backward-looking variety of moral responsibility in favor of its forward-looking counterpart. With this understanding of moral responsibility in tow, the argument goes, the skeptic can now explain how moral terms like "good" and "bad," "right" and "wrong," and "obligations" can properly apply to human agents and their actions even though concepts like basic desert, (genuine) moral praiseworthiness, and (genuine) moral blameworthiness cannot. The idea is that certain moral labels would do no more than indicate that certain individuals, through their decisions and actions, are proper targets of particular kinds of forward-looking responses (e.g., imprisonment geared toward rehabilitation) that can be expected to increase the likelihood of influencing their (or perhaps others') future behavior in order to achieve consequentialist ends such as increasing overall happiness. Might this serve as an effective strategy for defending a skeptical account of morality?

I previously discussed how the folk notion of moral responsibility is constituted in part by backward-looking elements such that, for example, the judgment that one is morally responsible for a morally bad action is often (if not most of the time) associated with the judgment that he/she ought to suffer retribution. It follows from this that any attempt to revise the term "moral responsibility" in a way that eliminates all backward-looking elements would leave us with a notion of moral responsibility that is fundamentally different from the folk meaning of the term. In light of this, we need an explanation of how such a revisionist account of moral responsibility can be justified by the skeptic given how a desire for preserving the folk understanding of key philosophical terms

serves to motivate the skeptic's position in the first place. Beyond this concern, another issue presents itself: if it is open to the skeptic to revise moral responsibility in a way that eliminates the backward-looking attributes it contains in the eyes of the folk, then why is it not equally open to a compatibilist to revise *free will* in a similar way—that is, in a way that denotes no more than the possession of the types of capacities (e.g., the ability to apprehend and respond appropriately to a society's norms) that are needed to render one an appropriate target for forward-looking (but not backward-looking) treatment? Since philosophers generally construe free will as being the control condition for moral responsibility, it is reasonable to expect that where one construes moral responsibility in a sense that is stripped of any backward-looking elements, they should also conceive of free will in a similar fashion. But if this is true, then it is unclear why the skeptic should choose to reject free will rather than accept a revised notion of it, something along the lines of that favored by Daniel Dennett.

Dennett's compatibilist account of free will shares many similarities with the most common forms of skepticism, including a naturalistic account of human decisions and behaviors, a rejection of basic desert moral responsibility, and the idea that punishments can only be justified by forward-looking considerations.[67] Unlike the skeptic, however, Dennett asserts the existence of free will, which he understands as basically the kind of capacity that allows us to conceive of the consequences of our actions—as well as to understand prevailing social norms—and respond to them accordingly. When it comes to the related issues of free will and morality, Dennett's concern is purely pragmatic in that he wants to know how we can justify punishment in order to ensure a well-functioning society. Since he believes that punishments based on forward-looking considerations are all we need to achieve this goal (and the only kind of punishments that are justifiable), he argues that it is necessary for us to identify which agents are proper subjects of such punishments and he uses the term "moral responsibility" to capture the status of such agents. Furthermore, he believes a well-functioning society needs a way to distinguish individuals whose capacities make them appropriate targets of these punishments from those who are not. He uses the term "free will" to refer to this capacity. The challenge for the skeptic is to

67 Dennett, *Freedom Evolves.* See also Clark, "Exchange on Waller's *Against Moral Responsibility*." While Dennett occasionally speaks as if his account of free will can accommodate a purely backward-looking type of moral responsibility, he has made it clear that any reactions toward either proper or improper behavior (e.g., rewards and punishments) that his account endorses are ultimately justified by forward-looking considerations. Hence, in a recent discussion with Gregg Caruso, he declares, "*Of course* it is the 'forward-looking benefits' of the whole system of desert (praise and blame, reward and punishment) that justifies it" (Warburton, "Just Deserts").

explain why, if it is legitimate to revise either the term "moral responsibility" or other moral terms like "morally wrong" so as to eliminate the backward-looking properties that are engrained in folk conceptions of them, we cannot also revise the term "free will" along similar lines.[68] Presumably, any skeptical resistance toward doing so would be based on the kinds of arguments Pereboom offered when addressing the dangers of revising key philosophical terms away from their folk meanings (the need to avoid terminological confusion, etc.). But similar to the objection I brought against Pereboom, this understandable apprehension toward revising key folk concepts would appear to be equally warranted against any attempts to revise moral terms in a way that jettisons the backward-looking elements that are built into the folk understanding of them.

4. Taking Agents Out of the Equation

Suppose that my foregoing arguments have persuaded a skeptic to agree that moral terms like moral responsibility, moral obligation, moral praise/blame, and even moral rightness/wrongness are never applicable to human agents since the use of such terms by the folk has implications that the skeptic rejects (e.g., the propriety of retributivist attitudes). A skeptic who desires to preserve our use of moral concepts might assert that while my arguments imply that we should abandon some of our moral concepts (specifically those that pertain to *human agents*), there are other moral concepts (namely, those that refer to positive *states of affairs*) that could still persist, and these moral concepts could form the basis of a revised moral perspective that sits comfortably alongside skepticism. Consider for instance how people often talk about the moral importance of reducing the impact of climate change or eradicating cancer. In speaking of such outcomes in moral terms such as being "morally good," the idea is that they are something that we should aspire to bring about.

My response to this hypothetical effort to defend morality in light of skepticism is twofold. To begin with, I would point out that this understanding of morality is a far cry from the kind of morality that skeptics such as Pereboom, Caruso, and Honderich have tried to defend. Each of them has made clear that the morality that they are interested in pertains to human beings and their actions. Thus, if the only kinds of moral properties a skeptic could defend concerned states of affairs rather than human agents, I doubt that these skeptical defenders of morality would find much to celebrate. Beyond this, however, it

68 One could even make the case that revising free will is more justifiable than revising these moral terms since while it is beyond reasonable doubt that the moral terms as used by the folk have certain backward-looking elements built into them, there is less certainty as to whether the folk notion of free will consists of essential backward-looking features.

is not clear to me what is entailed by the claim that a certain future outcome is "morally good." Presumably, in saying that eradicating cancer would be "a morally good state of affairs," one implication—and one that the folk would likely draw from such a statement—is that *we as humans* are somehow *morally responsible* or *morally obligated* to work toward eliminating cancer. But since this runs against my prior arguments that neither moral responsibility nor moral obligations are applicable to human agents from the skeptic's perspective, this interpretation would place the burden on the skeptic to explain how any state of affairs could be morally good in this way. Perhaps we could construe any claim about an outcome being "morally good" as amounting to nothing more than a single person, group of people, complete population, etc., being positively disposed to it. In this case, it is unclear what role the term "moral" is playing here. Under such an interpretation, it seems reasonable to assert that the individual(s) positively disposed to the outcome in question has *prudential* or *self-interested* reasons for seeing that it is brought about and nothing more. That being the case, an argument is called for to explain why it would be appropriate to add a moral element to the strictly prudential claim since it does not appear to accomplish anything other than imparting onto certain individuals particular moral obligations, etc., that do not appear justifiable in light of my previous arguments. At any rate, even if a skeptical account of morality that applies only to states of affairs and not to persons could be constructed in a way that does not stray too far from ordinary folk moral attitudes—an unlikely prospect in my opinion—I would still consider the arguments of this paper a success insofar as they persuaded skeptics to refrain from making moral ascriptions to people.

In the end, it appears that skeptics wishing to preserve the use of moral terms are faced with the horns of a dilemma. On the one hand, they can argue for the importance of retaining the folk understanding of key terms like free will and moral responsibility with their attendant backward-looking elements. In which case, it seems the proper course of action (given that the skeptical perspective is the correct one) is to deny the existence of free will, moral responsibility, or any moral properties whatsoever, since any attempt to preserve these concepts in philosophical discourse would seem to require changing their folk meanings too drastically. On the other hand, they can argue that such a revisionism of folk notions like moral responsibility is justified since it is necessary for some useful purpose, such as ensuring that society functions well. In this case, however, it would be unclear why the skeptic should not switch allegiance to Dennett-style compatibilism since the case for revising the folk understanding of free will can be justified on the same grounds. For my part, I believe that philosophers like Pereboom and Nadelhoffer make a strong case for not revising terms like free will and moral responsibility in a way that fundamentally

changes the folk meaning of these terms. I agree with them that doing so runs the risk of muddying up the waters with regard to the philosophical debates. I also worry that the folk might misinterpret the outcome of such debates such that, for example, when a philosopher affirms the existence of a type of moral responsibility that is devoid of backward-looking elements, a layperson may nonetheless take this to justify their retributive attitudes. And while I believe that this sort of problem is likely to outweigh whatever advantages would come by revising folk terms in such fundamental ways, this subject requires a deeper analysis than I have provided here. Regardless, it would appear that the burden is on the skeptic to explain why a substantial revisionism of key philosophical terms away from their folk meanings is acceptable in some cases (e.g., morality, moral responsibility) but not others (e.g., free will).

5. CONCLUSION

I have argued that skeptics have yet to succeed in their attempts to construct a compelling case for how the rejection of free will can be reconciled with a worldview that retains traditional moral concepts such as moral wrongness or moral responsibility. To this end, I analyzed Pereboom's defense of morality and argued for why it falls short of explaining how moral properties can plausibly be attributed to human agents lacking free will. I also made the more general point that any skeptical defense of morality is likely to fail insofar as the morality it ends up defending will almost certainly break too drastically from traditional folk moral notions that are heavily embedded with features that skeptics believe are untenable, including backward-looking elements such as basic desert. Given the skeptic's emphasis on the importance of retaining folk concepts, they shoulder the burden of explaining how they can justify significantly revising folk moral terms. I have argued that were they to allow revising folk moral terms (e.g., moral responsibility) in such a way as to eliminate the kinds of backward-looking properties that skepticism prohibits, it would seem that they should also allow revising free will in a similar manner. In doing so, however, they would essentially be undermining the case for free will skepticism.[69]

College of Staten Island (CUNY)
stephen.morris@csi.cuny.edu

69 I am thankful to Robert Lovering, Al Mele, Gregg Caruso, Thomas Nadelhoffer, and an anonymous reviewer for providing helpful comments on earlier drafts of this paper. I am also thankful to Ish Haji and the editorial staff at the *Journal of Ethics and Social Philosophy* for providing helpful correspondence during the writing process.

REFERENCES

Buckwalter, Wesley, and John Turri. "Inability and Obligation in Moral Judgment." *PLOS ONE* 10, no. 8 (August 2015): 1–20. https://doi.org/10.1371/journal.pone.0136589.

Campbell, C. A. *On Selfhood and Godhood*. London: Allen and Unwin, 1957.

Caruso, Gregg. *Free Will and Consciousness: A Determinist Account of the Illusion of Free Will*. Lanham, MD: Lexington Books, 2012.

———. "Free Will Skepticism and Criminal Behavior: A Public Health-Quarantine Model." Southwest Philosophy Review 32, no. 1 (April 2016): 25–48.

———. "Skepticism about Moral Responsibility." *Stanford Encyclopedia of Philosophy* (January 2018). https://plato.stanford.edu/entries/skepticism-moral-responsibility.

Chituc, Vladimir, Paul Henne, Walter Sinnott-Armstrong, and Felipe De Brigard. "Blame, Not Ability, Impacts Moral 'Ought' Judgments for Impossible Actions: Toward an Empirical Refutation of 'Ought' Implies 'Can.'" *Cognition* 150 (May 2016): 20–25.

Clark, Corey, Jamie B. Luguri, Peter H. Ditto, Joshua Knobe, Azim Shariff, and Roy F. Baumeister. "Free to Punish: A Motivated Account of Free Will Belief." *Journal of Personality and Social Psychology* 106, no. 4 (April 2014): 501–13.

Clark, Tom. Exchange on Waller's *Against Moral Responsibility*. *Naturalism* (October 2012). https://www.naturalism.org/resources/book-reviews/exchange-on-wallers-against-moral-responsibility.

Clarke, Randolph. "An Argument for the Impossibility of Moral Responsibility." *Midwest Studies in Philosophy* 29 (July 2005): 13–24.

Dennett, Daniel C. *Freedom Evolves*. New York: Viking, 2003.

Double, Richard. *The Non-Reality of Free Will.* Oxford: Oxford University Press, 1991.

Everett, Jim, Corey Clark, Peter Meindl, Jamie Luguri, Brian Earp, Jesse Graham, Peter H. Ditto, and Azim Shariff. "Political Differences in Free Will Belief Are Associated with Differences in Moralization." *Journal of Personality and Social Psychology* 120, no. 2 (April 2020): 461–83.

Fehr, Ernst, and Simon Gächter. "Altruistic Punishment in Humans." *Nature* 415 (January 2002): 137–40.

Haji, Ishtiyaque. *Luck's Mischief: Obligation and Blameworthiness on a Thread.* New York: Oxford University Press, 2016.

———. *Moral Appraisability*. New York: Oxford University Press, 1998.

Hamlin, J. Kiley, Karen Wynn, Paul Bloom, and Neha Mahajan. "How Infants and Toddlers React to Antisocial Others." *Proceedings of the National*

Academy of Sciences 108, no. 50 (November 2011): 19931–36.
Harris, Sam. *Free Will.* New York: Free Press, 2012.
Honderich, Ted. *A Theory of Determinism*, vol. 1. Oxford: Clarendon Press, 1990.
Joyce, Richard. *The Evolution of Morality*. Cambridge, MA: MIT Press, 2006.
Kant, Immanuel. *Groundwork for the Metaphysics of Morals*. 1785. Translated by Mary Gregor. New York: Cambridge University Press, 1998.
———. *Critique of Practical Reason*. 1788. In *Kant's Critique of Practical Reason and Other Works on the Theory of Ethics*. Translated by T. K. Abbott. New York: Longmans, Green, and Co., 1898.
Kurthy, Miklos, Holly Lawford-Smith, and Paulo Sousa. "Does Ought Imply Can?" *PloS One* 12, no. 4 (April 2017): 1–24.
Levy, Neil. *Hard Luck: How Luck Undermines Free Will and Moral Responsibility*. New York: Oxford University Press, 2011.
McKenna, Michael. "Basically Deserved Blame and Its Value." *Journal of Ethics and Social Philosophy* 15, no. 3 (July 2019): 255–82.
———. "Ultimacy and Sweet Jane." In *Essays on Free Will and Moral Responsibility*, edited by Nick Trakakis and Daniel Cohen, 186–208. Newcastle upon Tyne, UK: Cambridge Scholars Publishing, 2008.
McKenna, Michael, and D. Justin Coates. "Compatibilism: State of the Art." *Stanford Encyclopedia of Philosophy* (Winter 2019). https://plato.stanford.edu/entries/compatibilism/supplement.html.
Mele, Alfred. "Acting Intentionally: Probing Folk Notions." In *Intentions and Intentionality: Foundations of Social Cognition*, edited by Dare Baldwin, Bertram Malle, and Louis J. Moses, 27–43. Cambridge, MA: MIT Press, 2001.
———. *Effective Intentions.* Oxford: Oxford University Press, 2009.
———. *Free Will and Luck.* Oxford: Oxford University Press, 2008.
Mill, John Stuart. *Utilitarianism*. 1863. In *Ethical Theory: Classical and Contemporary Readings*, 2nd ed., edited by Louis P. Pojman, 171–91. Belmont, CA: Wadsworth, 1995.
Morris, Stephen G. *Science and the End of Ethics.* New York: Palgrave Macmillan, 2015.
Nadelhoffer, Thomas. "The Threat of Shrinking Agency and Free Will Disillusionism." In *Conscious Will and Responsibility: A Tribute to Benjamin Libet*, edited by Walter Sinnott-Armstrong, 173–88. Oxford: Oxford University Press, 2010.
Nahmias, Eddy. "Response to Misirlisoy and Haggard and to Bjornsson and Pereboom." In *Moral Psychology: Free Will and Moral Responsibility*, edited by Walter Sinnot-Armstrong, 43–58. Cambridge, MA: MIT Press, 2014.
Nahmias, Eddy, Stephen G. Morris, Thomas Nadelhoffer, and Jason Turner. "Is Incompatibilism Intuitive?" *Philosophy and Phenomenological Research* 73,

no. 1 (August 2007): 28–53.
Nielsen, Kai. "The Compatibility of Freedom and Determinism." In *Free Will*, edited by Robert Kane, 34–46. Malden, MA: Blackwell Publishing, 1971.
Nietzsche, Friedrich. *Twilight of the Idols, or How to Philosophize with a Hammer*. London: Penguin Classics, 1889.
O'Connor, Timothy. *Persons and Causes: The Metaphysics of Free Will*. New York: Oxford University Press, 2000.
Pereboom, Derk. *Free Will, Agency, and Meaning in Life.* Oxford: Oxford University Press, 2014.
———. "Free Will Skepticism and Criminal Punishment." In *The Future of Punishment*, edited by Thomas Nadelhoffer, 49–78. Oxford: Oxford University Press, 2013.
———. "Hard Incompatibilism." In *Four Views on Free Will*, edited John M. Fischer, Robert Kane, Derk Pereboom, and Manuel Vargas, 85–125. Malden, MA: Blackwell Publishing, 2007.
———. "Hard Incompatibilism and Its Rivals." *Philosophical Studies* 144, no. 1 (May 2009): 21–33.
———. *Living without Free Will*. New York: Cambridge University Press, 2001.
———. "Response to Kane, Fischer, and Vargas." In Fischer, Kane, Pereboom, and Vargas, *Four Views on Free Will*, 191–203. 2007.
Shariff, Azim, Joshua Greene, Johan Karremans, Jamie Luguri, Corey Clark, Jonathan Schooler, Roy Baumeister, and Kathleen Vohs. "Free Will and Punishment: A Mechanistic View of Human Nature Reduces Retribution." *Psychological Science* 25, no.8 (June 2014): 1563–70.
Strawson, Galen. *Freedom and Belief*. Oxford: Oxford University Press, 1986.
———. "The Impossibility of Moral Responsibility." *Philosophical Studies* 75, no. 1 (August 1994): 5–24.
Van Inwagen, Peter. *An Essay on Free Will*. Oxford: Clarendon Press, 1983.
Vargas, Manuel. "Desert, Responsibility, and Justification: A Reply to Doris, McGreer, and Robinson." *Philosophical Studies* 172, no. 10 (March 2015): 2659–78.
Warburton, Nigel. "Just Deserts." *Aeon* (October 2018). https://aeon.co/essays/on-free-will-daniel-dennett-and-gregg-caruso-go-head-to-head.
Wolf, Susan. "The Importance of Free Will." *Mind* 90, no. 359 (July 1981): 386–405.

Journal of Ethics and Social Philosophy
VOL. 23, NO. 3 · JANUARY 2023
https://doi.org/10.26556/jesp.v23i3.1564

WHEN TO START SAVING THE PLANET?

Frank Hindriks

GLOBAL WARMING reduces crop yields, increases species extinction, and threatens the future of Pacific Island Nations.[1] Intuitively, such alarming climate harms call for immediate action. However, the claim that individuals have a duty to prevent climate harms faces two important problems. First, individuals can rarely if ever avert a climate harm on their own. Second, often too few people are willing to contribute.[2] Climate duty skeptics take the first problem to entail that individuals are never obligated to reduce their carbon footprint, unless their government forces them to.[3] A couple of considerations suggest that things are not so bleak. Individuals can in some cases *help* prevent climate harms. Furthermore, they can often *activate* others and thereby increase the number of people who are willing to contribute. But there is a further problem. The process of activating enough people takes time. This poses a threat to what I call the "urgency intuition" according to which preventive action is required soon, if not immediately.

Saving this intuition requires a new conception of the duty not to harm, or, more precisely, of its causal and epistemic preconditions. Robert Goodin, Holly Lawford-Smith, and Stephanie Collins have argued that an individual is obligated to contribute to a collective outcome only if enough others are willing to do so as well.[4] The underlying idea is that the morally desirable outcome can be brought about successfully only if there is a critical mass of willing individuals. This forms the core of what I call the "success proviso." However, as just

1 Intergovernmental Panel on Climate Change, "Global Warming of 1.5° C."

2 Mitigating climate change involves many other challenges, such as power asymmetries between the rich and the poor and the fact that most of those who will be affected have not yet been born (Gardiner, *A Perfect Moral Storm*).

3 Sinnott-Armstrong "It's Not *My* Fault"; Cripps, *Climate Change and the Moral Agent*. As I discuss in section 2, Cripps allows for such obligations in exceptional circumstances, when all possibilities for promoting collective action have been exhausted (*Climate Change and the Moral Agent*, 164).

4 Goodin, "Excused by the Unwillingness of Others?"; Lawford-Smith, "Unethical Consumption and Obligations to Signal" and "What 'We'?"; and Collins, *Group Duties*.

mentioned, often too few individuals are willing to contribute and take preventive action.[5] When this is the case, the success proviso entails that people become obligated to take preventive action only once enough others have been activated. For instance, turning off your air conditioning at night or taking a train rather than a plane will be required only once enough others are willing to do the same. Because of this, the success proviso fails to preserve the urgency intuition.

The alternative that I propose here turns on the prospect of success. In order for an individual to be obligated to take preventive action, this prospect must be good enough. By this I mean that it must be reasonably likely and suitably clear that the individual can help prevent the harm. As I argue below, this "prospect proviso" sometimes requires taking preventive action right from the start. And when others have to be activated first, this activation process might progress in such a promising manner that preventive actions are already required before it has been completed. For this to be the case, individuals must have enough reason to believe that a sufficient number of others will be activated. Strikingly, an individual will then be required to initiate preventive action already before the harm can in fact be prevented. In this way, the prospect proviso preserves the urgency intuition.

In section 1, I introduce what I call the "timing question," which concerns the time at which preventive action is required in relation to the process of activation. In section 2, I critically discuss skepticism about climate duties. And in section 3, I discuss the non-skeptical positions mentioned and argue that the prospect proviso is to be preferred to the success proviso. Finally, in section 4, I distinguish different forms of activation and discuss how they influence the requisite timing of harm prevention. In these ways, attending to the timing question helps to shed light on the scope of the duty not to harm.

1. THE TIMING QUESTION

Many climate harms are caused by greenhouse gas emissions. Those harms are closely intertwined. I will assume, however, that there are particular climate harms that can be prevented when enough people reduce their carbon footprint by a certain amount. Perhaps the sea level will rise less, such that fewer islands are submerged. Maybe a storm will be less severe, and some people who would otherwise have died will now survive. Thus, either people's livelihoods are at stake or their lives are. Because climate harms are caused by several individuals, they are collective harms.

5 Batson, *What's Wrong with Morality*; and Bandura, *Moral Disengagement.*

Preventing a collective harm is a collective action problem. Perhaps the major obstacle to solving collective action problems is that often too few individuals are willing to contribute to a solution.[6] Although they might feel pressure to act, they frequently fail to do so. Psychologists have discovered that the wider the pool of required contributors, the less concerned people tend to be, and the less inclined to act.[7] The common excuse is that others are not doing anything either. Albert Bandura captures this phenomenon of the diffusion of responsibility as follows: "When everybody is responsible, no one feels responsible."[8] Strikingly, he takes it for granted that people are responsible for collective harms.

When too few people are willing to contribute, preventing a collective harm is a two-step process.[9] The first step is to activate others and increase the number of willing individuals. This serves to form a critical mass such that the individuals can prevent the harm by combining their efforts. The second step is to take preventive action, for instance by insulating your house and buying green energy. Both activation and prevention are ways of contributing to a morally desirable outcome. However, a preventive action is a direct or unmediated contribution. In contrast, someone who activates someone else makes an indirect contribution that is mediated by the other person. Its success is contingent on whether the other person makes a direct contribution. People can have an obligation to contribute directly, indirectly, or both.[10]

Intuitively, climate harms obligate people to take preventive action soon, if not immediately. This is the urgency intuition that I mentioned in the introduction. Anne Schwenkenbecher expresses it when she observes that "demands to reduce individual GHG emissions have a strong intuitive appeal."[11] Although, or perhaps because, she does not mention time, I take her to mean that, intuitively, people should do so now. The fact that preventing a collective harm is often a two-step process presents a challenge to this intuition. Activation is often a

6 Batson, *What's Wrong with Morality*; and Bandura, *Moral Disengagement.*

7 Darley and Latané, "Bystander Intervention in Emergencies"; Batson, *What's Wrong with Morality*; Philpot et al. "Would I Be Helped?"

8 Bandura, *Moral Disengagement*, 62.

9 Hindriks, "The Duty to Join Forces."

10 To contribute to a collective outcome is to perform an action that would generate that outcome if one or more other actions were also performed. In other words, a contribution is a necessary element of a set of actions that is sufficient for the outcome (Hart and Honoré, *Causation in the Law*; Mackie, *The Cement of the Universe*; Wright, "Causation, Responsibility, Risk, Probability, Naked Statistics, and Proof").

11 Schwenkenbecher, "Is There an Obligation to Reduce One's Individual Carbon Footprint?" 170.

time-consuming process, in particular when many people are involved. This means that, if activation has to precede making a contribution, this intuition has to be given up. In light of this, I ask what I call the "timing question":

> *Timing Question* (TQ): When do individuals acquire an obligation to take preventive action? Before or after the activation process has been completed?

Suppose that individuals acquire an obligation to activate others at t_0. One answer to TQ is that they acquire the obligation to take preventive action at the same time. In this case, the activation and prevention stage overlap. Another is that they acquire it at t_1, the moment at which enough others have successfully been activated. In this case, the period during which activation is required precedes that in which people ought to make a contribution. There is, however, no guarantee that the activation process will be successful. This entails that, in contrast to the first answer, the second answer leaves open the possibility that no one ever becomes obligated to contribute.

By way of illustration, consider the following example, which I adapt from Walter Sinnott-Armstrong:[12]

> *Joyguzzlers*: A number of people living in an area occasionally drive their gas-guzzling cars for fun. Each can stop doing so and thereby reduce greenhouse gas emissions. However, the harm they cause will decrease only if all of them stop driving their cars for fun.

This example raises a number of questions. Are joyguzzlers obligated to refrain from driving their gas-guzzling cars for fun? And how, if at all, does this depend on the number of joyguzzlers who are already willing to do so? Furthermore, if too few of them are willing, do they have an obligation to activate others? And, if so, what should they do in order to activate others? Finally, if they have a duty to activate, is it permissible to continue to joyguzzle until enough are willing to stop doing so such that their combined efforts prevent a climate harm? The last one of these closely connected questions is an instance of TQ.

Because livelihoods or even lives are at stake, it seems that joyguzzlers have no time to waste and should refrain from joyguzzling immediately. But they might object that doing so is futile as long as it is not possible to prevent any harm. The idea would be that, for it to make sense for one of them to stop joyguzzling, enough others must be ready to do so as well. This in turn requires that enough have been activated such that there is a critical mass of willing

12 Sinnott-Armstrong, "It's Not *My* Fault"; see also Kingston and Sinnott-Armstrong, "What's Wrong with Joyguzzling?"

individuals. And this will take time. As it might be difficult to sway enough joyguzzlers, this may in fact take a long time. These two considerations support different answers to TQ. What is at stake immediately is whether the urgency intuition can be preserved. Ultimately, however, these issues bear on the causal and epistemic preconditions of the duty not to harm.

Two existing proposals are the Harm Proviso and the Success Proviso:

> *Harm Proviso* (HP): A harm obligates an agent only if she knows that she has control over it.

> *Success Proviso* (SP): A harm obligates an agent only if she knows that she can help prevent it.

In what follows, I argue that these two provisos are too demanding. In light of this, I defend an alternative proposal, the Prospect Proviso:

> *Prospect Proviso* (PP): A harm obligates an agent only if the prospect that she can help prevent it is good enough.

As I argue below, PP supports obligations in a plausibly wide range of cases. Furthermore, it is the only proposal that preserves the urgency intuition.

2. CLIMATE DUTY SKEPTICISM

The skeptical answer to TQ is: *never*. The underlying idea is that obligations presuppose causal control. An agent has control over an outcome exactly if she is able to bring it about and able to prevent it.[13] This requires that she can perform an action that is both necessary and sufficient for it. Causal control can plausibly be combined with the similarly restrictive epistemic requirement of knowledge. Together, these requirements form HP.

Climate harms are collective harms that can typically be prevented only if a substantial number of individuals contribute. Because of this, HP is rarely met for such harms. An individual cannot prevent such a harm on her own. Furthermore, it is not necessary that she contributes when there are others who could do so.

Even so, it is in principle possible for an individual to have control over a collective harm. This will be the case when the following two conditions are met: first, enough others have already contributed such that only one more contribution is required; second, there is no one else around who will make it. Suppose opening a vault requires each of two individuals to turn a key. One has

13 Frankfurt refers to the ability to prevent the outcome as the "principle of alternate possibilities" ("Alternate Possibilities and Moral Responsibility").

already done so. In this situation, the individual who possesses the other key has control over whether the vault is opened. However, in particular when it comes to climate harms, such situations will be rare. Climate harm prevention usually require many contributions, which means that a particular individual will hardly ever play a pivotal role. It follows that, if HP is correct, individuals will rarely if ever have a duty to take preventive action with respect to a climate harm. Thus, HP supports skepticism about individual climate duties.

In this vein, Sinnott-Armstrong argues that a single individual does not as such cause climate harms and cannot avert them.[14] It follows that joyguzzlers do not have to stop driving their gas guzzlers for fun. More generally, climate harms do not obligate individuals to take preventive action.[15] Sinnott-Armstrong goes on to argue that global warming is a problem that "governments need to fix."[16] His argument is that "global warming is such a big problem that it is not individuals who cause it or who need to fix it."[17] The underlying idea must be that governments have control over climate harms. Sinnott-Armstrong also argues that people should encourage their governments to prevent climate harms and work for political candidates who are intent on changing government policies.[18]

These two ways of mitigating his skepticism fail. First, if HP is correct, it applies to direct as well as indirect contributions. And individuals do not control the outcomes of either of these.[19] For instance, someone who works for the campaign team of a political candidate does not control whether she is elected. It follows that individuals cannot be obligated to contribute indirectly either.[20] Second, HP also applies to collective agents. And there may well be climate harms over which governments do not have control. When this is the case, HP implies that they are not required to do anything, even if they could help prevent them by collaborating with other governments. This reveals that, in order for his position to be coherent, Sinnott-Armstrong should support

14 Sinnott-Armstrong, "It's Not *My* Fault."

15 Although he is not concerned with the climate, Jackson presents an argument that has the same implication ("Group Morality").

16 Sinnott-Armstrong, "It's Not *My* Fault," 312.

17 Sinnott-Armstrong, "It's Not *My* Fault," 312.

18 For an overview of criticisms of Sinnott-Armstrong's argument, see Fragnière, "Climate Change and Individual Duties."

19 Hiller, "Climate Change and Individual Responsibility," 364–65.

20 Kingston and Sinnott-Armstrong ("What's Wrong with Joyguzzling?" 185n22) claim that Cripps (*Climate Change and the Moral Agent*) has solved this problem. However, what they need is an argument that establishes that indirect contributions meet HP. And, as I discuss below, Cripps only argues that indirect contributions are often more effective than direct contributions.

skepticism about almost all individual climate duties, whether they concern direct or indirect contributions. Furthermore, it should also extend to some climate duties of collective agents.

Elizabeth Cripps is also rather skeptical about individual preventive obligations with respect to climate.[21] However, her argument turns on effectiveness rather than control. She compares preventive actions to what she calls "promotional actions," which are indirect contributions that range from writing letters to members of Congress to running for office. Such actions are meant to promote collective action, for instance by the government. Cripps argues that such indirect contributions are often more effective than direct contributions, because they can "contribute to a stockpile of impetus for collective change."[22] And she concludes that preventive actions are required only in exceptional circumstances, when all possibilities for promoting collective action have been exhausted.[23] Unfortunately, she is not very specific about the causal and epistemic preconditions of collective obligations. Her argument presupposes a proviso that is weaker than HP. But it remains unclear on exactly which proviso she relies. The upshot is that HP entails skepticism about climate duties. Because of this, it fails to preserve the urgency intuition.[24]

3. THE PROSPECT OF SUCCESS

3.1. The Success Proviso

Individuals can sometimes help bring about a morally desirable collective outcome. Think, for instance, of helping a neighbor jump-start his car. Similarly, a paramedic might save someone's life while being assisted by an emergency medical technician who drives an ambulance. And someone might form part of a human chain that ends up saving a drowning swimmer. It appears that in

21 Cripps, *Climate Change and the Moral Agent.*

22 Cripps, *Climate Change and the Moral Agent*, 148.

23 Cripps, *Climate Change and the Moral Agent*, 164.

24 Parfit, *Reasons and Persons*, 70. When discussing the preventive obligations of collective agents, Cripps considers situations in which a collective agent who is able to prevent the harm is yet to be formed (*Climate Change and the Moral Agent*, 3, 51–57). In such situations, individuals cannot take preventive action before they have incorporated themselves. It follows that the incorporation process necessarily precedes the prevention process, irrespective of which proviso is correct. Note that in many cases no individual has control over the incorporation process. Even so, it may well be that individuals can have a duty to incorporate (Held, "Can a Random Collection of Individuals Be Morally Responsible?"; Cripps, *Climate Change and the Moral Agent*; Collins 2013, *Group Duties*). This provides another reason for rejecting HP.

such cases individuals have a duty to take preventive action. If so, HP must be mistaken. Derek Parfit rejects it in effect when he argues that "even if an act harms no one, this act may be wrong because it is one of a *set* of acts that *together* harm other people."[25]

The natural alternative is that the agent must be in a position to *help* prevent the harm. Crucially, "to help" is a success verb. Someone helps save a victim's life, for instance, only if she actually survives. Because of this, I call this alternative proposal the Success Proviso (SP).[26] An agent can help prevent a harm precisely if, given the dispositions of the others, her preventive action is sufficient for preventing the harm.[27] In contrast to HP, SP can be met for several individuals at once. It could be that your preventive action is sufficient for preventing a harm given my disposition and vice versa. When this is the case, one possible state of affairs obligates several individuals. In other words, the pending collective harm gives rise to an obligation that is at least weakly collective in that it pertains to multiple agents.[28]

Holly Lawford-Smith and Stephanie Collins give further substance to SP when they argue that an individual should take preventive action on the condition that enough others are prepared to do so.[29] Furthermore, if too few individuals are willing to do so, those that are ought to signal their conditional willingness to others. The underlying idea is that when enough have signaled their willingness, they know that the harm can be prevented by combining their efforts. At this point, SP is met, which means that they are required to take preventive action. Crucially, this entails that, in the kind of situation at issue, preventive obligations are conditional and have the following content: to take preventive action

25 See also Braham and Van Hees, "An Anatomy of Moral Responsibility"; Spiekermann, "Small Impacts and Imperceptible Effects"; Pinkert, "What If I Cannot Make a Difference (and Know It)"; and Nefsky, "How You Can Help, without Making a Difference."

26 See Parfit, *Reasons and Persons*, 77–78. Schwenkenbecher proposes a slightly weaker proviso when she argues that individuals have a duty to reduce greenhouse gas emissions only if "they know that enough other people are highly likely to act this way" ("Is There an Obligation to Reduce One's Individual Carbon Footprint?" 178). The prospect proviso that I propose below is weaker still in that it requires this probability to exceed a threshold that need not be high at all.

27 Furthermore, an individual can have an obligation to take preventive action even if someone else were ready to do so in case she would fail to fulfill it.

28 As an obligation is a forward-looking responsibility, SP supports the idea that people might bear collective responsibility with respect to the harm. Note, however, that this notion of collective responsibility is consistent with reductionism (Narveson, "Collective Responsibility").

29 Lawford-Smith, "Unethical Consumption and Obligations to Signal," 322, and "What 'We?'" 229; and Collins, *Group Duties*, 120–21.

if enough others have signaled their conditional willingness to do so as well. Thus, the obligation to signal is part of the obligation to take preventive action.

Importantly, this entails that preventive obligations come into existence only after the signaling process has been completed. And this raises the question of what reason individuals have to signal. Not the preventive obligations, because obligations cannot pertain to the past. Collins argues that preventive obligations are partly constituted by mutual commitments.[30] To make sense of this, she invokes Goodin's account of how obligations can be created by exchanging conditional commitments.[31] Each must say to the others: "I will if you will" and "I will if (you will if I will)."[32] These statements express conditional commitments that become unconditional when all parties have made both of them. Crucially, nobody is obligated to do anything prior to this point. This is unproblematic when the conditional commitments are exchanged for independent reasons, as when two individuals want to do something but need each other's assurance to make it happen. However, in the case at hand, the condition pertains to signaling and is, at the same time, meant to be constitutive of the very obligation to signal, which is incoherent. The problem is that commitments cannot be constitutive of the obligation to communicate them, because obligations cannot pertain to the past.

To solve this problem, proponents of SP could say that, when someone has a conditional obligation, she is thereby obligated to satisfy the condition.[33] In the case under consideration, this means that someone who has the conditional obligation to take preventive action thereby has the obligation to signal her willingness. This reveals that SP can be defended in a coherent manner. However, at this point, another problem surfaces. Preventive obligations become unconditional only once the signaling process has been completed. In Collins's words, signaling "often serves as a precursor to more substantive coordinating actions."[34] This reveals that SP fails to preserve the urgency intuition. Thus, the way SP answers TQ is unsatisfactory.[35]

30 Collins, *Group Duties*, 119–21.

31 Goodin, "Excused by the Unwillingness of Others?"

32 Goodin, "Excused by the Unwillingness of Others?" 24.

33 Goodin, "Excused by the Unwillingness of Others?" 23.

34 Collins, *Group Duties*, 120.

35 Collins maintains that preventive obligations presuppose group abilities (*Group Duties*, 217). This secures that the group members can generate the relevant outcomes in a robust manner. But this robustness requirement is too demanding. Consider a drowning swimmer who is rescued by a human chain that almost fell apart. Even though the rescue process was anything but robust, the individuals were obligated to help save him. It follows that group abilities are not required.

3.2. The Prospect of Success

What an individual can reasonably be expected to do in the face of a collective harm depends not on success, but on the prospect of success. This prospect has to be good enough. For this to be the case, it must be reasonably likely and suitably clear that they can thereby help prevent the harm. This is what I called the Prospect Proviso (PP) in section 1.

Just as SP, PP supports collective obligations in that a single harmful outcome can entail that several individuals have a duty to contribute to harm prevention. PP requires that there is some probability that the agent's contribution is sufficient for the outcome, given the contributions of the others. As such, it is weaker than SP, which requires sufficiency. Just as the other provisos, PP consists of a causal and an epistemic requirement. These can be analyzed and developed in more detail as follows:

Prospect: A harmful outcome obligates an agent to do *A* if and only if:

1. doing *A* sufficiently increases the probability that the harm will be prevented,
2. the probability that the harm will be prevented if the agent does *A* is high enough,
3. the agent has adequate reason to believe that conditions 1 and 2 are met, and
4. there are no defeaters.[36]

The first two conditions constitute the causal requirement. First, the agent's contribution has to increase the probability of the outcome to a non-negligible and sufficient extent. This entails that a contribution can be too insignificant to be worthwhile. Second, the overall probability of success must be high enough. It may be high enough due to what everybody else is disposed to do. But it can also be that the contribution of the agent is needed in order for the threshold to be met. If this condition is met, the risk of failure is not too high. The third condition is the epistemic requirement that the agent must have enough reason to believe that the causal requirement is met. Finally, according to the fourth condition, the agent has no excuses or justifications that defeat the obligation.

36 Prospect solves what I call the "problem of insignificant hands" (Hindriks, "The Problem of Insignificant Hands"). This is the problem of why anyone would be obligated to contribute to a morally significant outcome, even though the consequence of an isolated individual contribution is morally insignificant. Nefsky calls this "the inefficacy problem" ("Collective Harm and the Inefficacy Problem").

Each of the first three conditions features a threshold, as indicated by the phrases "high enough," "sufficiently," and "adequate." Because of this, each of the conditions is either satisfied or not, depending on whether the threshold is met. An agent has a duty of the kind at issue only if all three of the conditions are met. Furthermore, the height of these thresholds is determined by two factors: first, how harmful the outcome is, and second, how costly the required action is.[37] Importantly, harms and costs can differ between cases. But the conditions of Prospect are meant to apply to all of them. Because of this, an account that is meant to be general cannot be specified in more precise terms. I should add that, just as in SP, PP is best understood in terms of a conditional obligation. The idea is that, at t_0, individuals have the conditional obligation with the following content: to take preventive action if the prospect of success is high enough. Together with Prospect, this claim constitutes what I call the "prospect account." Finally, this account is normative. To be sure, it features objective causal and epistemic facts. However, because of the thresholds, the ultimate question is whether these facts are weighty enough to constitute an obligation. This turns on the two factors just mentioned. And it is a normative question how weighty a harm is. The same holds for how much weight should be attached to the cost of making a contribution.

To illustrate how such normative factors can make a difference, consider a child who has lost her teddy bear in a mall. The parents trace their steps, and lots of people in different places help them look for it. After looking for a considerable amount of time, they have little reason to believe that they will find it, and the probability of finding it has become rather low. At this point, the prospect proviso ceases to be met and they are no longer obligated to look for the teddy bear. Suppose, however, that the parents lose track of their *child* in the mall. In that case, even a small chance of finding the child would justify continuing the search effort. Thus, when the stakes are high, even a very small probability of success can warrant preventive action. As an example of how costs might be relevant, Christian Baatz maintains that the level to which people should reduce their greenhouse gas emissions depends on how carbon dependent they are.[38]

The scope of the prospect account is considerably larger than that of SP. Whereas SP requires that the agent be in a position to help prevent the harm, Prospect is met when the probability of her helping to prevent the harm is high enough. In other words, the agent's contribution has to be pivotal in order for SP

37 How exactly they do so depends on the normative theory with which Prospect is combined. Here I remain neutral about this.

38 Baatz, "Climate Change and Individual Duties to Reduce GHG Emissions."

to be met, whereas Prospect only requires that the risk that this is not the case is morally acceptable. In Joyguzzlers, no individual knows that by refraining from joyguzzling she will help prevent or mitigate a climate harm. However, it could be that she has adequate reason to believe that she might be. Similarly, when forming a human chain, an individual will rarely know whether her contribution is pivotal. This means that joining it is not required if SP is correct. However, the fact that it might be is sufficient for Prospect to be met.[39]

Due to the thresholds that it features, Prospect supports what I call a "prospect range." At one extreme lies the contribution of which an agent has adequate reason to believe that it makes the total probability that the harm be prevented high enough. From that point onward, any contribution will be required that increases that probability to a non-negligible and sufficient extent, as long as the agent also has adequate reason to believe that this is the case. However, at some point the additional increase that a contribution makes is too small to be worth the effort. The other extreme is formed by the contribution of which the agent has adequate reason to believe that this is the case.

But how does the prospect account answer TQ? And does it preserve the urgency intuition? As I discuss in section 4, an individual who has a preventive obligation ought to activate others if need be. The thing to appreciate is that the activation process influences the prospect of prevention. Consider a number of individuals who have successfully activated a few others. These in turn set out to mobilize yet other individuals. Now suppose that along the way it becomes likely that they will succeed in creating a critical mass of willing individuals. The prospect of success will then be good enough at some point during the activation process. Thus, at least in some cases, PP supports the following answer to TQ: individuals acquire a duty to take preventive action before the activation process has been completed.

This answer to TQ can be illuminated in terms of the following analogy. Suppose a baker has to make a wedding cake, but the wedding cake topper has not arrived yet. Because of this, she is not yet in a position to finish it. Even so, she might as well begin. She can add the topper once it arrives. This illustrates that people can sometimes have good reason to start a process that they cannot yet finish. Suppose, next, that the baker is on a tight schedule. If she does not start making the cake before the topper arrives, she will not be able to finish it in time. In this situation, the baker should start baking right away if she is to finish on time. This reveals that someone can be rationally required to start a process

39 The obligations that people might have in the human-chain example do not fall under the duty not to harm, which provides for the focus of this paper. Even so, PP can plausibly be taken to extend to it. However, a more detailed analysis of how and when it applies to the duty to benefit others must take into account that this is an imperfect duty.

before she is in a position to complete it. The arrival of the topper stands for the last member's joining the collective of the willing. Thus, the idea is that an individual can be morally required to take preventive action before enough people have been mobilized.

In these first two versions of the story, the baker has every reason to expect that the topper will arrive soon. Next consider a version in which the topper will in all likelihood be too late. The stakes for the baker are high. They include a profitable long-term arrangement that is conditional on the wedding cake being perfect, which means that it must feature a topper. Although she has little reason to expect success, the baker may still have enough reason to start baking the cake, hoping that the topper will arrive in time. This illustrates that, in order for it to be prudent for the baker to start baking the cake, the prospect of success has to be good enough. Furthermore, it reveals that what is good enough depends in part on what is at stake. I propose that, also in this respect, what is morally required is analogous to what is rationally required.[40]

This supports the idea that what is required is not success but the prospect thereof. And the prospect of prevention can be good enough prior to the arrival of the "topper," that is, before the activation process has been completed. Thus, someone can be obligated to take preventive action already before enough others have been activated. To make this more precise, assume that the activation process is completed at t_2. At that point, enough individuals have been mobilized to prevent the harm. As before, the individuals acquire the duty to activate at t_0. If SP were correct, people would never become obligated to take preventive action before t_2. I have argued, in effect, that the prospect to prevent the harm successfully can be good enough already at t_1, after t_0 and before t_2. This means that the prevention stage sometimes overlaps with the activation stage. Because of this difference, Prospect is to be preferred to SP.

The question that remains is how soon after the activation process has started an individual acquires a preventive obligation. The human chain example suggests that this could happen sooner rather than later, which would mean that the period between t_0 and t_1 is short. But there may be other cases in which it is long. Furthermore, in order to fully accommodate the urgency intuition, there must be cases in which people have preventive obligations already at t_0. To determine whether this is possible, I go on to investigate the relation between activation and prevention.

40 This reveals that collective obligations depend on what is feasible (Hindriks, "The Problem of Insignificant Hands"). According to Wiens, what is feasible is a function of what is possible in the circumstances ("Political Ideals and the Feasibility Frontier"). This in turn is influenced by skills, resources, and (other) external conditions, including history (Jensen, "The Limits of Practical Possibility*").

4. ACTIVATION

When a harm is collective, obligations to prevent it are conditional. Their content is: to take preventive action if the prospect of success is high enough. As Goodin argues, someone who has a conditional obligation is thereby obligated to satisfy the condition.[41] Now, activation can increase the prospect of prevention. It follows that someone who has a conditional preventive obligation may be obligated to activate others. For this to be the case, the prospect of activation must be good enough.[42] The next thing to appreciate is that activation and prevention are not always independent processes. In order to be successful, an activator typically has to practice what she preaches. This insight forms the key to saving the urgency intuition, or so I argue in section 4.2. But first I explain in more detail what activating someone entails and how it is done (section 4.1). In section 4.3, I briefly discuss how the proposal generalizes to situations in which people have a temptation to freeride.[43]

4.1. Signaling, Persuasion, and Moralization

To activate someone is to make it the case that he is willing to contribute to the cause, either unconditionally or conditionally. Three important ways of activating others are: signaling, persuasion, and moralization.[44] Signaling is, in this context, a matter of indicating that one is willing to take preventive action. People can do so, for instance, by signing an online petition, wearing a printed T-shirt, or boycotting an unethically produced product.[45] Because of its communicative function, a signal is meant to contribute to the satisfaction of the epistemic condition. It gives those who pick up on it reason to believe that more people are willing to contribute than they thought before. Now, it could be that, because of the signal, Prospect becomes satisfied, which means that the relevant individuals become obligated to take preventive action. Furthermore, a signal can also inspire others who did not want to contribute at first

41 Goodin, "Excused by the Unwillingness of Others?" 23.

42 This entails that, if the prospect of activation is not good enough, people will not even have conditional preventive obligations.

43 Young touches on similar issues when she discusses the idea that people can be obligated to form or join a collective ("Responsibility and Global Justice," and *Responsibility for Justice*). But she does not address the question of activation directly.

44 In principle, activation can also be done by means of manipulation or coercion. I set these possibilities aside here because they raise moral concerns of their own, which makes addressing them too complicated at this stage. For the same reason, I assume that activators are sincere when they communicate.

45 Lawford-Smith, "Unethical Consumption and Obligations to Signal," 322, 325.

to change their mind. For instance, a consumer boycott can gain momentum when more and more people learn about it. Thus, signaling can contribute to the satisfaction of the causal condition.

As I discuss at greater length elsewhere, activation can also proceed by means of persuasion.[46] This requires reaching out to someone and communicating with her. What is distinctive of persuasion is that the activator presents (apparent) pros and cons and engages with the person to be activated. The paradigmatic context of activation by persuasion is that of a mutually respectful conversation. One person talks to another and tries to get the other to support the cause. He attempts to convince, persuade, or entice her to do so. And he listens and responds when the other person objects. As I discuss shortly, it will often be important that the activator expresses his support for the cause in the process.

Finally, activation can also proceed by moralizing the activity that contributes to the harm. This serves to delegitimize it or make it less attractive in other ways. Think, for instance, of how eating meat and smoking have been or are being moralized.[47] The moralization process can involve signaling and persuasion. However, what is distinctive about it is that it involves creating a new norm. So-called first movers or norm entrepreneurs, who have a strong moral identity, take the initiative to do so.[48] They embrace a norm that proscribes the harmful activity. This means that they set an example and comply with it. Furthermore, they approve of those who do so as well and disapprove of those who do not. In these ways, individuals signal their support for the norm.[49] As just indicated, they might also try to convince others by means of arguments. Crucially, successful moralizers get others to adopt the norm too. They might become convinced by the arguments, or they might be concerned with what others think if they do not follow suit. A third option is that they simply discover that enough others are willing to do what it takes.

Communication plays an important role in activation. Signaling just *is* a matter of communicating willingness to take preventive action. And persuasion and moralization are often ineffective if the activator does not convey her willingness to do so.[50] But such willingness might be conditional. And this leaves

46 Hindriks, "The Duty to Join Forces."

47 Rozin, "The Process of Moralization."

48 Van Zomeren, Postmes, and Spears, "Toward an Integrative Social Identity Model of Collective Action"; and Bicchieri, *Norms in the Wild*. For the notion of moral identity, see Aquino and Reed, "The Self-Importance of Moral Identity."

49 Cf. Lawford-Smith, "Unethical Consumption and Obligations to Signal," 323.

50 Also, people should advertise their willingness widely, if this is possible without too much effort. By doing so, they give more people adequate reason to believe that they should

open that prevention becomes obligatory only sometime after the activation process has started. I go on to argue, however, that people should often activate others by taking preventive action. And when this is the case, prevention is obligatory soon if not immediately.

4.2. Activation by Means of Prevention

Activation typically involves communication aimed at making someone willing to take preventive action, or so I have just argued. In order to be effective, such communication must be credible, or at least credible enough. And this typically requires that the activator practices what she preaches. Because of this, effective activation often involves preventive action. Consider Joyguzzlers once again. Imagine that you are at a party and someone tries to talk you out of driving a gas-guzzling car for fun. However, you just saw this person pulling up the driveway in an SUV. You point this out to him. And he responds by saying that he will stop driving his gas guzzler as soon as enough others have become willing to do so as well. You are not impressed, let alone convinced. And you find yourself another conversation partner.

The problem is not conceptual. It is in fact perfectly coherent for an activator to express conditional willingness. Instead, the problem is practical. As a matter of fact, attempts at mobilizing others tend to be more credible when the activator expresses an unconditional commitment or has already taken preventive action. Thus, the best way to get others to stop joyguzzling may well be to stop doing so yourself. This could at least be the first step of the activation process. In cases such as this one, you activate in part by means of taking preventive action. Presumably, doing so is required only as part of the activation process and not as a preventive action.

Cripps makes a similar point in relation to promotional actions.[51] She considers situations in which taking preventive action is the best means to promoting a cause. And she argues that performing it is required only as a promotional action. In other words, it is at that point never required in its capacity of a preventive action. I disagree. Suppose that one of the joyguzzlers is a trendsetter. She knows that she has this status. And she realizes that when she changes her lifestyle, many people will follow suit. Because of her influence, she can activate many others simply by trading her gas-guzzling car for an electric car.[52] This means that for her the prospects of activation are rather good.

contribute. Thus, it helps satisfy the epistemic condition for others.

51 Cripps, *Climate Change and the Moral Agent*, 144.

52 To make the example more realistic, it can be assumed that the idea of driving an electric car had already been gaining in popularity among her neighbors, perhaps because it is such a visible way of showing that you care about the environment.

Hence, the trendsetter is obligated to buy an electric car. She thereby activates others. However, it may be that the action is also required in its capacity of a preventive action. This will be the case if she has adequate reason to believe that it is sufficiently likely that by driving an electric car she can help prevent environmental harm. If this condition is indeed satisfied, she will be obligated to take preventive action from the start. And not just because it is the most effective means to getting others to support the cause, but also because of the effect it has on greenhouse emissions.

The human-chain example discussed in section 3.2 provides another illustration of this idea. Suppose that you approach some others and say that you will go into the water if they do so as well. In principle, you could wait for a few of them to get up before you take further action. You then start forming the human chain only once you have suitable reason to believe that enough people will join. However, you could also start running toward the water hoping that others will follow. This might actually be a rather effective way of engaging them. By running toward the water, you initiate preventive action. But you also activate others. Thus, your action plays two roles. And you might be obligated to perform it under both descriptions. To begin with, it is the most effective means to activating others. Suppose, however, that you have a sense that others will follow. Given that someone's life is at stake, this could mean that the prospect of success is good enough right from the start. If so, your running toward the water is also required in its guise as a preventive action.

Thus, preventive action can play an important role as part of the activation process. Nothing signals commitment more than enacting it. Such signaling can stand on its own or be part of a process of persuasion. And it often plays a significant role in moralization. Norm entrepreneurs are so committed to the cause that they will hardly violate the relevant norm, if at all.[53] They are not concerned with what others do. Furthermore, a moralizer is unlikely to be credible as an influencer if he does not practice what he preaches. If caught, he will be perceived as a hypocrite. Because of this, moralizers better comply with the norms they advocate.

However, often the prospect of prevention is not yet high enough. If so, then preventive action is required at best as a means to activation. But the prospect of moralization might not be high enough either. In that case, no one is obligated to do anything. Even so, this is unlikely to stop norm entrepreneurs.[54] They are convinced of their actions and tend to believe that they should act

53 Bicchieri, *Norms in the Wild*.

54 Van Zomeren, Postmes, and Spears, "Toward an Integrative Social Identity Model of Collective Action"; and Bicchieri, *Norms in the Wild*.

irrespective of what others do. Strikingly, this means that the moralization process is frequently initiated by people who go above and beyond the call of duty. In other words, acts of moralization are often supererogatory. But irrespective of whether it is obligatory, effective moralization typically requires compliance, which presupposes unconditional willingness and entails preventive action.

The upshot is that the urgency intuition can be preserved. In section 4.1, I argued that preventive action is sometimes required *soon* after people acquire a duty to activate others. Here I conclude that preventive action is often required *immediately* (at t_0). In some cases, this is merely because it is an effective means to activation. In others it is also required as such. For this to be the case, the prospect of successful prevention must be good enough right from the start. Finally, sometimes activation and prevention are supererogatory rather than obligatory.

4.3. Conflicts of Interests

Thus far, I have abstracted from the temptation to freeride, which people might experience in the kind of situations at issue. Instead, I assumed that their interests align and that all they need to do is contribute to harm prevention in a coordinated fashion. However, people's interests often conflict.[55] This can be the case even when everybody supports the cause. Suppose, for instance, that more individuals are willing to contribute than needed for preventing the harm. This entails that if one or a few individuals were to refrain from contributing, the harm would still be prevented. In such a situation, many will want to be among the exceptions. Another possibility is that the costs of taking preventive action are so high that some are tempted not to contribute. These two cases illustrate that harm prevention can involve a conflict of interest. Resolving it requires people to cooperate.

A conflict of interest affects the prospect of success. People will, in all likelihood, be less inclined to believe that others will cooperate and take preventive action. Norms of cooperation can provide a solution to this problem. They can enable cooperation by changing people's motivation. Sanctions can make it less attractive to violate a norm. Furthermore, if a norm is regarded as legitimate, this can increase people's motivation to comply with it.[56] Finally, when a norm is well-established, expectations about compliance will be in place and provide individuals with the requisite assurance. In these ways, a norm can even increase the prospect of success. Thus, norms do not only serve to convince

55 Olson, *The Logic of Collective Action*; and Hardin, "The Tragedy of the Commons."

56 Bicchieri, *The Grammar of Society*; and Hindriks, "Norms that Make a Difference."

people that engaging in a particular activity is wrong but they can also motivate them to comply in spite of a temptation to freeride on the efforts of others.[57]

5. CONCLUSION

So, *when should we start saving the planet?* Given the seriousness of climate harms, there is no time to waste. The main virtue of the prospect account I have proposed here is that it does justice to this sense of urgency. It entails that we should indeed start saving the planet *soon, if not immediately.* Strikingly, one of its rivals, which revolves around HP, implies that collective harms never obligate. The reason for this is that it insists on individual control (section 2). In contrast, both SP and PP do support collective obligations. They only require that an individual can help prevent it (section 3).

The main challenge that these two views face concerns situations in which too few individuals are willing to take preventive action. I have argued that people can be obligated to activate others by signaling their willingness to them, by persuading them, or by moralizing the harmful activity. According to PP, people have activation obligations if the prospect of success is good enough (section 3.2). SP fails to account for such obligations because it takes them to be constituted by mutual commitments. Such commitments will be in place only after the activation process has been completed (section 3.1).[58]

This has consequences for the time at which preventive action is required. As SP requires success, this is the case only once a critical mass of individuals is willing to take preventive action. This entails that the activation process and prevention process are temporally distinct. However, the prospect of success

57 When some fail to cooperate, the prospect account can require people to take up the slack (see also Collins, *Group Duties*, 119). This conflicts with the fair-shares view of obligations, according to which an individual ought to do only that which would prevent the harm if everybody did it (Murphy, *Moral Demands in Nonideal Theory*). The fair-shares view also implies that people cannot have a duty to activate others. It would be unfair to require someone to put effort into getting someone else to do what she should do anyway, as this would require them to do more than their fair share. For critiques of the fair-shares view, see Johnson, "Ethical Obligations in a Tragedy of the Commons"; Baatz, "Climate Change and Individual Duties to Reduce GHG Emissions"; and Karnstein, "Putting Fairness in Its Place."

58 Instead of a constitutive role, commitments play an instrumental role in the prospect account. First, someone who is committed to performing a particular action is more likely to do so, other things being equal (Bratman, *Intention, Plans, and Practical Reason*). Second, someone who expresses this commitment thereby provides someone else a (defeasible) reason to believe that they will perform the action. Third, expressing a commitment can be conducive to activating others. Finally, commitments to norms play a central role in the moralization process, in particular when interests conflict.

can be good enough much earlier. In fact, activation and prevention can and should often go hand in hand. In such cases, people are obligated to take preventive action soon, if not immediately. It follows that only PP accounts for the urgency intuition (section 4.2).

A distinctive feature of PP is that it includes normative thresholds. Their height depends on the moral significance of the pending harm and on the burden that taking preventive action places on individuals (section 3.2). It follows that people will not have preventive obligations when the harm is small and the burden is large. Furthermore, it entails that even a rather small probability of success can be high enough to support such obligations if the harm is rather large, as in the case of climate change.

But how many individuals are obligated to take preventive action? Three salient answers are: no one, everyone, and exactly the number of individuals needed for averting the harm. If PP is correct, all of these answers are mistaken. Instead, this number depends on the circumstances. Suppose that some number of individuals is obligated to take preventive action. It may be that the prospect of success would still be good enough if circumstances change such that fewer individuals are in a position to take such action, or more for that matter (section 3.2). Thus, there is a prospect range within which individuals are required to contribute.[59]

University of Groningen
f.a.hindriks@rug.nl

REFERENCES

Aquino, Karl, and Americus Reed II. "The Self-Importance of Moral Identity." *Journal of Personality and Social Psychology* 83, no. 6 (December 2002): 1423–40.

Baatz, Christian. "Climate Change and Individual Duties to Reduce GHG Emissions." *Ethics, Policy and Environment* 17, no. 1 (2014): 1–19.

Bandura, Albert. *Moral Disengagement*. New York: Worth Publishers, 2016.

Batson, C. Daniel. *What's Wrong with Morality? A Social-Psychological Perspective*. New York: Oxford University Press, 2015.

Bicchieri, Cristina. *The Grammar of Society: The Nature and Dynamics of Social*

59 I thank Gunnar Björnsson, Stephanie Collins, Niels de Haan, Martin van Hees, Holly Lawford-Smith, Abe Roth, Kai Spiekermann, Titus Stahl, and Bill Wringe for inspiring and thought-provoking discussions on the topics discussed in this paper. I am also grateful for the constructive comments offered by a referee.

Norms. Cambridge University Press, 2006.
———. *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*. New York: Oxford University Press, 2016.
Braham, Matthew, and Martin van Hees. "An Anatomy of Moral Responsibility." *Mind* 121, no. 483 (July 2012): 601–34.
Bratman, Michael. *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press, 1987.
Collins, Stephanie. "Collectives' Duties and Collectivization Duties." *Australasian Journal of Philosophy* 91, no. 2 (2013): 231–48.
———. (2019). *Group Duties: Their Existence and Their Implications for Individuals.* Oxford: Oxford University Press.
Cripps, Elizabeth. *Climate Change and the Moral Agent: Individual Duties in an Interdependent World*. Oxford: Oxford University Press, 2013.
Darley, J. M., and B. Latané. "Bystander Intervention in Emergencies: Diffusion of Responsibility. *Journal of Personality and Social Psychology* 8, no. 4 (April 1968): 377–83.
Fragnière, Augustin. "Climate Change and Individual Duties." *Wiley Interdisciplinary Reviews: Climate Change* 7, no. 6 (2016): 798–814.
Frankfurt, Harry. "Alternate Possibilities and Moral Responsibility." *Journal of Philosophy* 66, no. 23 (December 1969): 829–39.
Gardiner, Stephen M. *A Perfect Moral Storm: The Ethical Tragedy of Climate Change.* Oxford: Oxford University Press, 2013.
Goodin, Robert E. "Excused by the Unwillingness of Others?" *Analysis* 72, no. 1 (January 2012): 18–24.
Hardin, Garrett. "The Tragedy of the Commons." *Science* 162, no. 3859 (December 1968): 1243–48.
Hart, H. L. A., and Tony Honoré. *Causation in the Law*. Oxford: Oxford University Press, 1959.
Held, Virginia. "Can a Random Collection of Individuals Be Morally Responsible?" *Journal of Philosophy* 67, no. 14 (July 1970): 471–81.
Hiller, Avram. "Climate Change and Individual Responsibility." *Monist* 94, no. 3 (July 2011): 349–68.
Hindriks, Frank. "The Duty to Join Forces: When Individuals Lack Control." *Monist* 102, no. 2 (209), 204–20.
———. "Norms that Make a Difference: Social Practices and Institutions." *Analyse & Kritik* 41, no. 1 (2019): 125–46.
———. "The Problem of Insignificant Hands." *Philosophical Studies* 179, no. 3 (March 2022): 829-54.
Intergovernmental Panel on Climate Change. "Global Warming of 1.5° C." 2018. https://www.ipcc.ch/sr15.

Jackson, Frank. "Group Morality." In *Metaphysics and Morality: Essays in Honour of J. J. C. Smart,* edited by Philip Pettit, Richard Sylvan, and Jean Norman, 91–110. Oxford: Oxford University Press, 1987.

Jensen, Mark. "The Limits of Practical Possibility*." *Journal of Political Philosophy* 17, no. 2 (June 2009): 168–84.

Johnson, Baylor L. "Ethical Obligations in a Tragedy of the Commons." *Environmental Values* 12, no. 3 (August 2003): 271–87.

Karnstein, Anja. "Putting Fairness in Its Place: Why There Is a Duty to Take Up the Slack." *Journal of Philosophy* 111, no. 11 (November 2014): 593–607.

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

Lawford-Smith, Holly. "Unethical Consumption and Obligations to Signal." *Ethics and International Affairs* 29, no. 3 (Fall 2015): 315–30.

———. "What 'We'?" *Journal of Social Ontology* 1, no. 2 (2015): 225–49.

Mackie, John Leslie. *The Cement of the Universe: A Study of Causation.* Oxford: Oxford University Press, 1974.

Murphy, Liam B. *Moral Demands in Nonideal Theory.* Oxford: Oxford University Press, 2000.

Narveson, Jan. "Collective Responsibility." *Journal of Ethics* 6, no. 2 (June 2002): 179–98.

Nefsky, Julia. "Collective Harm and the Inefficacy Problem." *Philosophy Compass* 14, no. 4 (April 2019): e12587–17.

———. "How You Can Help, Without Making a Difference." *Philosophical Studies* 174, no. 11 (November 2017): 2743–67.

Olson, Mancur. *The Logic of Collective Action*. Cambridge, MA: Harvard University Press, 1965.

Parfit, Derek. *Reasons and Persons*. Oxford: Oxford University Press, 1984.

Philpot, Richard, Lasse Suonperä Liebst, Mark Levine, Wim Bernasco, and Marie Lindegaard. "Would I Be Helped? Cross-National CCTV Footage Shows that Intervention Is the Norm in Public Conflicts." *American Psychologist* 75, no. 1 (January 2019): 66–75.

Pinkert, Felix. "What If I Cannot Make a Difference (and Know It)." *Ethics* 125, no. 4 (July 2015): 971–98.

Rozin, Paul. "The Process of Moralization." *Psychological Science* 10, no. 3 (May 1999): 218–21.

Schwenkenbecher, Anne. "Is There an Obligation to Reduce One's Individual Carbon Footprint?" *Critical Review of International Social and Political Philosophy* 17, no. 2 (2014): 168–88.

———. "Joint Duties and Global Moral Obligations." *Ratio* 26, no. 3 (September 2013): 310–28.

Sinnott-Armstrong, Walter. "It's Not *My* Fault: Global Warming and Individual Moral Obligations." In *Perspectives on Climate Change: Science, Economics, Politics, Ethics*, edited by Walter Sinnott-Armstrong and R. B. Howarth, 293–315. Bingley, UK: Emerald Group Publishing Ltd., 2005.

Spiekermann, Kai. "Small Impacts and Imperceptible Effects: Causing Harm with Others." *Midwest Studies in Philosophy* 38, no. 1 (September 2014): 75–90.

Van Zomeren, Martijn, Tom Postmes, and Russell Spears. "Toward an Integrative Social Identity Model of Collective Action: A Quantitative Research Synthesis of Three Socio-Psychological Perspectives." *Psychological Bulletin* 134, no. 4 (July 2008): 504–35.

Young, Iris Marian. "Responsibility and Global Justice: A Social Connection Model." *Social Philosophy and Policy* 23, no. 1 (January 2006): 102–30.

———. *Responsibility for Justice.* Oxford: Oxford University Press, 2011.

Wiens David. "Political Ideals and the Feasibility Frontier." *Economics and Philosophy* 31, no. 3 (November 2015): 447–77.

Wright, Richard W. "Causation, Responsibility, Risk, Probability, Naked Statistics, and Proof: Pruning the Bramble Bush by Clarifying the Concepts." *Iowa Law Review* 73 (1988): 1001–77.

Journal of Ethics and Social Philosophy
VOL. 23, NO. 3 · JANUARY 2023
https://doi.org/10.26556/jesp.v23i3.1900

NO DISRESPECT—BUT THAT ACCOUNT DOES NOT EXPLAIN WHAT IS MORALLY BAD ABOUT DISCRIMINATION

Frej Klem Thomsen

ALMOST EVERYONE agrees that paradigmatic cases of discrimination are morally bad.[1] The employer who refuses to hire women, or the police officer who arrests Black citizens while letting White citizens off with a warning for similar offenses—these figures are universally (or near enough) condemned.

Underneath this consensus, however, lies a series of further questions where unanimity rapidly evaporates. For example, what *exactly* is discrimination? When should discrimination be legally prohibited? And, perhaps most important, *why* is discrimination morally bad (when it is)?

These questions have attracted increased philosophical attention over the past decade, resulting in a rapidly expanding literature.[2] Among answers to the question of what makes discrimination morally bad (when it is), two accounts in particular stand out. The first, harm-based account holds (roughly) that discrimination is morally bad when and to the extent that it brings about harm to the discriminatee or others.[3] The second, disrespect-based account

1 I use moral badness here to denote the quality of there being a *pro tanto* (moral) reason against an action. Since such reasons are defeasible, an action that is morally bad need not be morally wrong all things considered. Cf. Lippert-Rasmussen, "The Badness of Discrimination."

2 See for example Collins and Khaitan, *Foundations of Indirect Discrimination Law*; Eidelson, *Discrimination and Disrespect*; Hellman, *When Is Discrimination Wrong?*; Hellman and Moreau, *Philosophical Foundations of Discrimination Law*; Khaitan, *A Theory of Discrimination Law*; Lippert-Rasmussen, *Born Free and Equal?* and *Routledge Handbook of the Ethics of Discrimination*; and Moreau, *Faces of Inequality.* For overviews see Altman, "Discrimination"; and Thomsen, "Discrimination."

3 See Arneson, "Discrimination and Harm"; Lippert-Rasmussen, "Private Discrimination," "The Badness of Discrimination," and *Born Free and Equal?*; Berndt Rasmussen, "Harm and Discrimination"; Ishida, "What Makes Discrimination Morally Wrong?"; and Thomsen, "Stealing Bread and Sleeping Beneath Bridges."

holds (roughly) that discrimination is morally bad when and to the extent that it is disrespectful.[4]

Few will deny that causing harm is morally bad, and there are obvious ways in which discrimination can bring about harm, e.g., through offense, stigmatization, or the imposition of avoidable, unjust disadvantage. As such, even proponents of alternative accounts tend to acknowledge that one way in which discrimination can be morally bad is that it causes harm.[5] This is compatible with what we have said of the disrespect-based account so far—"when and to the extent" defines an entailment, and the proponent need claim only that disrespect is sufficient for moral badness, not that it is necessary. Arguably, then, the most defensible version of the disrespect-based account claims only that the harm-based account does not *exhaust* the ways in which discrimination can be morally bad, since discrimination can *also* be morally bad when and because it is disrespectful.[6]

Although one of the most prominent accounts of what makes discrimination morally bad, it seems to me both that the disrespect-based account remains underdeveloped, and that upon reflection it faces objections so powerful that ultimately we ought to abandon it. This article attempts first to provide some clarification of how we can best understand the disrespect-based account, and thereupon to present and develop the objections that jointly show why it should be abandoned.

4 See Alexander, "What Makes Wrongful Discrimination Wrong?"; Beeghly, "Discrimination and Disrespect"; Eidelson, *Discrimination and Disrespect*; and Glasgow, "Racism as Disrespect." The account at stake is different from the related account that holds (roughly) that discrimination is morally bad when and because it *expresses* disrespect of ("demeans," in Deborah Hellman's phrasing) the discriminatee. See Hellman, *When Is Discrimination Wrong?* and "Discrimination and Social Meaning"; and Shin, "The Substantive Principle of Equal Treatment." I intend to set the expressive disrespect account aside. It bears mentioning, however, that there are what seem to me overwhelmingly strong arguments against that account. See Arneson, "Discrimination, Disparate Impact, and Theories of Justice," 91–94; Eidelson, *Discrimination and Disrespect*, 85–90; Ekins, "Equal Protection and Social Meaning"; and Lippert-Rasmussen, *Born Free and Equal?* ch. 5.

5 See Alexander, "What Makes Wrongful Discrimination Wrong?"; Eidelson, *Discrimination and Disrespect*; Slavny and Parr, "Harmless Discrimination"; cf. Beeghly, "Discrimination and Disrespect," 89.

6 In combination with the assumption that what we are looking for is an account of moral badness, this leads to the view summarized by Richard Arneson: "There are wrong-making characteristics of discrimination, such that if an act of discrimination embodies any of these characteristics, its doing so is a pro tanto consideration against its moral permissibility. . . . These characteristics can be outweighed by countervailing factors, and whether a given act of discrimination is wrong, all things considered, depends on the overall balance of considerations" ("Discrimination, Disparate Impact, and Theories of Justice," 103).

Section 1 clarifies the disrespect-based account by making precise the meaning of disrespect and disrespectful discrimination. Section 2 introduces the first challenge, in the shape of the competing thesis that disrespectful discrimination speaks to the moral character and blameworthiness of the agent. Section 3 sketches a powerful objection launched by Kasper Lippert-Rasmussen, which shows disrespectful discrimination to be intuitively no worse than respectful discrimination, and demonstrates that the objection can be applied to the version of the disrespect-based account developed in section 1. Section 4 adds the objection that disrespect appears to provide the intuitively wrong answer in cases of "right actions for the wrong reasons," specifically by condemning at least some cases of disrespectful nondiscrimination. Section 5 confronts an argument advanced by Adam Slavny and Tom Parr that there are cases of intuitively bad harmless discrimination, and argues that our intuitions about such cases can be explained without reference to the disrespect-based account. Section 6 summarizes and concludes with some perspectives on the implications of abandoning the disrespect-based account for our understanding of discrimination specifically and moral theory more generally.

1. WHAT IS THE DISRESPECT-BASED ACCOUNT OF MORALLY BAD DISCRIMINATION?

Let us assume for the purposes of this article a direct, generic, descriptive definition of discrimination (loosely) based on Lippert-Rasmussen's work: an agent *A* discriminates against persons with property *P* iff:

1. *A* treats persons with *P* differently than she treats (or would treat) persons without,
2. *A*'s treatment of persons with *P* is disadvantageous as compared with her treatment of persons without, and
3. the difference in treatment is suitably explained by the fact that persons do and do not possess *P* (or that *A* believes this to be the case).[7]

The definition is direct in that it concerns standard cases of differential treatment, not cases where treatment that does not differentiate on the basis of *P* nonetheless results in disparate impact.[8] It is generic, in that it does not delimit discrimination to differential treatment of a particular set of properties, such

7 Cf. Eidelson, *Discrimination and Disrespect*; Hellman, *When Is Discrimination Wrong?*; Lippert-Rasmussen, "The Badness of Discrimination"; Thomsen, "But Some Groups Are More Equal than Others" and "Direct Discrimination"; and Moreau, *Faces of Inequality.*

8 Thomsen, "Direct Discrimination." Cf. Thomsen, "Stealing Bread and Sleeping Beneath Bridges"; Cosette-Lefebvre, "Direct and Indirect Discrimination"; Doyle, "Direct

as gender, race, ethnicity, religion, sexuality, disability, and/or age, or the properties that are in the appropriate context "socially salient."[9] It is descriptive in that it does not require that an act be morally bad, not even *prima facie*, for it to qualify as discrimination.[10]

I do not want to claim that this is the "right way" to define discrimination, in part because I am not persuaded that there is one right way to define discrimination. It seems to me more true to say that we sometimes speak of discrimination in the sense I give it here, and at other times in narrower senses that restrict it along one of the parameters I have noted above, e.g., discrimination that targets socially salient groups specifically, or discrimination that is at least *prima facie* morally bad. This diversity of conceptions makes stipulating the sense at stake helpful, and this particular, simple definition will make certain points easy to state. However, nothing in the argument of this article hinges on the stipulated definition; we could, I think, make the same points, only somewhat more cumbersomely, while employing any reasonable alternative definition.

The question at the heart of moral analysis of discrimination is this: What might make an act of discrimination (as defined above) morally bad? And the answer we want to discuss is the *disrespect-based account*:

> *Disrespect-Based Account*: Discrimination is morally bad when and to the extent that it is disrespectful.

There are variations on this account in the literature on the ethics of discrimination. In his seminal piece, Larry Alexander argues: "When a person is judged incorrectly to be of lesser moral worth and is treated accordingly, that treatment is morally wrong regardless of the gravity of its effects. It represents a failure to show the moral respect due the recipient, a failure which is by itself sufficient to be judged immoral."[11]

Similarly, in a piece on the definition and moral badness of racism, Joshua Glasgow argues that racial differentiation becomes morally bad racism "if and only if [the act or policy] is racially disrespectful."[12]

Discrimination, Indirect Discrimination, and Autonomy"; Khaitan, "Indirect Discrimination"; Lippert-Rasmussen, "Indirect Discrimination Is Not Necessarily Unjust."

9 Thomsen, "But Some Groups Are More Equal than Others"; cf. Lippert-Rasmussen, *Born Free and Equal?*

10 Cf. Lippert-Rasmussen, *Born Free and Equal?*; and Eidelson, *Discrimination and Disrespect*.

11 Alexander, "What Makes Wrongful Discrimination Wrong?" 159. It is worth noting that Alexander has since rejected the disrespect-based account. See Alexander, "Is Wrongful Discrimination Really Wrong?"

12 Glasgow, "Racism as Disrespect," 81. While Glasgow's analysis focuses on racism, I believe Lippert-Rasmussen is right to suggest that it is sympathetic to Glasgow's work to extend

Finally, in arguably the most sophisticated development of the disrespect-based account, Benjamin Eidelson writes that "acts of discrimination are intrinsically wrong when and because they manifest a failure to show the discriminatees the respect that is due to them as persons."[13]

Stated in such general terms, the disrespect-based account requires clarification. Specifically, we need to know what *precisely* disrespect is, as well as what it means for an act of discrimination to be disrespectful. Only once we have filled out these details can we evaluate whether the account is plausible.

1.1. What Is Disrespect?

Answers to the first question generally focus on how the agent responds to the moral status of the discriminatee. To be disrespectful, Glasgow suggests, is "something like a failure to adequately recognize autonomous, independent, sensitive, morally significant creatures."[14] Eidelson defines respect in light of his "interest thesis": "To respect a person's equal value relative to other persons one must value her interests equally with those of other persons, absent good reason for discounting them."[15]

Alexander's phrasing, particularly in comparison with the just-cited passages by Glasgow and Eidelson, illustrate two possible ways of understanding disrespect. On one interpretation, disrespect consists in the discriminator having a particular mental state related to the moral status of the discriminatee, such as the discriminator *judging* or *believing* that the discriminatee has lower moral status.[16]

On a different interpretation, disrespect need not consist in the agent having any particular offending mental state. Disrespect, Eidelson suggests, arises "not simply by the presence of some positive factor of animus or a defamatory belief, but by the *absence* of appropriate recognition of someone's personhood."[17] On this interpretation, disrespect can consist in the mere failure to have a required mental state related to moral status.

---

it from racism to potentially applying to other groups and forms of discrimination. See Lippert-Rasmussen, *Born Free and Equal?* 116–17.

13 Eidelson, *Discrimination and Disrespect*, 73. The disrespect-based account of morally bad discrimination can draw on broader theories of morally bad disrespect. As Eidelson makes explicit, the notion of (dis)respect at stake is similar and indebted to the notion of recognition respect developed by Stephen Darwall, which requires that agents "take seriously and weigh appropriately the fact that [other persons] are persons in deliberating about what to do" ("Two Kinds of Respect," 38). Cf. Frankfurt, "Equality and Respect."

14 Glasgow, "Racism as Disrespect," 85.

15 Eidelson, *Discrimination and Disrespect*, 97

16 Alexander, "What Makes Wrongful Discrimination Wrong?" Cf. Arneson, "What Is Wrongful Discrimination?"; and Beeghly, "Discrimination and Disrespect," 85.

17 Eidelson, *Discrimination and Disrespect*, 75. Cf. Beeghly, "Discrimination and Disrespect," 86.

Between the two, the latter, Eidelsonian conception of disrespect is the more powerful version of the account. It can include cases where the discriminator holds an offending mental state, on the grounds that these explain how disrespect is brought about, e.g., that the presence of a false belief about lower moral status causes the agent to fail to adequately recognize the discriminatee's moral status. However, unlike the first of the two conceptions, it can also include cases where the agent fails to recognize moral status in spite of having no such offending mental state.[18]

Moral status, in turn, might be interpreted in different ways. It might pertain, for example, to interests, autonomy, virtues, or desert. For present purposes, I shall assume that we are speaking of disrespect as it pertains to interests.[19]

Furthermore, one can assume the Kantian view that all persons and only persons have equal moral status, or the (arguably more plausible view) that there can be differences in moral status and that it is not restricted to persons.[20] Between these two alternatives, Eidelson appears to favor the former approach, while Alexander favors the latter.

Finally, *lower* moral status is a relative term, and as such might mean lower absolutely—lower than the discriminatee actually has, or lower comparatively, that is, lower than the group that is treated differently.

Table 1. Disrespect

Mental state is …	Presence of offending state	Absence of required state
Moral property is …	Interests	Autonomy, desert, virtues, etc.
Actual status is …	Equal (Kantian)	Varied
Lower than …	Absolutely	Comparatively

Even restricting our attention to interests, there are thus eight possible variants of the disrespect-based account. I will suggest below that some versions are more attractive than others, but also that all versions face very serious challenges.

1.2. What Is Disrespectful Discrimination?

Before discussing the challenges, we must address the second issue of what it means for discrimination to be disrespectful, that is, what role must disrespect play in relation to discrimination for the action to be disrespectful? Let

18 Cf. Eidelson, *Discrimination and Disrespect*, 98–99; Lippert-Rasmussen, "Respect and Discrimination," 324–25.

19 Eidelson extensively discusses disrespect that does not adequately recognize a person's autonomy. I set aside here separate treatment of this version mostly due to constraints of space, but it seems to me that the challenges I present below will (with suitable adjustments) affect other versions. However, for focused critical discussion of disrespect of autonomy, see Lippert-Rasmussen, "Respect and Discrimination" and *Born Free and Equal?*

20 See Lippert-Rasmussen, *Born Free and Equal?* 119–20, 124–25.

us review three possible answers. The first of these ties disrespect to beliefs about moral status:

> *Epistemic Background*: Discrimination is disrespectful if the discriminator holds a false belief about the lower moral status of the discriminatee, or if she does not hold a true belief about the moral status of the discriminatee.[21]

Epistemic background is vulnerable to two objections. First, many cases of what we might intuitively want to label disrespectful discrimination appear to be compatible with the discriminator holding true beliefs about the equal moral status of the discriminatees because, again, such beliefs need not prevent the discriminator from, e.g., giving less weight to the interests of the discriminatees.[22] Consider:

> *Friedrich Wilhelm*: FW accurately believes that men and women have equal moral status. However, his repressed neurotic shame at his own sexuality makes him loathe and fear the objects of his attraction. As a result of these feelings, he often fails to adequately recognize women's moral status when acting in spite of his beliefs.

Second, it seems implausible that an action becomes disrespectful because of the presence or absence of a belief even when that belief is causally inert, that is, if the presence or absence of the belief in no way affects the discriminator's actions.[23] Consider:

> *Statistics*: Agents *A* and *B* discriminate in identical fashion against members of a group for statistical reasons. *A* holds a true belief about the equal moral status of the discriminatees. *B* holds a false belief about the lower moral status of the discriminatees. The beliefs in no way affect the actions of either agent.

It seems very strange to say that *B*'s discrimination is disrespectful while *A*'s discrimination is respectful (supposing that there are no other differences between *A* and *B* and their actions than the difference in beliefs). Plausibly, both are disrespectful if they both fail to adequately recognize the moral status of persons from the group at stake, and disrespectful if the opposite.[24]

21 Either version can further require that the belief be conscious in the discriminator's mind, but this makes no difference to the challenges that epistemic background faces.

22 Cf. Lippert-Rasmussen, *Born Free and Equal?* 116.

23 See Lippert-Rasmussen, *Born Free and Equal?* 126 and "Respect and Discrimination," 325.

24 In the latter case, discrimination might still be morally permissible—perhaps the statistical reasons are valid and sufficient to outweigh the interests of the discriminatees—but the

As a different suggestion, some might say that discrimination is disrespectful when it treats the discriminatee as if she had lower moral status in the sense that the agent discriminates although there are reasons grounded in the discriminatee's moral status that count against the permissibility of the action. Call this:

> *Contrary to Reasons*: Discrimination is disrespectful if the discrimination is contrary to reasons grounded in the discriminatee's moral status.

There are passages in Eidelson's work where he appears to lean in this direction. Thus, Eidelson claims, "one acts disrespectfully . . . by failing to act *on* the reasons that would be given by recognition respect."[25] One problem for this version is that it seems clear that there can be situations where the reasons grounded in a person's moral status that count against an act are outweighed by other reasons. It sounds strange to say that an agent who carries out the (permissible) act in such cases is being disrespectful, particularly if we suppose that she is conscious of and gives accurate weight to the reasons grounded in moral status. Second, on this version of the account, disrespect presupposes and appears to add nothing to an independent account of the relevant reasons. Or as Lippert-Rasmussen puts it: "the suspicion is that respect turns out to be parasitic on a prior account of what these moral requirements are."[26] As such, we cannot use disrespect to explain the moral badness of discrimination, since it is only possible to determine whether an act is disrespectful once we have established whether it is for independent reasons, in a certain respect, morally bad. Third, even more so than in Epistemic Background, the mental state of the discriminator plays no part. She is disrespectful simply by virtue of acting contrary to certain reasons, regardless of how and why she does so.

We can apply the lessons learned from the failures of the first two suggestions to state a more plausible understanding of disrespectful discrimination. A common thrust of the objections above is that for discrimination to be disrespectful it must *be based upon* disrespect. The cases where the presence or absence of relevant beliefs intuitively makes an action disrespectful are cases where this affects what the agent does.[27] And the cases where acting contrary

---

issue at stake here is only whether the discrimination is *disrespectful*.

25 Eidelson, *Discrimination and Disrespect*, 78, emphasis added.

26 Lippert-Rasmussen, *Born Free and Equal?* 117. Cf. Beeghly, "Discrimination and Disrespect," 92–95; and Pettit, "Consequentialism and Respect for Persons."

27 Cf. Lippert-Rasmussen, *Born Free and Equal?* 119: "Accordingly, an act can be based on an assumption about the moral worth of the affected individual if, and only if, this act is somehow motivated by the actor's judgment of the individual's moral worth."

to reasons is intuitively disrespectful are cases where the agent does not give these reasons appropriate weight.

The third suggestion thus places greater emphasis on the agent's decision-making, to hold that discrimination is disrespectful not merely when it is contrary to reasons grounded in the discriminatee's moral status, but when the discriminator does not act *for* these reasons.[28] Specifically, the disrespect-based account can assume:

> *Responsive to Reasons*: Discrimination is disrespectful of the discriminatee if the agent gives reasons grounded in the moral status of the discriminatee lower weight in her decision making.

This seems to me the most attractive of the three suggestions, and I shall assume in the following that it is the understanding of what it means to be disrespectful at stake in the disrespect-based account.

1.3. The Baseline for Lower Moral Status

We must consider one final issue before turning to the challenges: the choice of baseline for lower moral status. Consider perhaps the two most obvious suggestions, an absolute and a comparative baseline.

> *Absolute Baseline*: Discrimination is disrespectful if the discriminator gives reasons grounded in the discriminatee's moral status *lower weight than these reasons actually have*.

> *Comparative Baseline*: Discrimination is disrespectful if the discriminator gives reasons grounded in the discriminatee's moral status *lower weight than she gives to the reasons grounded in the moral status of non-discriminatees*.

Each of these baselines has certain disadvantages.

The main disadvantage for the absolute baseline is that it rules out labeling discrimination as disrespectful of the discriminatee in scenarios where the discrimination is comparatively disrespectful while respectful of the discriminatee according to the absolute baseline. Consider:

28 Eidelson writes that "failure to recognize someone as a person of equal value as others *may* be expressed in a belief or cognitive judgment that has a misestimate of her value as its content. Whatever you believe, however, the interest thesis implies that respecting someone as a being of equal value also entails *responding* to her status as a bearer of interests with presumptively equal normative weight. And to act consistently with what that presumption requires—to actually succeed in respecting it—it is not enough to reason in good faith. Your deliberation and action must actually track the relevant moral facts" (*Discrimination and Disrespect*, 103).

> *Brahmin and Dahlit*: Employers 1 and 2 both consistently favor members of group *B* over members of group *D* in hiring. Employer 1 does so because she considers *D*-persons to be morally unworthy, and assigns the reasons grounded in their interests less than their actual weight, while she considers *B*-persons to be morally worthy, and assigns the reasons grounded in their interests their actual weight. Employer 2 does so because she considers *B*-persons to be morally super-worthy, and assigns their interests far greater than their actual weight, while she considers *D*-persons to be morally worthy, and assigns the reasons grounded in their interests their actual weight.

Those attracted to the disrespect-based account will presumably want to say that the two employers' discrimination is equally disrespectful of *D*-persons. The absolute baseline precludes drawing this conclusion because employer 2 does not give the reasons grounded in the moral status of *D*-persons lower than their actual weight. The comparative baseline avoids this issue, because both employers give lower weight to the reasons grounded in the moral status of *D*-persons than to the reasons grounded in the moral status of *B*-persons.

The comparative baseline, however, has the disadvantage that it entails labeling discrimination as disrespectful of discriminatees in scenarios where the discriminator gives different weight to reasons grounded in moral status because the reasons have different weight. Suppose that nonhuman animals have lower moral status than humans, but that many nonhuman animals, including all vertebrates, do have moral status.[29] Consider:

> *Babies and Parrots*: A team of firefighters attempts to rescue inhabitants from a burning house. Each firefighter can carry either a caged parrot or a baby out of the house. Firefighters assign the actual weight to reasons grounded in the interests of babies and parrots, respectively. As a result, the firefighters all rescue babies.[30]

It sounds absurd to say that the firefighters are disrespectful of parrots—surely they ought to grant every set of reasons exactly the weight to which it is entitled—yet that is what the comparative baseline entails.[31]

29 This challenge to the comparative baseline is easily overlooked if one assumes the Kantian view that all persons and only persons have equal moral status. The assumption that many non-human animals have moral status seems to me obviously true. However, even Kantians should be willing to admit that the mere conceptual possibility of nonpersons with higher or lower moral status makes the disadvantages of the comparative baseline apparent.

30 Cf. Kagan, *The Limits of Morality*, 16.

31 Note that as the comparative baseline avoided the first disadvantage, so the absolute baseline avoids this particular problem.

In light of the disadvantages, neither baseline appears satisfactory. A possible solution is to adopt a combination of the two in the shape of the Comparative Ratio of Actual to Given Weight as baseline:

> *Comparative Ratio of Actual to Given Weight*: Discrimination is disrespectful if the discriminator gives the reasons grounded in the discriminatee's moral status lower weight relative to their actual weight *as compared to the weight relative to actual weight she gives to the reasons grounded in the moral status of non-discriminatees*.

We can abbreviate this to say that the disrespectful discriminator *discounts* some status-based reasons but not others, or that she employs different discount rates for different status-based reasons.[32] This allows the employers to be equally disrespectful in Brahmin and Dahlit, and the firefighters to avoid being disrespectful in Babies and Parrots. Perhaps there are disadvantages to this suggestion in turn, but I will assume for the purposes of the subsequent discussion that it is the sense of "giving lower weight" at stake in the disrespect-based account.

This completes our review of the disrespect-based account of morally bad discrimination. In the next four sections, I will present three challenges to the account and critically discuss a recent argument in favor of it. Sadly, after all our efforts at detailing it, the analysis in these sections supports the conclusion that we should abandon the disrespect-based account of morally bad discrimination.

2. WEAK VS. STRONG DISRESPECT

The first challenge to the disrespect-based account of discrimination stems from the similarity of two theses. The disrespect-based account, as I have reviewed it above, subscribes to what we can call the strong disrespect thesis:

> *Strong Disrespect Thesis*: Disrespect is morally relevant in the sense that there is a *pro tanto* reason against an action when that action is disrespectful.[33]

Compare:

32 The discriminator could employ a negative discount rate, which would *magnify* the weight of reasons. In such cases, it remains disrespectful to discount reasons at different rates such that the weight of one type of reason is overestimated relative to the other. For simplicity, I shall assume we are discussing examples of a positive discount rate.

33 Cf. Lippert-Rasmussen, *Born Free and Equal?* 160, 173; Eidelson, *Discrimination and Disrespect*, 80–84.

> *Weak Disrespect Thesis*: Disrespect is morally relevant in the sense that it reflects poorly on the agent's character, and/or makes her liable to blame when the agent's action is disrespectful.[34]

The distinction between these differing ideas of how mental states are or might be morally relevant is, of course, familiar from broader debates within moral philosophy, in part due to Thomas Scanlon's influential work.[35] Regardless of one's views on the broader issue, the weak disrespect thesis seems to me very plausible. Clearly, it is also possible consistently to hold that both the weak disrespect thesis and strong disrespect thesis are true. However, the combination of the weak thesis' plausibility and similarity to the strong thesis puts obstacles in the path of arguing for the disrespect-based account.

To illustrate these obstacles, consider how we might interpret disrespect according to the weak disrespect thesis in the light of different background conditions, i.e., conditions that explain *why* the agent is disrespectful. Specifically, consider what we might say of an agent who gives lower weight to someone's interests in her decision making (i) while holding a true versus while holding a false belief about moral status, and (ii) while justifiably versus unjustifiably holding a belief about moral status. The concept of justified belief is, of course, notoriously difficult, but let us say for present purposes (very loosely) that an agent justifiably believe that *P* iff the agent believes that *P* because she has reasoned about the evidence for *P* in an epistemically responsible manner. If we assess what these different possibilities mean for how disrespect speaks to the agent's moral character and blameworthiness, there is, it seems to me, a natural hierarchy of sins.[36]

For a start, consider an agent who discounts status-based reasons because she holds the false but justified belief that the relevant beings have lower moral status. Such an agent might be said simply to be unfortunate. Suppose, for example, that the agent lives in a cultural and scientific environment in which available evidence supports the belief that fish have no moral status, thinks carefully about this evidence, and draws the reasonable conclusion that fish have no moral status. Suppose also (as seems to me very plausible) that this belief is false. If the agent discriminates against fish, she will do so disrespectfully

34 Cf. Lippert-Rasmussen, *Born Free and Equal?* 124. I do not mean to presuppose any particular theoretical commitments about the moral role of blame, but it is worth noting that even consequentialists partial to the harm-based account could accept the weak disrespect thesis and follow the present analysis, on a suitable account of the moral role of blame (e.g., Arneson, "The Smart Theory of Moral Responsibility and Desert").

35 See Scanlon, *Moral Dimensions*.

36 I do not mean for this analysis to be comprehensive; I intend only to illustrate a point by covering certain of the most interesting possibilities.

on the account developed above, but she does not display an objectionable moral character, nor does holding her belief in any uncontroversial way make her liable to blame.

The situation is different for an agent who discounts status-based reasons because she holds the false and *unjustified* belief that the relevant beings have lower moral status. If she reasons in a way that is defective but unbiased, then we can reasonably blame her for her careless reasoning; however, it is simply bad luck that she happened to arrive at this particular false belief.[37] If her reasoning process is defective in a way that systematically distorts beliefs in a particular way, e.g., because she employs motivated reasoning to shape negative beliefs about a certain group to fit her animosity toward them, then deriving this particular false belief is not merely unfortunate. In such cases, we might reasonably blame her to a greater degree, and say that both her animosity and her proclivity for motivated reasoning reflect poorly on her character.[38]

We can also imagine an agent who discounts status-based reasons through sheer negligence, that is, because she omits to entertain the pertinent reasons at all. The agent might, let us suppose, decide too hastily or while distracted. In so doing, we might say that she displays an objectionable recklessness in reasoning, and she is presumably liable to blame, perhaps to roughly the same extent as the careless reasoner above.

Finally, we can imagine an agent who discounts status-based reasons *in spite* of holding and being conscious of the belief that the relevant being has equal or higher moral status. Eidelson suggests, in the context of his analysis of the strong thesis, that such disrespect is a form of contempt.[39] Plausibly, in some paradigmatic cases of racism or misogyny the discriminator is well aware that discriminatees have equal moral status, but nonetheless consciously and deliberately discounts the weight of reasons grounded in their interests, for example because of animosity toward them. Intuitively, and to the extent that we can

37 Interestingly, on plausible theories of moral luck, we might want to say something similar about an agent who gives *equal* weight to someone's interests based on a true but unjustified belief. See Nagel, *Mortal Questions*, ch. 3; Williams, *Moral Luck*; and Zimmerman, "Luck and Moral Responsibility."

38 Such biased belief formation plausibly occurs in many cases of, e.g., racists and misogynists. As Larry Alexander notes about the related process of generating biased beliefs about other properties: "One who realizes that his biases cannot be justified on their own terms, such as one who realizes the invalidity of his judgment that blacks are inherently morally inferior, may, rather than relinquish the judgment fully, merely replace it with a belief that blacks very frequently have trait *X*, trait *X* being a perfectly respectable basis for discrimination. Thus, many irrational proxies are the products of bias-driven tastes for certain erroneous beliefs" ("What Makes Wrongful Discrimination Wrong?" 170).

39 See Eidelson, *Discrimination and Disrespect*, 106.

meaningfully rank such things, this strikes me as the type of disrespect that reflects most poorly on the agent's character and makes her most liable to blame.

As is evident from even this cursory analysis, the weak disrespect thesis allows a nuanced moral evaluation of disrespect. Furthermore, it is able to track several differences that proponents of the strong disrespect thesis claim are relevant, as in the difference between disrespect based on biased and merely unfortunate false beliefs, and negligent versus contemptuous disrespect.[40] This symmetry means that, although the theses are not incompatible, they are often in competition. Specifically, it is or at least often will be possible to explain our moral intuitions about cases with reference to both one and the other. This places a tall stumbling block in the path of arguments for the disrespect-based account, which relies on the strong thesis. When an argument for the account relies on intuitions about disrespect, the proponent must establish that the intuition is at least in part attributable to the factors at stake in the strong thesis, rather than deriving simply from the weak thesis. Barring such clarification, the intuition cannot count as evidence for the strong thesis specifically because it is possible that the intuition is tracking the moral relevance of disrespect in the sense stated by the weak thesis.

3. Disrespectful Discrimination Can Be at Least No Worse

The most sophisticated argument against the disrespect-based account of morally bad discrimination in the literature is Kasper Lippert-Rasmussen's demonstration that there are cases of disrespectful discrimination that are intuitively at least no worse than otherwise identical cases of respectful discrimination.[41] Although developed in great detail by Lippert-Rasmussen, it seems to me worthwhile rehearsing it here, in part because the force of the challenge appears not to have been fully appreciated, and in part to show its applicability to the analysis of disrespect set out above.[42]

Lippert-Rasmussen's argument is (roughly) the following:

40 On the former, see Alexander, "What Makes Wrongful Discrimination Wrong?"; and Arneson, "What Is Wrongful Discrimination?" On the latter, see Eidelson, *Discrimination and Disrespect*.

41 See Lippert-Rasmussen, "The Badness of Discrimination," "Intentions and Discrimination in Hiring," *Born Free and Equal?* and "Respect and Discrimination."

42 Richard Arneson ("Discrimination, Disparate Impact, and Theories of Justice") does not discuss it in his critical review of deontological accounts of morally bad discrimination, Adam Slavny and Tom Parr ("Harmless Discrimination") make no mention of the challenge in their recent argument for the disrespect-based account, and Erin Beeghly ("Discrimination and Disrespect") does not discuss it in her reference article on the account.

1. All else equal, the presence of a wrong-making factor makes an action intuitively morally worse.
2. There are cases where the presence of disrespect, leaving all else equal, does not intuitively make discrimination morally worse.

C. Disrespect is not a wrong-making factor for discrimination.

The first premise presupposes that intuition is generally capable of tracking moral differences, but this is widely accepted in applied ethics. The argument is valid, such that if the premises are true, so is the conclusion. This leaves the second premise: Are there cases where all else equal the presence of disrespect does not make discrimination intuitively worse?

Lippert-Rasmussen advances a first set of cases against the version of the disrespect-based account associated with Larry Alexander, where disrespect is based on a false belief that the discriminatee has lower moral status. In this set, two persons both conduct painful experiments on animals to provide a small benefit to humans. The *inegalitarian experimenter* justifiably holds the false belief that animals have lower moral status, while the *egalitarian experimenter* justifiably holds the true belief that animals have equal moral status. As Lippert-Rasmussen observes:

> If Alexander's account is correct, the inegalitarian experimenter acts in a way that is disrespectful—he harms animals on the basis of his false belief about the unequal moral status of animals and human beings—unlike the egalitarian experimenter, who holds true beliefs about the comparative moral status of animals and human beings.... However, intuitively, *if* there is a difference in terms of wrongfulness between the two acts of experimentation, the case involving what I stipulated to be true—egalitarian beliefs about moral status—is morally more wrong.[43]

Benjamin Eidelson objects to this set of cases that both experimenters equally fail to give appropriate weight in their decision making to the interests of animals: "Lippert-Rasmussen's attempt at a controlled comparison ... fails if the relevant judgment is understood as constituted by taking certain considerations as reasons for certain kinds of acts, rather than as simply a propositional attitude."[44] *Pace* Lippert-Rasmussen's intention, Eidelson claims, the two cases do not differ in that only one involves disrespect.

If the two cases are equally disrespectful, how does Eidelson explain the intuition that, if anything, the egalitarian experimenter acts *worse*? Eidelson argues that the experimenter who holds the true belief that animals have equal

43 Lippert-Rasmussen, "Respect and Discrimination," 321

44 Eidelson, *Discrimination and Disrespect*, 104.

moral status evinces a particularly egregious form of disrespect, "contempt," which explains our intuition that her discrimination may be morally worse.[45]

In response, Lippert-Rasmussen has shown that there are comparison cases where contempt does not make disrespectful discrimination morally worse. Consider this (lightly rephrased) version:

> *Roses:* Red and White both perform painful experiments on persons. Each is motivated primarily by conformist reasons, but justifiably holds the false belief that Yorks have lower moral status than Lancasters. Red experiments only on Yorks, in line with her beliefs, while White experiments only on Lancasters, in contravention of her beliefs.

Lippert-Rasmussen concludes: "In Eidelson's sense, both agents disrespect the individuals on whom they experiment, since both experimenters fail to give proper weight in their deliberations to the value, as perceived by them, of those persons they experiment on. . . . Only the [latter] case involves contempt. Yet it is unclear that the [latter] case is more wrongful than the [former]."[46]

Interestingly, there is an apparently promising response, which abandons Eidelson's idea that contempt affects permissibility in favor of the weak thesis.[47] The intuitive difference in the first set of cases is explained, on this response, by the fact that, although equally disrespectful, the egalitarian experimenter displays a morally worse character and is more liable to blame. The intuitive similarity in *Roses*, by the fact that while White's action is contemptuous, it is not based on a disrespectful belief about the discriminatee (White's belief is disrespectful of Yorks, not of the Lancasters on whom she experiments). Thus, White and Red might be intuitively (roughly) equally blameworthy.

Can we extend Lippert-Rasmussen's line of argument to cover the disrespect-based account in combination with the weak thesis? I believe we can. Consider:

> *Speciesist Scientist*: A very serious disease affects many humans but no other animals. Researchers *A* and *B* both want to perform painful and dangerous tests for a potential cure. The cure can be tested equally well on either human volunteers or lab rats. The benefits of the potential cure are such that in spite of the pain and risk it would be morally permissible

45 See Eidelson, *Discrimination and Disrespect*, 105–7.

46 Lippert-Rasmussen, "Respect and Discrimination," 328–29.

47 Lippert-Rasmussen briefly discusses this possibility in the context of a related challenge, that our intuitions about the weak thesis "drown out" our intuitions about the strong thesis ("Respect and Discrimination," 322–23).

to test it on human volunteers. Nonetheless, because rats have lower moral status than humans, both choose to test on rats.

Compare:

Disrespect: Researcher *A* discounts the reasons grounded in the interests of rats.

No Disrespect: Researcher *B* does not discount the reasons grounded in the interests of rats.[48]

Intuitively, researcher *A*'s discrimination against rats is not morally worse than researcher *B*'s. If there is any difference between the two, it seems to concern the factors at stake in the weak disrespect thesis. Presumably, *A* is liable to some blame for giving lower weight to the reasons grounded in the interests of rats.

A possible objection is that we cannot explain why both researchers would choose to experiment on rats when one gives lower and the other equal weight to the reasons grounded in their interests. This is mistaken. Since rats *actually* have lower moral status than humans, the *actual* balance of reasons to which researcher *B* is responding may favor experimenting on rats. This touches upon a different challenge, which we consider next: Does the disrespect-based account allow that agents can do right for the wrong reasons?

4. Can Discrimination Not Be Right for the Wrong Reasons?

The third challenge for the disrespect-based account of morally bad discrimination concerns the counterintuitive implication that intuitively permissible actions can become wrong simply by virtue of the malignant mental state of the agent.[49] We can bring the challenge into focus by comparing a trio of cases. Consider:

Study Group 1: Adam is a student who is considering whether to invite his fellow students Fatima and Christopher to form a study group. As an extrovert, Adam has no problem forming the group, but his fellow students are shy introverts, who would not form a group without his initiative. Forming a group will benefit all students included. Fatima is Arabic, while Christopher is Caucasian. Because Fatima is Arabic, Adam

48 Recall that on the baseline we have adopted, for researcher *B* to give equal weight to the interests of rats does not mean that she holds their interests to be equal to human interests or to ground equally strong reasons (which would contradict their lower moral status).

49 Arneson, "Discrimination and Harm," 157–58; and Lippert-Rasmussen, *Born Free and Equal?* 126. Cf. Parfit, *On What Matters*, 1:216.

> gives the benefit to Fatima of joining the group less weight than the comparable benefit to Christopher. The difference in weights causes Adam to invite Christopher, but to not invite Fatima.[50]

On the disrespect-based account, Adam's discrimination of Fatima is morally bad *because* it is disrespectful. This is true even if it would not be morally bad for Adam not to form the group at all.[51]

Compare this with a similar case of respectful equal treatment:

> *Study Group 2*: As Study Group 1, except that Adam gives equal weight to benefits to Arabic persons and Caucasian persons. Furthermore, Adam enjoys socializing with Arabic persons. Therefore, Adam invites both Fatima and Christopher to join the group.

Intuitively, Adam's actions in Study Group 2 are morally benign. Perhaps the most obvious difference between the two cases is that Adam does not discriminate against Fatima, but the disrespect-based account entails that another important difference is that Adam does not give lower weight to Fatima's interests. Meanwhile, the introduction of a preference for socializing with Arabic persons does not intuitively affect permissibility, even if this preference is one reason why Adam invites Fatima. This is important, because we can now reintroduce disrespect without varying the other factors. Consider:

> *Study Group 3*: As Study Group 2, except that because Fatima is Arabic, Adam gives the benefit to Fatima of joining less weight than the comparable benefit to Christopher. However, the lower weight is exactly balanced by his preference for socializing with Arabic persons, such that Fatima's probability of being invited to join is the same as if she had been Caucasian. Therefore, Adam invites both Fatima and Christopher.

In Study Group 3, Adam is (by stipulation) as disrespectful of Fatima as in Study Group 1, in that he equally discounts benefit to her because of her ethnicity. If the presence of disrespect makes an action *pro tanto* morally bad, then Study Group 3 is as bad as Study Group 1 in the specific dimension of disrespect. Yet, intuitively, this is not the case. Adam's inviting Fatima in Study Group 3 seems to me not merely better than his action in Study Group 1, which could be explained by the fact that Fatima is disadvantaged in the former case, but to

50 The case is loosely based on a case discussed by Eidelson, *Discrimination and Disrespect*, 96–97.

51 Consequentialists will conclude that since the group provides only benefits, Adam is obligated to form the group (unless there is an even better action alternative), but friends of the disrespect-based account are likely to think doing so is supererogatory.

be not in any respect morally bad. Study Group 3 is rather a case of doing the right thing for the wrong reasons, which is to say that it is an action that is not in any particular respect morally bad, but where we might nonetheless find fault with the agent's character and decision making.[52] This again suggests that we should adopt the weak disrespect thesis, which holds only that disrespect is relevant to moral assessment of the agent, but not the strong disrespect thesis, which holds that disrespect is relevant to the permissibility of the action.

5. IS DISRESPECTFUL HARMLESS DISCRIMINATION INTUITIVELY MORALLY BAD?

Above, we considered three challenges to the disrespect-based account of morally bad discrimination. In this penultimate section, we critically review a recent argument in favor of it, in order to show that it does not support the account.

The argument is due to Adam Slavny and Tom Parr, who present a series of cases that are meant to provide intuitive support for the disrespect-based account by showing that harmless disrespectful discrimination can be morally bad.[53] This is an important challenge. Much of the work for friends of the harm-based account consists in showing how apparently harmless, morally bad discrimination is either actually harmful or actually not morally bad (although perhaps discrimination that we have harm-based reason to prohibit or support a norm against).[54]

The most compelling case, developed after considering some possible objections, is:

> *Cambridge University 3* (CU3): Helen is an admissions officer at Cambridge University. As a result of her racist prejudices, she is averse to spending time around students with dark skin tone. Having read Kasper Lippert-Rasmussen's *Born Free and Equal?* she believes that it would be wrong for her to harm these applicants, so she uses her connections to

52 It may also be worth noting that the present argument avoids a counter presented by Tom Parr against a related argument by Richard Arneson. Parr claims that disrespect only affects permissibility when the agent's actions affect the target of disrespect. This condition is not satisfied in Arneson's case, where a spiteful philosopher stabs a Justin Bieber voodoo doll, because this in no way affects the unwitting Justin Bieber, but is satisfied in the Study Group cases. Parr, "Revisiting Harmless Discrimination," 2–3. Cf. Arneson, "Discrimination and Harm," 157.

53 Slavny and Parr, "Harmless Discrimination"; and Parr, "Revisiting Harmless Discrimination."

54 See, e.g., Arneson, "Discrimination, Disparate Impact, and Theories of Justice"; and Thomsen, "Iudicium ex Machinae" and "The Art of the Unseen."

> ensure that qualified dark-skinned applicants are also offered a place at Oxford. (The places Helen secures for these students are *additional* ones such that no one else is denied a place at Oxford as a result of Helen's actions.) Applicants prefer Oxford to Cambridge, and they would not have received an offer from Oxford but for Helen's intervention.[55]

CU3 is constructed so as to ensure that Helen's actions are harmless, indeed even beneficial to the dark-skinned students, on any plausible account of harm. Slavny and Parr believe that "despite benefiting the applicants, Helen's actions remain wrongful. Although there may be differing explanations for this wrongfulness, the most promising is that Helen's actions are wrong because they are motivated by the desire not to spend time around dark-skinned students."[56] According to Slavny and Parr, then, CU3 establishes both that the harm-based account does not explain all cases of morally bad discrimination, and that there are cases of discrimination that are morally bad *because* of the discriminator's disrespect for the discriminatee.

The first and most immediate challenge for CU3 is that it is not clear that it need involve disrespect.[57] On the face of it, Helen's discrimination is best understood as based on a brute desire not to be around dark-skinned persons. On the disrespect-based account, as I set it out in section 2, desires are not themselves respectful or disrespectful.[58] Disrespect is a matter of what weight the agent gives to reasons grounded in moral status, not of what the agent likes, prefers, or wants. Even desires for or against sharing the company of certain persons need not lead to or be accompanied by disrespect. If I strongly dislike racists and posh snobs, for example, I might prefer to avoid their company, but I need not (I hope) give lower weight to reasons grounded in their moral status. To circumvent this issue, let us suppose that CU3 is a case of genuine disrespect, that is, that Helen's preference against associating with dark-skinned students is accompanied by, perhaps causally connected with, giving the reasons grounded in their moral status lower weight than she gives reasons grounded in the moral status of light-skinned students.

55 Slavny and Parr, "Harmless Discrimination," 109. I have here reconstructed the case, integrating parts that the authors present in discussing the first and second versions of it.

56 Slavny and Parr, "Harmless Discrimination," 109.

57 It is also not a case of discrimination *against* dark-skinned applicants on the definition I have adopted, but a case of discrimination *in favor* of dark-skinned applicants. This, I take it, is only a terminological issue, since I have not assumed and do not think that there is a moral asymmetry between discrimination against and discrimination in favor of.

58 Cf. Eidelson, *Discrimination and Disrespect*, 115–26.

I have three more serious concerns with CU3, however, all of which pertain to the presence of potentially confounding factors. The first is that, in spite of Slavny and Parr's efforts to construct the case so as to avoid it, Helen's discrimination might be harmful. Thus, we might think that increased racial segregation can have bad aggregate effects. In the most extreme example, it seems reasonable to suppose that an all-light-skinned Cambridge and an all-dark-skinned Oxford would create or reinforce racial schisms, even if the educations they offer are equally good. A related concern is the risk of causing offense. Recipients of the offers, sensing the underlying motive, may reasonably feel hurt and humiliated. We can eliminate the first of these potential confounders by altering the scenario to avoid any increase in racial segregation, e.g., by supposing that barring Helen's discrimination, dark-skinned students would be underrepresented at Oxford and overrepresented at Cambridge. However, it seems to me difficult to alter the scenario so as to reduce the risk of offending dark-skinned applicants without introducing deception, which might itself affect our intuitive response to the scenario.

The second confounding factor is the violation of the norms of the admissions system. I suspect that intuitions might be affected by the notion that Helen's duties as an admissions officer require her to set aside any and all personal preferences. Thus, we might find it similarly intuitively troubling if she gave weight to other, more idiosyncratic desires, such as the desire not to be around persons whose names begin with a consonant, even if we suppose that she in no way holds such persons to have different moral status or gives less weight to reasons grounded in their moral status.[59] These professional duties might in turn be related to or based upon a meritocratic norm, which many find intuitively appealing in the context of admissions to higher education. The meritocratic norm, substituting "position" for "job" in David Miller's formulation, is that "justice demands that the [position] be offered to the best-qualified applicant. We express this by saying that the best-qualified applicant deserves the [position], or, in a slightly different formulation, that the principle involved is one

59 Slavny and Parr briefly consider an objection along these lines, and reject it with reference to a sketched case involving a millionaire donating selectively to white persons, but not Black persons (see Slavny and Parr, "Harmless Discrimination," 111). The problem with this response is, of course, that discrimination here is not harmless. Black persons suffer real costs, in the shape of being deprived of benefits they otherwise would have received, from the millionaire's differential treatment. They also note that the claim that the case involves a violation of professional duties is compatible with the claim that the case involves morally bad disrespect. The problem with this response is that the objection does not deny the compatibility of these claims. It simply points out that since our intuitions about the case could be caused by either of the moral factors, these intuitions cannot be taken to support the disrespect-based account.

of merit."[60] Note that the meritocratic norm is both different from the strong disrespect thesis and not itself a plausible account of what makes discrimination morally bad.[61] It is also worth noting that there are powerful arguments against the meritocratic norm as a principle of justice.[62] Nonetheless, its intuitive appeal is likely to affect our response to CU3.

Third, I think it is indisputable that the factors identified by the weak disrespect thesis affect our intuitions about CU3. We can confidently say of Helen's actions that they reflect her morally bad character, and we can criticize that character, e.g., by blaming Helen for her racist prejudice. I suspect that it is very difficult to tell to what extent our intuition in CU3 is triggered by the factors at stake in the weak and the strong disrespect thesis, respectively.

This might suggest that we are at an impasse. Our intuition is plausibly affected by confounding factors, but it could also be triggered by disrespect. How do we tell whether it is one or the other? One way is to compare CU3 with other scenarios. Consider:

> *Cambridge University 4* (CU4): Like CU3, except that Helen has no racial prejudice, and does not give lower weight to reasons grounded in the moral status of dark-skinned students. Instead, her offer to dark-skinned applicants is based on her having made a drunken bet with friends that she could subvert the admissions process along racial lines without being discovered.

CU4 is like CU3 in that Helen risks causing racial segregation and offense, that she fails to respect her professional duties and the meritocratic norm of the admissions system, and that we can criticize her moral character. However, she does not give lower weight to reasons grounded in the moral status of dark-skinned students. In fact, we can assume that her careful construction of a beneficial offer is made because she gives their interests exactly the same weight as the interests of light-skinned students, and is genuinely concerned to ensure that they are no worse off for her actions.[63] In spite of this, the two cases seem to me intuitively very similar, such that removing disrespect from the scenario has made no discernible difference.

60 Miller, *Principles of Social Justice*, 156.

61 See Lippert-Rasmussen, *Born Free and Equal?* 108–9

62 See Segall, "Should the Best Qualified Be Appointed?"

63 The same point applies if we adopt one of the alternative versions of the disrespect-based account discussed in section 2. For example, it does not appear to me to make any intuitive difference to the moral permissibility of Helen's actions whether we suppose that she holds racist beliefs about differences in moral status or an irresponsible willingness to shirk her professional duties to win a bet.

6. CONCLUSION

In the course of this article, I have attempted to clarify the disrespect-based account of discrimination, only to argue that it faces challenges so severe it seems reasonable to conclude that we should abandon it.

Disrespectful discrimination, I have argued, is perhaps most appealingly understood as discrimination where the discriminator gives less weight to reasons grounded in the discriminatees' moral status, compared to their actual weight, than she does to reasons grounded in the moral status of non-discriminatees. This version of the account avoids problems plaguing versions that focus on the discriminator's beliefs or the reasons at stake, or that adopt the absolute or comparative baselines.

However, arguments for the disrespect-based account face a serious obstacle in that intuitions that might support it can often be equally or more plausibly explained by reference to the fact that disrespect reflects poorly on the moral character of the discriminator (the weak disrespect thesis). Simultaneously, there are cases of disrespectful discrimination that are intuitively no worse than respectful discrimination, and cases of disrespectful nondiscrimination that are not intuitively morally bad because of disrespect. Both types suggest that disrespect does not make actions morally bad.

Finally, I reviewed an argument by Adam Slavny and Tom Parr that attempted to show that there are cases of intuitively morally bad harmless discrimination, where the moral badness can best be explained by disrespect. I argued that, in line with the preceding analysis, intuitions about these cases can better be explained by the presence of confounding factors.

It is worth addressing one final point. Where does abandoning the disrespect-based account leave the ethics of discrimination specifically and the debate on the moral relevance of mental states more generally?

For the ethics of discrimination, deontologists need not despair. Although it is often interpreted as such, the harm-based account of discrimination is not consequentialist.[64] And there remain alternatives to both the disrespect- and harm-based accounts, such as luck egalitarian or liberal accounts.[65]

64 Moreau and Slavny and Parr are just two examples of authors who insist on associating the harm-based account with consequentialism. Friends of consequentialism might hope as much. Given the intuitive importance of harm doing, it would constitute a decisive blow to deontology if only consequentialism could account for its moral relevance. Clearly, however, this is not the case. See Moreau, *Faces of Inequality*; Slavny and Parr, "Harmless Discrimination." Cf. Arneson, "Discrimination and Harm."

65 See Segall, "What's So Bad about Discrimination?"; Knight, "Discrimination and Equality of Opportunity"; and Moreau, *Faces of Inequality*.

The situation is broadly the same with respect to the broader debate. Slavny and Parr argue that arguments for and against the strong disrespect thesis have ties to broader debates such that commitments to deontological accounts of the role of mental states in determining moral permissibility have implications for how we should assess the strong disrespect thesis, and conversely that abandoning the disrespect-based account should be resisted because doing so would weaken the general case for mental states affecting permissibility.[66] Both claims are mistaken.

The second claim is dangerously close to a fallacy *ad consequentiam*. "So much the worse for the general case for mental states affecting permissibility," one might say. Indeed, those unimpressed with general arguments for the claim that mental states have any such role might consider any such negative implications of abandoning the disrespect-based account a feature, not a bug.

While tempting, this response would be misguided. There is no immediate reason why deontologists committed to affirming the claim that mental states affect moral permissibility need to accept the disrespect-based account, and denying it does not conflict with either the general claim or popular specific theories.

Consider for illustration probably the most widely debated version of a theory that mental states affect moral permissibility: the intention principle, which is at the heart of the doctrine of double effect (DDE). The intention principle can be stated in different ways, but one way that fits our purposes here is to say that an action can be morally worse when and because it is performed with a bad intention.[67]

Clearly, the intention principle is not the disrespect-based account, nor does either entail the other. Consider, for example, cases of intentional and unintentional indirect discrimination.[68] A prospective employer might employ a hiring procedure that disproportionately disfavors women. She might do so without

66 Slavny and Parr, "Harmless Discrimination."

67 See FitzPatrick, "The Doctrine of Double Effect"; and Liao, "Intentions and Moral Permissibility."

68 Some draw the distinction between direct and indirect discrimination on the basis of intentions (or, perhaps, a slightly broader set of mental states). See Altman, "Discrimination." On this way of drawing the distinction, there is no such thing as intentional indirect discrimination. This seems to me an unhelpful way of distinguishing the cases we tend to label direct and indirect discrimination. I prefer to draw the distinction depending on whether the discriminator differentially or equally treats persons, in the sense of employing the relevant property as a distinguishing criterion for performing different actions. See Thomsen, "Stealing Bread and Sleeping Beneath Bridges." This is compatible with the discriminator directly discriminating in deciding to employ a particular decision procedure, which is itself only indirectly discriminatory. Cf. Eidelson, *Discrimination and Disrespect*, 41–45.

intending to indirectly discriminate against this group, or she might do so while intending this discrimination. Importantly, however, even intentional discrimination against the group need not involve disrespect. She might, for example, believe (let us assume, correctly) that the company's profits will increase as a result of the discrimination, and consider the discrimination an instrument to this goal, while holding members of the group to have equal moral worth. In this case, according to the intention principle, the moral status of the discrimination might vary between the intentional and unintentional cases, without varying in terms of disrespect. Thus, whatever theoretical commitments one might have to the general idea that mental states can affect the moral permissibility of actions, they are not *necessarily* challenged by the arguments against the strong disrespect thesis specifically.[69] The disrespect-based account of morally bad discrimination stands—or, more plausibly, falls—on its own.[70]

Danish National Centre for Ethics
fkt@dketik.dk

REFERENCES

Alexander, Larry. "Is Wrongful Discrimination Really Wrong?" San Diego Legal Studies Paper no. 17-257 (May 15, 2016). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2909277.

———. "What Makes Wrongful Discrimination Wrong? Biases, Preferences, Stereotypes and Proxies." *University of Pennsylvania Law Review* 141, no. 1 (1992): 149–219.

Altman, Andrew. "Discrimination." *Stanford Encyclopedia of Philosophy* (Winter 2020). https://plato.stanford.edu/entries/discrimination.

Arneson, Richard J. "Discrimination and Harm." In Lippert-Rasmussen,

69 There are what seem to me very persuasive arguments against DDE and the intention principle, such that we have reasons unrelated to the failure of the disrespect-based account to reject both. See Nelkin and Rickless, "So Close, Yet So Far"; Scanlon, *Moral Dimensions*; Steinhoff, "The Secret to the Success of the Doctrine of Double Effect" and "Wild Goose Chase"; and Thomson, "Physician-Assisted Suicide."

70 My work on this article has benefitted from discussing it at seminars with the CEPDISC Centre of Excellence at Aarhus University and the Research Group for Criminal Justice Ethics at Roskilde University. I owe thanks for very valuable comments and discussion on these occasions to Didde Boisen Andersen, Benjamin Eidelson, Sebastian Holmen, Søren Flinch Midtgaard, Viki Møller Lyngby Pedersen, Thomas Søbirk Petersen, Kasper Lippert-Rasmussen, and Jesper Ryberg. The research was conducted as part of a one-year visit to CEPDISC, generously funded by the Danish National Research Foundation (DNRF144).

Routledge Handbook of the Ethics of Discrimination, 151–63.
———. "Discrimination, Disparate Impact, and Theories of Justice." In Hellman and Moreau, *Philosophical Foundations of Discrimination Law*, 87–111.
———. "The Smart Theory of Moral Responsibility and Desert." In *Desert and Justice*, edited by Serena Olsaretti, 233–58. Oxford: Clarendon Press, 2007.
———. "What Is Wrongful Discrimination?" *San Diego Law Review* 43, no. 4 (2006): 775–808.
Beeghly, Erin. "Discrimination and Disrespect." In Lippert-Rasmussen, *Routledge Handbook of the Ethics of Discrimination*, 83–96.
Berndt Rasmussen, Katharina. "Harm and Discrimination." *Ethical Theory and Moral Practice* 22, no. 4 (August 2019): 873–91.
Collins, Hugh, and Tanurabh Khaitan, eds. *Foundations of Indirect Discrimination Law*. Oxford: Hart Publishing, 2018.
Cosette-Lefebvre, Hugo. "Direct and Indirect Discrimination." *Public Affairs Quarterly* 34, no. 4 (October 2020): 340–67.
Darwall, Stephen L. "Two Kinds of Respect." *Ethics* 88, no. 1 (October 1977): 36–49.
Doyle, Oran. "Direct Discrimination, Indirect Discrimination, and Autonomy." *Oxford Journal of Legal Studies* 27, no. 3 (Autumn 2007): 537–53.
Eidelson, Benjamin. *Discrimination and Disrespect*. Oxford: Oxford University Press, 2015.
Ekins, Richard. "Equal Protection and Social Meaning." *American Journal of Jurisprudence* 57, no. 1 (2012): 21–48.
FitzPatrick, William. J. "The Doctrine of Double Effect: Intention and Permissibility." *Philosophy Compass* 7, no. 3 (March 2012): 183–96.
Frankfurt, Harry. "Equality and Respect." *Social Research* 64, no. 1 (Spring 1997): 3–15.
Glasgow, Joshua. "Racism as Disrespect." *Ethics* 120, no. 1 (October 2009): 64–93.
Hellman, Deborah. "Discrimination and Social Meaning." In Lippert-Rasmussen, *Routledge Handbook of the Ethics of Discrimination*, 97–107.
———. *When Is Discrimination Wrong?* Cambridge: Harvard University Press, 2009.
Hellman, Deborah, and Sophia Moreau, eds. *Philosophical Foundations of Discrimination Law*. Oxford: Oxford University Press, 2013.
Ishida, Shu. "What Makes Discrimination Morally Wrong? A Harm-Based View Reconsidered." *Theoria* 87, no. 2 (April 2021): 483–99.
Kagan, Shelly. *The Limits of Morality*. Oxford: Oxford University Press, 1991.
Khaitan, Tarunabh. "Indirect Discrimination." In Lippert-Rasmussen, *Routledge Handbook of the Ethics of Discrimination*, 30–41.

———. *A Theory of Discrimination Law*. Oxford: Oxford University Press, 2015.
Knight, Carl. "Discrimination and Equality of Opportunity." In Lippert-Rasmussen, *Routledge Handbook of the Ethics of Discrimination*, 140–50.
Liao, S. Matthew. "Intentions and Moral Permissibility: The Case of Acting Permissibly with Bad Intentions." *Law and Philosophy* 31, no. 6 (November 2012): 703–24.
Lippert-Rasmussen, Kasper. "The Badness of Discrimination." *Ethical Theory and Moral Practice* 9, no. 2 (April 2006): 167–85.
———. *Born Free and Equal? A Philosophical Inquiry into the Nature of Discrimination*. Oxford: Oxford University Press, 2013.
———. "Indirect Discrimination Is Not Necessarily Unjust." *Journal of Practical Ethics* 2, no. 2 (December 2014): 33–57.
———. "Intentions and Discrimination in Hiring." *Journal of Moral Philosophy* 9, no. 1 (January 2012): 55–74.
———. "Private Discrimination: A Prioritarian Desert-Accommodating Account." *San Diego Law Review* 43, no. 4 (2007): 817–56.
———. "Respect and Discrimination." In *Moral Puzzles and Legal Perplexities: Essays on the Influence of Larry Alexander*, edited by Heidi M. Hurd, 317–32. Cambridge: Cambridge University Press, 2018.
———, ed. *Routledge Handbook of the Ethics of Discrimination*. Abingdon, UK: Routledge, 2017.
Miller, David. *Principles of Social Justice*. Cambridge MA: Harvard University Press, 1999.
Moreau, Sophia. *Faces of Inequality: A Theory of Wrongful Discrimination*. Oxford: Oxford University Press, 2020.
Nagel, Thomas. *Mortal Questions*. Cambridge: Cambridge University Press, 1979.
Nelkin, Dana K. and Samuel C. Rickless. "So Close, Yet So Far: Why Solutions to the Closeness Problem for the Doctrine of Double Effect Fall Short." *Noûs* 49, no. 2 (June 2015): 376–409.
Parfit, David. *On What Matters*. Vol. 1. Oxford: Oxford University Press, 2011.
Parr, Tom. "Revisiting Harmless Discrimination." *Philosophia* 47, no. 5 (November 2019): 1535–38.
Pettit, Phillip. "Consequentialism and Respect for Persons." *Ethics* 100, no. 1 (October 1989): 116–26.
Scanlon, Thomas. *Moral Dimensions: Permissibility, Meaning, and Blame*. Cambridge: Belknap Press, 2008.
Segall, Shlomi. "Should the Best Qualified Be Appointed?" *Journal of Moral Philosophy* 9, no. 1 (January 2012): 31–54.
———. "What's So Bad about Discrimination?" *Utilitas* 24, no. 1 (March 2012):

82–100.
Shin, Patrick S. "The Substantive Principle of Equal Treatment." *Legal Theory* 15, no. 2 (June 2009): 149–72.
Slavny, Adam and Tom Parr. "Harmless Discrimination." *Legal Theory* 21, no. 2 (June 2015): 100–14.
Steinhoff, Uwe. "The Secret to the Success of the Doctrine of Double Effect: Biased Framing, Inadequate Methodology, and Clever Distractions." *Journal of Ethics* 22, nos. 3–4 (December 2018): 235–63.
———. "Wild Goose Chase: Still No Rationales for the Doctrine of Double Effect and Related Principles." *Criminal Law and Philosophy* 13, no. 1 (March 2019): 1–25.
Thomsen, Frej K. "The Art of the Unseen: Three Challenges for Racial Profiling." *Journal of Ethics* 15, nos. 1–2 (June 2011): 89–117.
———. "But Some Groups Are More Equal than Others: A Critical Review of the Group Criterion in the Concept of Discrimination." *Social Theory and Practice* 39, no. 1 (January 2013): 120–46.
———. "Direct Discrimination." In Lippert-Rasmussen, *Routledge Handbook of the Ethics of Discrimination*, 19–29.
———. "Discrimination." In *Oxford Research Encyclopedia of Politics*, edited by William. R. Thompson. Oxford: Oxford University Press, 2017.
———. "Iudicium ex Machinae: The Ethical Challenges of Automated Decision-Making at Sentencing." In *Sentencing and Artificial Intelligence*, edited by Jesper Ryberg and Julian V. Roberts, 252–75. Oxford: Oxford University Press, 2022.
———. "Stealing Bread and Sleeping Beneath Bridges: Indirect Discrimination as Disadvantageous Equal Treatment." *Moral Philosophy and Politics* 2, no. 2 (2015): 299–327.
Thomson, Judith J. "Physician-Assisted Suicide: Two Moral Arguments." *Ethics* 109, no. 3 (April 1999): 497–518.
Williams, Bernard. *Moral Luck: Philosophical Papers 1973–1980*. Cambridge: Cambridge University Press, 1981.
Zimmerman, Michael J. "Luck and Moral Responsibility." *Ethics* 97, no. 2 (January 1987): 374–86.

Journal of Ethics and Social Philosophy
VOL. 23, NO. 3 · JANUARY 2023
https://doi.org/10.26556/jesp.v23i3.2707

AGENCY, STABILITY, AND PERMEABILITY IN "GAMES"

Elisabeth Camp

Chi Nguyen's "Games and the Art of Agency" is a landmark article, backed by an important and engaging book.[1] If they do not exactly inaugurate the philosophical study of games, they most certainly level it up considerably. While there is much to explore here about what counts as a game, when games constitute art, and why they are aesthetically valuable, I want to focus on what Nguyen's discussion reveals about agency. One significant contribution of his analysis is that it highlights a profound complexity in human motivation. I think it also thereby calls into question a traditional notion of selfhood—one that plays a crucial role in Nguyen's own analysis. Without this traditional conception, games look more like life, and both look riskier, than we might otherwise hope.

1. STRIVING PLAY AND NESTED AGENTS

Nguyen proposes that a game is a complex structure consisting of a goal, a profile of deployable abilities, and an environment (partially abstract, often also concrete) that presents obstacles to and opportunities for achieving that goal using those abilities. So, for instance, basketball is a physical game with the goal of scoring points by passing a ball through a small elevated hoop while dribbling and passing to teammates, on a court with marked zones, while avoiding noncontact obstacles constituted by the opposing team's bodies. By creating such a structure, a game designer invites players to exercise an *agential mode*: a pairing of a type of goal (here, scoring points) with a set of skills (dribbling, shooting, blocking) and patterns of attention for fulfilling it.

A game is designed to elicit a particular mode, which is especially apt for playing it. But it also thereby makes that mode, or an analogue of it, available for real-life action. Human agency in general is characterized by a duality of limitation and flexibility: deploying one agential mode precludes another, but

1 Nguyen, *Games*.

we also expand and refine our repertoire of modes over time, and we can at least sometimes choose which mode to activate at a time. Nguyen argues that games transform this duality of limited flexibility into art, by "sculpting" and "crystallizing" agential modes into stable, tangible forms that focus attention and skills in precise, well-defined ways , which we can then deploy elsewhere, in less scripted contexts.[2] Participating in a game's interlocking structure of goals, abilities, and obstacles also affords access to game-extrinsic, real life goods like exercise and social connection. And it can be aesthetically rewarding in its own right, by offering an experience of harmonious "flow," and the "existential balm" of engaging with a coherent environment in which success is possible but not guaranteed.[3]

However, Nguyen argues that unlocking these various extrinsic and intrinsic rewards requires a "peculiar motivational two-step" of coming to care about something we recognize to be pointless.[4] All game play involves tackling artificial obstacles under arbitrary constraints in pursuit of the artificial goal that constitutes winning—that's what makes it a game. Some players—*achievement players*—really want to win, either in their own right or as a means to fame or fortune; given this, they also really care, albeit only instrumentally, about achieving the game's internal goals. But others—*striving players*—do not care about winning; they just want to engage in the struggle, either for its own intrinsic reward or as a means to an extrinsic end like exercise or social connection. The conundrum is that playing games is defined by trying to win them. Given this, striving players must invert the ordinary structure of means-end motivation: they must (try to) win just in order to play.

How can striving play even be possible? Nguyen argues that it requires *temporarily* adopting winning as a *genuine* goal. Normatively, the striving player's behavior must be guided by trying to win. And given this, in order to play well—or even to "really play" at all—those goals must dominate their functional motivational structure and attention. This much is compatible with winning being a merely instrumental goal, as it is for the achievement player whose ultimate interest lies in fame or fortune. However, many of the ultimate goals that motivate striving players, like aesthetic appreciation or social connection, are "self-effacing": they cannot be pursued directly and "transparently."[5] This means, Nguyen thinks, that winning must become a *disposable end*, which he in turn analyzes as a goal that is genuine and noninstrumental but adopted temporarily and voluntarily, insofar as it is "partially detached from our normal

2 Nguyen, "Games and the Art of Agency," 427, 432.

3 Nguyen, "Games and the Art of Agency," 456.

4 Nguyen, "Games and the Art of Agency," 440.

5 Nguyen, "Games and the Art of Agency," 441.

ends" in such a way that "one can rid oneself of [it] without doing significant damage to one's enduring value system or core practical identity."[6]

Like instrumental ends, disposable ends in the service of self-effacing ulterior goals are not especially unusual: we regularly take up hobbies like knitting, kickboxing, or cooking for the sake of health or social connection. Nguyen argues that striving play's "motivational two-step" is more distinctive, though, because it often requires not just turning one's attention away from the ulterior goal and toward the implementing one, but actually modifying one's motivational structure to include goals that conflict with one's enduring ends. So, for instance, the ultimate goal of social connection may require a local goal of ruthless domination.[7] Likewise, within the game it is at least "odd" and perhaps incoherent to avoid a strategic move on the ground that doing so would prolong the pleasure of striving, whereas outside the game, it is reasonable to avoid acquiring additional game-relevant skills if doing so would make it too easy to win the next time one plays.[8]

Ngyen argues that in order to accommodate this divergence in motivational structures, we need to posit a layered or "nested" agent.[9] On the inside, dominating one's practical rationality and phenomenology, is a game agent wholeheartedly and single-mindedly focused on winning. Meanwhile, lurking in the background is an "enduring agent" who monitors the game agent's performance "in an interestingly distanced way."[10] Ultimately, Nguyen concludes that the existence of such "purposeful and managed agential disunity" reveals human agency to be more "fluid" and "modular" than philosophers have heretofore recognized.[11]

2. PRETENSE, QUARANTINE, AND PERMEABILITY

A natural alternative to Nguyen's analysis treats the striving player as temporarily adopting winning, not as a genuine, noninstrumental, disposable end, but as a merely *pretended* one. Nguyen argues against this alternative by pointing out, first, that the motivational structure of someone who is merely "acting as if" they care about the goal of winning will focus on producing observable behaviors that mimic caring, where this may come apart from or even conflict

6 Nguyen, "Games and the Art of Agency," 435, and *Games*, 34.
7 Nguyen, "Games and the Art of Agency," 445.
8 Nguyen, "Games and the Art of Agency," 437.
9 Nguyen, "Games and the Art of Agency," 443.
10 Nguyen, "Games and the Art of Agency," 447, 443.
11 Nguyen, "Games and the Art of Agency," 445.

with actual caring-type thoughts.[12] Second, he points out that the motivational structure of a striving player need not revolve around or even be couched in terms of a game's fictional goals, such as rescuing the princess.[13] Thus, both "acting as if" one wants to win and caring about fictional goals are at most optional for striving play. By contrast, he says, the goal of winning must occupy a "central and immediate role" within the striving player's psychology for the duration of game play in order for them to really play at all.[14]

The problem is that even if we grant that these psychological contrasts are apt, and that winning plays a dominant normative, functional, and phenomenological role in striving play, this does not suffice to establish winning as a genuine goal for the striving player; after all, this is just what a pretense theorist denies. On a pretense view, game play involves a complex interplay of real-world actions and mental states and corresponding fictional actions and mental states, linked by pretense. More specifically, the pretense theorist holds that I genuinely perform certain real-world actions that make it fictional that I accomplish (or fail to accomplish) certain game goals, and I pretend *of* those real-world actions and their effects that they have their prescribed in-game significance. Likewise, I pretend *of* my actual real-world psychological states that they are instantiations of psychological states that I really would have if the fiction were real.

Given this, the pretense theorist holds that we cannot read off the attitude and content of any individual psychological state, or cluster of states, in isolation. Rather, whether those actual states genuinely have a certain content depends on how they interact with the rest of the agent's psychology. Make-believe or simulated states are, by definition, "off-line," in the sense that they are *quarantined* from the rest of an agent's beliefs and actions.[15] Thus, the theatergoer's racing heart does not constitute real fear, but only quasi-fear, because they do not believe they are in danger or flee the theater.[16] Likewise, the striving player's very real "armpit sweats, jitters, and surge of adrenaline" do not constitute genuinely wanting to win, because the player does not undertake the full range of extra-game actions that would rationally support this goal.[17]

According to the pretense theorist, then, the striving player is just like the achievement player insofar as they both engage in the game's prescribed

12 Nguyen, "Games and the Art of Agency," 447.

13 Nguyen, "Games and the Art of Agency," 449.

14 Nguyen, "Games and the Art of Agency," 448.

15 Goldman, "Empathy, Mind, and Morals"; Currie, "The Moral Psychology of Fiction"; Walton, "Spelunking, Simulation, and Slime"; and Nichols and Stich, "A Cognitive Theory of Pretense."

16 Walton, "Fearing Fictions."

17 Nguyen, "Games and the Art of Agency," 436.

pretense in order to make it fictional that they have achieved the game goals, because doing so will make it actually true that they have won. The only difference is that for the striving player, the scope of their pretense also extends to include their caring about winning.

If this analysis of striving play is coherent, Nguyen and the pretense theorist appear to be locked in a dialectical impasse. They agree that the real, enduring agent does not really care about winning. They agree in their descriptions of the player's psychological states construed narrowly, in terms of physiology, phenomenology, and local functionality. And they agree that the player's actions are locally coherent but appear to conflict with their enduring goals. They differ only in their descriptions of these states and actions and their explanation of the putative conflict. Nguyen explains it by positing a nested agent who genuinely wants to win and who pursues that goal by undertaking actual actions (e.g., capturing a knight) whose reality is constituted by the game's rules plus more basic actions (e.g., moving a plastic piece three squares), because performing those actions in that context helps fulfill actual but nested winning-conducive goals (e.g., launching a debilitating assault by surprising their opponent), where pursuing these goals in this context in turn facilitates a genuine long-term goal (e.g., social connection). By contrast, the pretense theorist posits a single agent who merely pretends to want to win, and who implements that pretense by undertaking real-world actions (e.g., moving a plastic piece three squares) that implement merely fictional actions (e.g., capturing a knight) in the service of merely fictional winning-conducive goals (e.g., launching a debilitating assault by surprising their opponent), because the immersive pretense of pursuing those goals facilitates a genuine long-term goal (e.g., social connection).

Given all that Nguyen and the pretense theorist agree on, it is unclear who has the burden of proof, or what proof they could provide. Moreover, it would seem that the pretense theorist has the advantage of parsimony, and that Nguyen could capture all the data he adduces while avoiding the Meinongian profligacy of positing multiple agents by recasting the "motivational two-step" of striving play in terms of functional and phenomenological immersion in a merely pretended goal of winning.

I suspect that many will be tempted by this route. However, I would urge Nguyen to hold fast to the idea that winning is a temporary but genuine, non-instrumental goal for the striving player. But I advocate this option because I reject an assumption that both Nguyen and the pretense theorist endorse: that the local motivational structure of striving play is robustly *quarantined* from the enduring motivations of real life.

The pretense theorist holds that a mental state like quasi-fear constitutes a mere simulation because it is quarantined from the enduring agent's broader network of beliefs. Similarly, Nguyen holds that the goal of winning belongs only to the nested game agent because it is quarantined from the enduring agent's broader network of goals. According to him, striving play involves a "single-minded absorption" in which we "aggressively seal ourselves off from the vast majority of our usual ends and considerations."[18] While playing, the temporary game agent is in total control; the enduring agent is only engaged via "background monitoring processes," lounging in the wings to intervene if things go too far awry.[19] This is how games can be "morally transformative technologies" that "turn competition into cooperation" in shared pursuit of the experience of striving.[20]

I agree that robust quarantine happens and that it is theoretically revealing. But I also think such "aggressive sealing off" is relatively rare. In my experience, even highly competent and engaged players are often attentive to external social relations throughout the course of play. Their real-life expectations, hopes, and worries about their own and other players' game-extrinsic psychologies affect the intuitive salience and attractiveness of in-game moves, strategic choices, and emotional responses in pervasive and nuanced ways. Likewise, their own in-game and extra-game goals operate in more direct competition and interaction than Nguyen's overseer model predicts. And in those cases where players do achieve single-minded, wholehearted immersion, it is not obvious that they have not temporarily slipped into achievement play.

These intimate interactions between internal and external motivational structures arise partly because our knowledge of other players' game-extrinsic psychologies helps us predict their in-game actions, and because we care about how game play affects their post-game attitudes. This much is arguably compatible with the nested model. But we also take our enduring selves to bear at least some *responsibility* for our game actions inherently, apart from their in- and post-game effects on other players. Thus, Brenda Romero's installation-art board game, *Train*, is designed to induce an experience of moral complicity as players realize that in efficiently moving yellow pieces across the board they are fictionally shipping prisoners to Holocaust concentration camps.[21] At a smaller scale, one of my many reasons for hating Monopoly is that I do not like the agential mode of being "narcissistically bent toward the destruction of

18 Nguyen, "Games and the Art of Agency," 440, 441.

19 Nguyen, "Games and the Art of Agency," 443.

20 Nguyen, *Games*, 174.

21 Nguyen, *Games*, 103.

others for my own good," even if I am confident that I can put that mode aside after playing.[22] More specifically, the reason I do not like it is that my in-game behavior reveals something about my real character: that I am competent in, and able to deploy and even revel in, this agential mode. (And for that same reason, I do not like it when my kids enact it either.)

Nguyen focuses his analysis on highly formalized games with fixed, explicit rules and arbitrary goals. The permeability of the game-life boundary is underscored if we expand our purview to include more fluid games. Fluidity and permeability are especially palpable with children's games, which often begin as spontaneous sandbox play and evolve into something more constrained and articulated, with as much energy invested in haggling over rules as in actual play. Adult players are especially likely to experience permeability and to feel and impute in-game responsibility while playing open-ended, interactive, role-playing games like *World of Warcraft*—with more pro-social players feeling more in-game control and responsibility, and with skilled, young, male gamers apparently being more likely to engage in anti-social game play.[23]

I think the profile of quarantine and permeability with games closely parallels our engagement with fiction. Many readers of fiction regularly cultivate interpretive perspectives and attendant emotional and moral responses that differ markedly from those they would have if they encountered the same situations in real life; but at the same time, that interpretive flexibility also displays significant causal and normative limits, with different readers being more or less willing or able to bracket their real-world perspectives.[24] In both cases, I take the lack of robust quarantine plus the presence of constrained flexibility to suggest that our engagement with art often involves actually but temporarily trying on alternative modes, rather than merely pretending to do so.

However, acknowledging the permeability of the game-life boundary undermines quarantine as a criterion for demarcating a genuine interest in winning from a merely pretended or nested one. As Walton himself says:

> It will not always be obvious whether and to what extent a competitor or spectator engages in make-believe. . . . [It] may not be evident even to the pretender herself. Perhaps in some instances there is no fact of the matter about whether a person is engaging in pretense.[25]

22 Nguyen, *Games*, 90. Indeed, Monopoly originated as a game intended to drive home the moral and economic perils of landlording.

23 Banks and Bowman, "Avatars Are (Sometimes) People Too"; and Bowman, Schultheiss, and Schumann, "'I'm Attached, and I'm a Good Guy/Gal!'"

24 Camp, "Perspectives in Imaginative Engagement with Fiction."

25 Walton, "'It's Only a Game!'" 82–83.

Indeed, if permeability is as pervasive as I take it to be, this cuts against any clear segregation of motivational structures as genuine or either pretended or nested. Some players clearly do sometimes achieve the sweet spot of "absorbed, thrilling play" just for the experience of struggle. But for many more of us, our motivational structure is considerably more unstable: sometimes we engage in striving play, sometimes we fall into achievement play despite ourselves, and often we experience that "peculiar double-consciousness" of motivations, which may be more or less "anxious" depending on our personalities and circumstances.[26]

3. Stability and Selfhood

Stepping back from the debate between nesting and pretense analyses, these observations about fluidity and permeability largely support Nguyen's core conclusion that a kind of "purposeful and managed agential disunity" is not merely common but advantageous in human agency.[27] Indeed, I think they press us to push that conclusion further.

Construing agency primarily in terms of enduring beliefs and goals motivates an analysis in which game players and fiction readers do not *really* change their minds, insofar as their temporarily dominant phenomenology and functionality are not properly integrated with their long-term, reflective attitudes. This gets something right: we do have cross-contextually stable concepts, beliefs, and goals, which we deploy in the course of planning and executing such myriad activities as making meals, buying houses, and building bridges and constitutions.[28]

However, those stable attitudes do not exhaust who we are. More importantly, those long-term attitudes are formed, accessed, and revised in concert with intuitive dispositions to parse, prioritize, and respond to particular properties and possibilities as we encounter them within particular contexts. Where Nguyen emphasizes the role of intuitive *agential modes* in practical action, I have emphasized the role of intuitive cognitive *perspectives* in interpretation.[29] Both perspectives and agential modes are significantly more malleable than beliefs and goals as traditionally conceived. Moreover, both are partly, but only partly, under voluntary control, in a way that motivates an analogy with

26 Nguyen, "Games and the Art of Agency," 445.

27 Nguyen, "Games and the Art of Agency," 445.

28 Camp, "Logical Concepts and Associative Characterizations."

29 See eg., Camp, "Metaphor and That Certain 'Je Ne Sais Quoi,'" "Logical Concepts and Associative Characterizations," and "Perspectives and Frames in Pursuit of Ultimate Understanding."

gestalt perception: we can try to adopt or cast them off, but "getting" them is something that ultimately just happens. When it does, this makes a substantive phenomenological and functional difference, by activating an open-ended ability to "go on" in interpreting and responding to an indefinite range of further situations. By highlighting and fostering the flexibility of these intuitive, phenomenologically and functionally dominant aspects of our psychology, both games and fiction reveal human agency to be more "fluid and fleeting" than the traditional view maintains.[30]

Nguyen treats agents as stable, robust selves armed with "libraries" or "Swiss Army knives" of "modular" agential modes.[31] But given the permeability of agential fluidity, it might be more appropriate to think of persons as chameleons, morphing among modes of interpretation and action as they traverse disparate contexts. On this model, we develop selves by building repertoires of interpretation and action, within which beliefs and goals function as especially stable nodes. The locus of agency would then reside as much in one's choices about which contexts to enter and which modes to cultivate as in one's enduring, reflectively endorsed commitments or one's moment-to-moment choices. And we would achieve selfhood not necessarily by subsuming our lives under extended teleological structures, but rather by integrating our repertoires for engagement into coherent characters whose contextual variations hang together in complex higher-order wholes.[32]

I take it that this model is very much in the spirit of Nguyen's overall view, but that it moves at least one step further away from the traditional picture of autonomous rational liberalism. Applied to game play, it may even point in the opposite direction, by suggesting that the primary locus of agential stability resides not in an enduring agent who constructs a nested, winning-obsessed game agent as a means to fulfilling a long-term goal like social connection. Rather, agential stability resides *in the game itself*, precisely because and to the extent that the game constitutes a crystallized frame for "inscribing" and "storing" a well-defined agential mode.[33]

Here again, I take games to exhibit a close analogy with fictions, along with other species of interpretive frame, like metaphors and slurs, which crystallize perspectives.[34] Like interpretive frames in general, games schematize—or

30 Nguyen, *Games*, 79.

31 Nguyen, "Games and the Art of Agency," 426, 457, and *Games*, 86, 89.

32 Camp, "Wordsworth's Prelude, Poetic Autobiography, and Narrative Constructions of the Self."

33 Nguyen, "Games and the Art of Agency," 427.

34 Camp, "Metaphor and That Certain 'Je Ne Sais Quoi,'" "Showing, Telling, and Seeing," "Slurring Perspectives," and "Imaginative Frames for Scientific Inquiry."

"sculpt"—an otherwise amorphous mode of engagement in simpler, more discrete terms. More specifically, like mantras—such as "He's just not that into you," "What would Jesus do?" or "It's the economy, stupid"—games offer concrete, tangible touchstones for action that can be accessed by multiple agents across multiple contexts.[35] By functioning to coordinate intuitive engagement in ways that we can try to deploy but that ultimately function intuitively and beneath the level of voluntary control, both games and interpretive frames constitute powerful "social technologies," which can be used for good and for ill.[36]

4. Learning and Life

These observations—about the "flexible and fleeting" quality of agency in general, about the permeable boundary between games and life, and about frames as interpretive stabilizers—also push us to adopt more cautionary versions of Nguyen's lessons about how playing games helps us learn about life.

Nguyen argues that games are "yoga for your agency" in at least three ways.[37] First, playing a variety of games can enrich our practical resources by augmenting our *repertoire* of agential modes. Second, it can train us to be *flexible* in choosing goals and agential modes. And third, engaging specifically in *aesthetic striving play* "fosters a special form of agential fluidity, where we enter into, and then step back from, the narrowly practical state" of game play.[38] Here, once again, I find Nguyen's case for games' agency-building potential to be generally persuasive but overly tidy. Nguyen cautions that the lessons offered by games are contingent: games are "a resource for autonomy development, not a guarantee. . . . You can misuse games, just as you can misuse Jane Austen."[39] However, I think that acknowledging the permeability between games and life, and the variety in formalization among games, reveals the hazards of misuse to be considerably more subtle and pervasive than he acknowledges.

At the first order, there is the risk of habituation. Like fictions, games inculcate open-ended patterns of attention and response that can linger even if we intend to indulge them only temporarily and instrumentally, and even if we abstract away from their particular contents.[40] Thus, just as a researcher might intend to read *Lolita* merely in order to gain a better understanding of

35 Nguyen, "Games and the Art of Agency," 438.

36 Nguyen, *Games*, 1.

37 Nguyen, "Games and the Art of Agency," 458.

38 Nguyen, *Games* 216.

39 Nguyen, *Games*, 92.

40 Camp, "Perspectives in Imaginative Engagement with Fiction."

pedophilia but inadvertently end up more disposed to notice and interpret tween girls' pubescent features in sexual terms, so might a "good-hearted agent" intend to play Monopoly simply to placate their whining child or to predict the scheming of real estate moguls but end up genuinely more disposed to notice opportunities for exploiting other people's financial vulnerabilities.[41]

To combat habituation, we need a form of agency that is not just fluid, but actively flexible: one that enables us to "apply [our agential] inventory in the right circumstances."[42] The problem is that deploying an active, flexible agency requires selecting an agential mode that appropriately matches our goals and circumstances. But lurking beneath the risk of habituation into agential modes we reflectively reject lurks the deeper problem that we are often unclear or confused about which agential mode really is appropriate, given our goals and circumstances. Worse, it may be indeterminate what our goals and circumstances themselves really are. Games are satisfying because they set us right-sized goals in preestablished harmony with their environments. Insofar as they are explicit and formalized, with fixed goals and tightly sculpted agential modes, they obviate the need to form those goals or develop those modes for ourselves. Abstract, highly restricted games like chess define precise grids of interlocking choice points, with little room for rational deviation. At the limit, games like War or Chutes and Ladders offer no agential choice, but merely a narrative and phenomenology of striving. But this means that the sort of flexibility we gain by playing even a wide variety of games may not just fail to foster but actively hinder the development of an accurately perceptive and appropriately responsive species of agency.

One tempting way to manage the mess of life is to stick to our default modes of interpretation and action; after all, their success in getting us this far constitutes some evidence that we have accurately assessed our circumstances and selected commensurately effective perspectives and modes for handling them. However, this comforting complacency may itself be borne of myopia: we may be ignoring complexities we *should* notice, or failing to appreciate alternative values and strategies we could embrace. Open-minded exploration of the sort fostered by games and fictions is indeed the best antidote to such complacency. But it carries its own risk: of being seduced into modes that appear satisfying precisely because they are stable and schematic.[43]

Nguyen is deeply insightful about the risks of such "gamification." Much as we can fall into exporting particular open-ended perspectival patterns of

41 The term "good-hearted agent" is Nguyen's (*Games*, 91).

42 Nguyen, "Games and the Art of Agency," 458.

43 Camp, "Perspectival Complacency, Perversion, and Amelioration."

attention and response even while carefully bracketing a game or fiction's specific contents, so can we fall into exporting a more generalized assumption of "value clarity" even while bracketing the particular modes of the games we play.[44] Here again, highly formalized, "teleologically crisp" games like chess are especially seductive.[45] But even more amorphous games like *World of Warcraft* foster the primordial fantasy that one's environment contains a hidden meaning that, once unlocked, determines a right action.

In this vein, game designer Reed Berkowitz argues that the political conspiracy theory QAnon is so pernicious because it exploits three sources of cognitive reward that game designers also tap into: apophenia, or promiscuous pattern recognition; the phenomenology of "eureka!" insight; and social competition and validation.[46] But where actual game designers carefully channel these factors to keep players moving toward an ultimate goal that coherently integrates the game's environment, obstacles, and abilities, QAnon is "AI with a group-think engine," inciting unfettered apophenia in service of an alternate-reality-creating pyramid scheme. In this case, it is precisely the fluid, evolving nature of gamification that makes it so seductive and self-perpetuating, and hence so destructive when unleashed on the real world.

Ultimately, Nguyen's true hero for agential calisthenics is not games, but striving play. And indeed, striving play promises to provide a distinctively powerful tool for autonomy development, because it trains us to treat not just the various goals of the games we play, but also winning itself as a disposable end. However, precisely because winning is so cognitively and socially alluring, and because striving play requires a locally dominant focus on winning, striving play is also quite precarious. The risk of falling into achievement play always looms, and with it the risk of actively hindering our autonomy by blinding us to other, more profound but messier and more organic values.

Our last, best hope for building autonomy through games is *aesthetic* striving play: cultivating a form of "impractical and unfiltered attention" that staves off achievement play while nurturing deep open mindedness, in a way that can then equip us to notice subtle, neglected properties and values as we stumble across them in life.[47] Even here, though, it is not obvious that the type of disengaged self-reflection that characterizes the aesthetic attitude readily transfers to the type that is relevant for autonomous, critical self construction in life. As Richard Posner notes in his critique of Nussbaum's "moral imagination,"

44 Nguyen, *Games*, 199; see also "The Seductions of Clarity."

45 Nguyen, "Games and the Art of Agency," 457.

46 Berkowitz, "A Game Designer's Analysis of QAnon."

47 Nguyen, *Games*, 118.

aesthetic sophistication and wholehearted empathetic fictional engagement can serve as welcome escapes from an unpleasant reality and are all too compatible with real-life moral myopia and perversion.[48] So too, cultivating an aesthetic appreciation of the harmony between one's experience and environment within a game is not just compatible with but can actively hamper investment in more ethically pressing dimensions of assessment. Moreover, aesthetic reflection is arguably easiest and most rewarding to achieve with highly formalized, tightly sculpted games; but, if so, this very formalization makes transferring the aesthetic attitude from the game to practical engagement with messy reality that much more challenging. Thus, I take it that the risks of disuse and misuse from games for autonomy development are not just possible in principle, but pressing in practice.

As human agents, we need to be both fluid and persistent in our modes of engagement. As Nguyen demonstrates, games exploit and foster both of these capacities. Playing a rich variety of well-designed games, with the right attitude under the right circumstances, can expand and strengthen our agency in ways that other art forms and activities do not. But playing games offers no reliable recipe for crafting rich, sensitive, reflective persons. This should not surprise us: in real life—unlike games—there are no sure-fire recipes.

Rutgers University
elisabeth.camp@rutgers.edu

REFERENCES

Banks, Jaime, and Nicholas David Bowman. "Avatars Are (Sometimes) People Too: Linguistic Indicators of Parasocial and Social Ties in Player–Avatar Relationships." *New Media and Society* 18, no. 7 (August 2016): 1257–76.

Berkowitz, Reed. "A Game Designer's Analysis of QAnon: Playing with Reality." Curiouserinstitute, 2020. https://medium.com/curiouserinstitute/a-game-designers-analysis-of-qanon-580972548be5.

Bowman, Nicholas David, Daniel Schultheiss, and Christina Schumann. "'I'm Attached, and I'm a Good Guy/Gal!': How Character Attachment Influences Pro- and Anti-Social Motivations to Play Massively Multiplayer Online Role-Playing Games." *Cyberpsychology, Behavior, and Social Networking* 15, no. 3 (March 2012): 1–6.

Camp, Elisabeth. "Imaginative Frames for Scientific Inquiry: Metaphors,

48 Posner, "Against Ethical Criticism."

Telling Facts, and Just-So Stories." In *The Scientific Imagination*, edited by Arnon Levy and Peter Godfrey-Smith, 304–36. New York: Oxford University Press, 2019.

———. "Logical Concepts and Associative Characterizations." In *The Conceptual Mind: New Directions in the Study of Concepts*, edited by Eric Margolis and Stephen Laurence, 591–621. Cambridge, MA: MIT Press, 2015.

———. "Metaphor and That Certain 'Je Ne Sais Quoi.'" *Philosophical Studies* 129, no. 1 (May 2006): 1–25.

———. "Perspectival Complacency, Perversion, and Amelioration." In *Open-Mindedness and Perspective*, edited by Wayne Riggs and Nancy Snow. New York: Oxford University Press, forthcoming.

——— "Perspectives and Frames in Pursuit of Ultimate Understanding." In *Varieties of Understanding: New Perspectives from Philosophy, Psychology, and Theology*, edited by Stephen Grimm, 17–45. New York: Oxford University Press, 2018.

———. "Perspectives in Imaginative Engagement with Fiction." *Philosophical Perspectives* 31, no. 1 (December 2017): 73–102.

———. "Showing, Telling, and Seeing: Metaphor and 'Poetic' Language." *Baltic International Yearbook of Cognition, Logic, and Communication* 3 (2007): 1–24.

———. "Slurring Perspectives." *Analytic Philosophy* 54, no. 3 (September 2013): 330-349.

———. "Wordsworth's Prelude, Poetic Autobiography, and Narrative Constructions of the Self." *nonsite.org*, October 14, 2011. https://nonsite.org/wordsworths-prelude-poetic-autobiography-and-narrative-constructions-of-the-self/.

Currie, Gregory. "The Moral Psychology of Fiction." *Australasian Journal of Philosophy* 73, no. 2 (1995): 250–59.

Goldman, Alvin. "Empathy, Mind, and Morals." *Proceedings and Addresses of the American Philosophical Association* 66, no. 3 (November 1992): 17–41.

Nguyen, C. Thi. "The Seductions of Clarity." *Royal Institute of Philosophy Supplement* 89 (2021), 227-255.

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

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

Nichols, Shaun, and Stephen Stich. "A Cognitive Theory of Pretense." *Cognition* 74, no. 2 (February 2000): 115–47.

Posner, Richard. "Against Ethical Criticism." *Philosophy and Literature* 21, no. 1 (April 1997): 1–27.

Walton, Kendall. "Fearing Fictions." *Journal of Philosophy* 75, no. 1 (January

1978): 5–27.
———. "'It's Only a Game!': Sports as Fiction." In *In Other Shoes: Music, Metaphor, Empathy, Existence*, 75–83. Oxford: Oxford University Press, 2015.
———. "Spelunking, Simulation, and Slime." In *Emotion and the Arts*, edited by Mette Hjort and Sue Laver, 37–49. Oxford: Oxford University Press, 1997.

Journal of Ethics and Social Philosophy
VOL. 23, NO. 3 · JANUARY 2023
https://doi.org/10.26556/jesp.v23i3.2708

COVERAGE SHORTFALLS AT THE LIBRARY OF AGENCY

Elijah Millgram

In "Games and the Art of Agency," C. Thi Nguyen makes an intriguing and very plausible suggestion: games, or at any rate a great many of them, are artworks whose medium is, roughly, how one goes about doing what one does.[1] In assigning an objective, laying down the constraints under which it has to be achieved, and specifying the terrain on which it will be played out, a game sculpts the decision-making processes of its players, the ways they see their environment and option space, their motivations, and much else. Thus our by now quite extensive repertoire of games constitutes a *library of agency*. This library allows us to try on different modes of agency before deciding which is best for us—for a given type of occasion, or generally. It can help educate us into unfamiliar forms of agency by providing the sort of controlled exercises that allow beginners the practice they need, which is to say that games are exercise and preparation for autonomous agency. And it promises to broaden and enrich our philosophical treatments of the topic, in part by serving as a testbed for competing theories of practical rationality; if we want to get a realistic sense of what it would be like to decide what to do, in the way that one or another theory of practical deliberation says, we can experiment with it in an appropriately designed game.[2]

All of this seems on target to me, and an important step forward for, especially, the ongoing discussion of practical reasoning. However, in availing ourselves of this very valuable resource, it is important to remain aware of

1 Nguyen, "Games and the Art of Agency." The ideas are further developed in Nguyen, *Games*.

2 The recent back and forth about constitutivism is focused on the question of what agency essentially is, the presumption being that agency is *one* thing. See, for instance, Ferrero, "Constitutivism and the Inescapability of Agency"; for an overview of the action-theoretic variant of that debate, see Millgram, "Practical Reason and the Structure of Actions." Nguyen's treatment obviously pulls in a very different direction, but here we will not need to take up the question of whether there are aspects of the way one goes about doing things that are simply nonoptional.

its limitations, and so those are what I will be highlighting here. The point is neither to object to Nguyen's view, nor even to suggest that he has overlooked the issues I will be raising; on the contrary, some of them are central to his own discussion, and others he acknowledges in passing. Rather, I mean to contribute a helpful reminder to what might one day evolve into a user guide for the library of agency.

I

Although my very terse summary confined itself to the aspects of Nguyen's argument that bear most directly on work on agency itself, his essay is first and foremost a contribution to the philosophy of art, identifying a largely overlooked and underappreciated class of artworks. Works of art are produced and consumed for their aesthetic properties, and perhaps the entry ticket for the class—the aesthetic property that keeps a game in the library of agency—is *playability*. Pausing for a moment on that concept, by introducing it as an aesthetic property, that is, in the same logical family as, say, beauty or uncanniness, I mean to distinguish it both from what it takes, formally, for an activity to count as gameplay at all, and also from being simply enjoyable to play (although I do not mean to discount the presumptive links between the three conditions). If the analogy helps, a film may be unwatchable even though it is possible to watch it, and it may be compellingly watchable despite being morbidly unpleasant; important documentaries on difficult topics tend to fall into this latter category, and conversely, to foreshadow our next step, unwatchable documentaries all too easily end up being unimportant.

If games are works of art, then to the extent that there are forms which agency can take in the wild that make for unappealing play, the modes of agency induced by games will be unrepresentative of agency across the board. That is, the library of agency should be expected to exhibit *playability bias*.[3] Consider some of the ways in which the mix of agencies invoked by games will diverge from what we ought to find in the wild.

First and foremost, when someone sits down with you to introduce you to a new game, they will almost always start out by telling you what the objective

3 Since not all artworks are games, this is going to be a special case of a presumably more general phenomenon: there will be aesthetics-driven selection effects across the arts, with upshots for the uses that get made of artworks. For instance, we should be suspicious when moral philosophers appeal to snippets from famous works of fiction. What made the work famous? No doubt (although *inter alia*) its aesthetics, and we should be asking: What are we *not* going find in novels, because people would be very unlikely to want to read a novel like *that*?

of the game is—say, to checkmate your opponent's king. There are perhaps exceptions (think of *Minecraft* or *The Sims*), but games for the most part come with goals.[4] And goals are a distinguishing feature of the class of games that take center stage in Nguyen's discussion, those that are occasions for "striving play": the attempt to achieve a designated objective in the face of specified constraints and impediments, for the sake of the experience of doing so.

Two features of the way the objective of a game figures into it matter for our purposes. One, *all* of the in-game activity is to be directed toward achieving the objective of the game. For instance, if one of the players positions their pawns and rooks in an elegant pattern, not for the sake of the win but because that strikes them as a pretty way to arrange the board, they are no longer really playing chess. And two, the objective of the game is not negotiable; you do not, in the course of a game of chess, propose that perhaps instead of checkmating the king, it would suffice to weaken his armies and render them nonthreatening—or that it should be enough if bad publicity makes the king into a lame duck.

Because this will be a controversial claim, right now I neither want to insist on it, nor be detained by it. Nonetheless, it does seem to me that an important aspect of agency in the wild is figuring out what matters and what is important, and thus what one's goals or objectives are to be.[5] Deliberation of ends, as the old-school way of speaking designates it, is often a frustrating endeavor; there is no cut-and-dried procedure that gets you the right answer, and it is typically hard to tell that you *have* gotten the right answer. Consequently, people often will not agree on whether that sort of question has been successfully resolved. If

4 "Perhaps": this is a tricky question, and taking *Minecraft* as our illustration, first distinguish its "creative" and "survival" modes, the latter being an overlay of much more traditional game structure, goals and all, on the former. If we confine ourselves to that creative mode, which was what made *Minecraft* so popular in the first place, in its pure form it is something on the order of virtual Lego.

Now, and here is a suggestive distinction drawn from ordinary language, when we say that a child is playing with Lego, we do not say that they are playing a *game*—rather, they are playing with a *toy*. (As a matter of "grammar," as an old-school, ordinary-language philosopher might say it, what you do with a game is play it, but not all play is taking part in a game.) We do call *Minecraft* a "game," but apparently that is mostly a matter of commercial near-convention: recreational software is categorized this way even when the recreational activity it enables would not, if off-device, be considered playing a game. (For very helpful guidance from a native informant, I am grateful to Abie Millgram.)

5 I argue that we have to learn what matters from experience in *Practical Induction*, and survey the state of play in the instrumentalism debate as of about the turn of the millennium in *Varieties of Practical Reasoning*. Vogler makes what is still the best case in the literature for (a nuanced version of) the opposing view: that actions have to be directed toward objectives, and that practical reasons that are not generated by objectives are entirely optional (*Reasonably Vicious*).

only because that makes it hard to score, it is quite understandable that deliberation of ends does not generally figure into the demands that a game—anyway, a game that most people could enjoy—makes on its players.

In addition, people generally seem to have an appetite for vicarious activity that is solely end driven, and where the ends themselves are not up for reconsideration; witness not just games, but the many genres of popular fiction in which readers identify with a protagonist who strenuously overcomes obstacles in order to attain some antecedently given objective. (There are many variations on the structural theme: he must defuse the bomb, or win the affections of a romantic interest …) The appetite for single-minded, goal-driven activity in real life is much more muted; when it is not a game, we are much more liable to take a relaxed approach to our goals, procrastinate, and generally let other issues influence our choices and the way we execute them. But playability is enhanced when a game caters to a deeply rooted appetite, and we should anticipate that our repertoire of games will induce and exercise by and large only modes of agency from which—again, if I am right about what is a controversial topic—two significant aspects of agency in the wild have been excised.[6]

II

In Bill Watterson's deservedly famous comic strip, the child plays "Calvinball," a game where you make up the rules as you go. But while Calvin is playing, he

6 The instrumentalism debate is focused on whether you can reason about what your final ends are to be, rather than the possibility of activity that is not structured around ends or goals at all. It does seem to me that this latter is a possibility we should be taking very seriously, and the alternative control structure that we perhaps understand the best is the feedback loop (see, e.g., Millgram, *Ethics Done Right*, ch. 1). It is quite plausible that games built around feedback loops rather than goals can be playable, gripping, and even addictive; so while they are not a focus of Nguyen's discussion, we can bank on finding this sort of agency well represented within the agential library.

Bowman introduces an agential posture that is oriented toward "aspirations" rather than goals; although superficially similar, we expect, if we are at all self-aware, to abandon our aspirations long before they are achieved, and when we do we will not count that as failure (*Are Our Goals Really What We're After?*). When we abandon our goals, that *is* failure; our aspirations, Bowman argues, have a very different cognitive function, and the ways we pick them up and drop them make Bowman's aspirations resemble Nguyen's disposable ends in important respects.

Notice, however, that Bowmanian aspirations are unlikely to lend themselves to rewarding game play. Imagine a would-be game that, instead of objectives, had aspirations: rather than saying to the novice player, "The objective of the game is to checkmate the king" (something you could actually do), they say, "Your aspiration is to checkmate the king, and as the game goes on, you can anticipate that you will just give up on that, and keep playing, but with a new aspiration, which you will also give up…." Who would want to play *that*?

is not playing a *game*; Wittgenstein's observations about family-resemblance concepts notwithstanding, we expect games to come with rules. Be the formal point as it may, for the Suitsian gameplay that is the focus of Nguyen's discussion to be possible, a game must come with something that anyway serves the purpose of rules here, that of complicating and impeding what would otherwise be the too-straightforward achievement of the objective of the game.[7] To serve that function, the rules (or whatever does the job, but I will continue to refer to whatever it turns out to be as rules) must also be nonnegotiable, in much the way that the objective of the game is.[8] If someone asks his opponent whether he can move his pawn like a knight *just this once*, not only has he has given up on playing chess, he is eliciting something on the order of a disappointed sigh.

However, one of the more fraught but also unavoidable activities in life as we have to live it is renegotiating the rules.[9] If you are reading this essay, you are probably an analytic philosopher; in that case, you are working in a tradition that was produced when its founders did a drastic reset of the rules for philosophizing, and since that time, within that tradition, the rules of the game have been renegotiated on a fairly regular basis. For instance, part of that initial reset was the flat-out rejection of the coherentist arguments that had been the stock in trade of Russell's and Moore's British Idealist predecessors. That mode of argumentation has been reclaimed throughout analytic philosophy, sometimes under the label "reflective equilibrium," sometimes in a Davidsonian, and sometimes in a Lewisian accent.[10] If you *are* an analytic philosopher, you are a participant in a practice an essential part of which is renegotiating the rules of that very practice, and while the illustration is in some respects exotic, the phenomenon can be found throughout our social life.

Moreover, there are a good many occasions on which we are no longer in the business of adjusting the rules, or even substituting new rules for old, but rather of ignoring or systematically violating them. Revolution and civil disobedience are dramatic and large-scale examples that come in for attention on the part of political philosophers, and there are also unfortunately too many people who

7 Watterson, *It's a Magical World*, 101; Wittgenstein, *Philosophical Investigations*, secs. 66–71; Suits, *The Grasshopper*.

8 An observation that also requires qualification: the rules of a game can be changed, and recently we witness games being modified regularly, e.g., by the addition of new entities that alter the physics or landscape of a virtual world. But the players themselves do not generally get to adjust the rules in the course of a given game.

9 For a disconcerting illustration at perhaps the largest scale, see Millgram, "The Persistence of Moral Skepticism and the Limits of Moral Education."

10 And perhaps adjustments in this direction were inevitable, for reasons sketched in Millgram, "Relativism, Coherence, and the Problems of Philosophy."

will tell you that all's fair in love and war, but lower-key examples make it clear that this is a pervasive and basic aspect of agency. Just for instance, a great many of the novels or poems covered in a literature class are likely to have run roughshod over the constitutive rules of the genre in which they place themselves.

Briefly, disregarding or altering the metaphorical rules of the game in real time is an indispensable mode of agency. In games, or anyway the games that are the focus of Nguyen's discussion, the literal rules of the game cannot be disregarded, and they cannot be altered within the course of the game itself. Consequently, this mode of agency will also not be properly represented in a library of agency whose card catalog is confined to the Dewey Decimal 793–796 range.

We can now introduce an important complication. It is very plausible that the relevant form of connoisseurship, developed as one deliberates with one's gaming companions about what game to play next, exercises the aspects of agency we have been worried were missing from the library. Thinking about why we were not happy with last night's game, and what we should try instead, is likely to involve deliberation of ends. While you do take the rules for granted while you play a game—and thus games do train you, in one way, in accepting the rules in force as a given—choosing among games ought to hone your awareness that there are alternative sets of rules, and that you can move between them. Thus a connoisseur of games is training himself not to take the rules of a game as given, in a different way.[11] The use of the library of agency as a whole—one's engagement with it as a *library*—compensates for what is not actually on its shelves. And perhaps this mutes the concerns about playability bias we have been developing.

This seems right to me as far as it goes, and we will return to the point below. But when we are considering *bias*, we need to bear in mind not just how resources *can* be used, but how they are most *predominantly* used. Consider for a moment actual libraries, the ones stocked with books. No doubt engaging a library as a *library*, as a whole—browsing the stacks, exploring the many resources it offers, consulting with the librarians—develops skills and attitudes you do not necessarily come by just reading one or another book. But now, how often do you see this sort of engagement? For the most part, users take the fastest shortcut they can find to the volume they need. To think about the effects of libraries as institutions, and in particular about the way libraries shape the habits and dispositions of readers, will likely turn out to be, by and large, to think about how reading one after another book influences typical patrons.

11 For raising this latter point, I am grateful to C. Thi Nguyen.

III

Reasoning is *defeasible* when you would be correct in drawing a conclusion from the premises you have, but there are further things you might learn, or simply additional considerations that might come to mind, none of which would impugn those premises, but that would require you to retract the conclusion: supplemental information or assessment can *defeat* the inference.[12] As I type this, I am on the road, but in quarantine, imposed as part of the Israeli government's attempt to slow the progress of the coronavirus epidemic. My reasons for taking the trip were perfectly satisfactory support for the decision to embark on it, but they would quite properly have been overridden had I realized that I was going to spend my time in self-isolation. That is, the argument for taking my trip was defeasible, and one of the many potential defeaters for it has turned out, belatedly, actually to defeat it. Deductive inference guarantees the truth of its conclusions, given the truth of the premises; reasoning that is not deductive is defeasible; practical reasoning—to a first approximation, reasoning about what to do—is defeasible through and through, perhaps with negligible exceptions.[13]

Unsurprisingly, a player's deliberations in the course of a game are typically defeasible as well. ("Typically": in some extremely rigidly structured games, the argument for making one or another move can be put into deductive form.) During sheepdog trials, perhaps the border collie can hear her handler's whistle, telling her to bring that tiny flock down and left, taking them through the next obstacle on the course; but close up, and interacting directly with this particularly ornery group of sheep on an especially hot day, she is aware that pressing them in that direction will likely make them break and run. In such circumstances, the collie on a winning team overrides her master's defeasible inference, skips the panel, and brings the sheep directly to the shedding ring.

Despite the appearance of shared structure, however, defeasibility management in the world of games differs deeply from what agency in the wild has to muster up. In the real world, defeating conditions for an inference-in-waiting can come from just about *anywhere*. Who *knew* that epidemiology and public

12 Defeasibility travels under various labels: in philosophy of science, discussion centers on *ceteris paribus*—or "other things equal"—generalizations; in AI, this sort of reasoning is *nonmonotonic*. For overviews, see Reutlinger, Schurz, and Hüttemann, "*Ceteris Paribus* Laws"; Horty, *Reasons as Defaults*; and Hlobil, "Choosing Your Nonmonotonic Logic."

13 For a more leisurely introduction to defeasibility in practical inference, and support for that last claim, see Millgram, *The Great Endarkenment*, sec. 6.2. A delicate point that I will not develop further here: in generalizing the contrast between deductive and defeasible to cover practical reasoning, we will want to broaden the thumbs-up status of a premise or conclusion. There is no agreement on how the relevant statuses of steps of a practical argument are to be construed, but insisting that they are true or false is evidently procrustean.

health policy were going to bear on decisions about plane tickets and speaking engagements? But of course opting to take the trip would properly have been preempted by any number of conditions, had they proved to obtain: a pet emergency, an impossible-to-turn-down collaboration with a very tight deadline, the conference turning out to be academically disreputable, or the sudden discovery that a particular manufacturer's aircraft are prone to falling out of the sky . . . Lists of potential defeaters for a nondeductive argument are not only generally open ended; they pose a distinctive challenge, that of noticing the surprising ways that entirely unanticipated facts or evaluations can be relevant to a pending choice.

Defeasibility inside games is by contrast narrowly constrained. The objective of the game, together with the constraints imposed on meeting it, determine what counts as a salient defeater: defeating conditions *cannot* come from just anywhere. That your queen might be endangered if you move your rook is a legitimate defeater, but that castles are ugly vestiges of feudal social structure is not; that the day is too hot for your border collie to complete the course at full speed is something you can reasonably consider in deciding whether to stick to the drill, but that bringing her around to the stands would allow you to show her off to your family and friends is not. In-game defeaters are anchored in the objectives and the rules of the game, which a player is apprised of up front, whereas defeating conditions for inference conducted in the out-of-game world might, for all one knows, be anchored in just *anything*. Thus in-game agency requires a more minimal kind of attention to defeating conditions, one that does not make the qualitatively remarkable over-the-top demands that inference imposes on reasoners in the wild. Putting that point the other way around, the library of agency is unlikely to prepare us for—or prepare us to *understand*—full-fledged defeasibility management.

If the library of agency is stocked with games, then the library's accession policies select for playability. Nguyen emphasizes the importance for playability of fit between the challenges that a game poses and the abilities it bestows on the players: games are fun to play when they are neither too easy, nor exercises in futility.[14] But defeasibility, if I am understanding the phenomenon rightly, is a mark of a deep mismatch between the complexity of the world, and thus of

14 But this is another observation that requires a complicated qualification. A great deal of what we do in our lives is boring routine, and so much of what we have to do does not nearly engage the abilities we are able to marshal. Surely here we will find another massive lacuna in the library of agency: How many games are going to reproduce the endless commutes, tedious errands, and all the rest of it?

I think that is correct, but the claim requires contouring. A great many quite undemanding games have the function of (merely) keeping one occupied: think solitaire, *Tetris*, and the seemingly endless variants on jewel-matching games. That said, while undemanding, they differ substantially in agential structure from the tasks that characteristically make

the problems it poses for the agents in it, and human competences. We can neither adequately represent the problem spaces we face, nor calculate the way our actions will play out in them, and so we need to anticipate the fact that, no matter how hard we try, we can all too easily turn out to have overlooked one or another vital consideration. To make room in our logic for having overlooked indefinitely many vital considerations is to treat inference and reasoning as defeasible. So it is no accident that Nguyen's library of agency gives short shrift to this aspect of it.[15]

As before, there is a complication to introduce: perhaps the experience of a game to which you are new allows you to experience something like the surprisingness of defeasibility in the wild.[16] There is something to this, but the point only carries so far, at least if you find plausible another admittedly controversial view that I will not now defend, but just put on the table. In the world at large, you have to learn what matters from experience, and there are no *a priori* boundaries we can place on what you might discover to be important—or unimportant. In the game of life, it is not that you know what winning *would* be . . . but then there are surprises about what it takes to get there and what you need to pay attention to on the way. Life would be a very different matter if it came with a rule book that told you what counted as a successful finish.

In a striving game, it cannot happen that you come to understand that the objective of the game is to be simply disregarded; even if a game is new to you, you know *a priori*, so to speak, that however surprising the connections you need to make, moves in the game are to be adjudged by their relevance to that objective. And that is the case even if, as in *Bag on the Head* (a party game that turns up in Nguyen's fascinating discussion), winning the game does not *matter*. In a game, there can turn out to be intermediate objectives that one does not initially realize are called for by the objective of the game, and these

---

up the background processes of everyday life: people play *Jewel Crush* in waiting rooms precisely because one's agential configuration *qua* player and *qua* waiting are *different*.

15 Although it is important to have this point in front of us, I want to emphasize that this is not an issue Nguyen himself overlooks. On the contrary, and laying out his train of thought, in a game, agents act strategically, on the basis of their own self-interest, as that is defined by the scoring rules for the game. Most of real life is hard to face up to because it is not like this; not only is your own self-interest not transparent to you, there is no presumption that other agents share your priorities and objectives. So once we have Nguyen's characterization of games on the plate, it is suddenly clear that moral theory, as analytic philosophers practice it (but it is not just them), is for the most part the very same fantasy of moral clarity purveyed by games, only less enjoyably packaged; it is suddenly clear that the sort of economic theory that we learn in that introductory econ class is a theory of decision making inside a video game, but not an account of choice in real life. This reframing sets an extensive and novel agenda for moral theory.

16 Once again, the point is due to Nguyen himself.

intermediate ends can give rise to surprises about what defeaters are relevant to some course of action you are considering within the game. In life as we live it, however, you can notice that something matters, in a way you had overlooked, and not because it serves some goal or other you are pursuing; rather, in view of what you can now see to be important, you may begin rethinking what your goals—your ultimate goals—are to be. That is, the intellectual demands imposed by defeasibility, in games and out in the world, differ in the direction they can require your thoughts to move: within a game, to notice a defeater is to notice a connection to an already given objective; in life, one can come by a new objective—a new final end, as the jargon has it—by noticing a defeater.

However, to the extent that a game you are just learning your way around does simulate the surprises you can encounter in real life, it tells us something about what drives defeasibility—namely, that to live life is to encounter the unfamiliar. If you had, as Andrew Marvell once put it, "world enough and time," not to mention the computational power, to familiarize yourself with everything there is, that sort of defeasibility would presumably gradually vanish. And since it never does vanish, what is made vivid by this qualification is how much the world is always new and unfamiliar to us.

IV

Turning to a fourth area in which the library's coverage is likely to be minimal, one of the very exciting contributions made by Nguyen's piece is the observation that, in playing a game for the sake of the experience of overcoming the obstacles to a goal (striving play, as opposed to "achievement play"), we adopt throwaway ends. This is an important contribution to the theory of practical rationality precisely in that it brings into view a hitherto neglected mode of agency.

But this mode of agency will also rarely or never appear *within* a game—as opposed to being invoked in order to enter the game in the first place. It is not that we cannot imagine a game in which a player must pause for a game within a game. ("In order you proceed, you must challenge Death to a game of chess!") But because the demands of playability so strongly impress objective-oriented structure on games, a game within a game will be played as a step toward the organizing end of the game it is in; that is, it will prompt achievement play rather than striving play. The rules of the game will not tell players to take time out to play another game, purely for the enjoyment of that game itself, rather than in order to advance toward the goal set by the top-level game. A game that did make the demand would be lackadaisical, and so annoying rather than gripping. Accordingly, the very mode of agency that is Nguyen's dramatic

contribution to the theory of agency and of practical rationality is not itself represented in the library of agency we are now considering.

There is a second layer to the problem. A precondition of agents inserting themselves into a game by taking on the objectives it specifies is someone having made up the game in the first place. That is, inventing games is something that agents do, and while I do not know that the activity counts as a natural kind within the world of agency, the categories under which it is natural to subsume it—invention more generally, one would think—look different, and are responded to differently, inside games. Consider *Sign*, described in Nguyen's essay, a game in which invention plays a prominent role or, again, the ingenuity that has been devoted at one time or another to coming up with new chess openings. Genuine invention introduces novelty; what is genuinely novel is likely to fit whatever concepts and rules one already has to hand poorly; scoring rules deploy the concepts one has to hand when they are being laid down. So for a game that demands invention to be playable, the novelty it elicits cannot be scored directly; certainly to adopt anything like the stance of a Nobel Prize committee within a game would make it confusing and frustrating. Instead, we score the cleanly designated objectives that the novelty is to promote: checkmates, in chess; successfully transmitting a given message, in *Sign*.

This means that in-game invention is in the service of previously designated objectives.[17] Now in the world at large, the most impressive innovations are, often enough, not too closely tied to antecedently available targets, and that is true of games as well: the games that are most likely to evoke novel forms of agency are also more likely to be products of the more freewheeling, less goal-focused modes of deliberation. So we should be concerned about the aspects of agency exercised in the course of inventing games being underrepresented within the library of agency.

V

When we engage in striving play, Nguyen points out, we adopt ends for the sake of the experience of struggling to achieve them, and he takes time out to push back against an anticipated objection, that these are not really the agent's *ends*, but rather some sort of second-rate imitation. Suppose that is right: we ought *not* to think in terms of a two-tier system, containing the properly so-called ends that agents adopt for real reasons, and then also the mimic ends they take on merely in play. Then in my view there is a possibility that it is methodologically important

17 But as discussed above, one can not only play games—one can play *with* them, as one does in *Minecraft*, or when making maps for other players, as in *Halo*.

to leave open and explore to the fullest—namely, that *all* of our final ends are ultimately underwritten by the capacities that make it possible for us to play games. Perhaps we take the goals that structure our lives seriously in the same manner that we take the goals that structure our games seriously, but because we have been immersed in our own lives for so long, we forget that this is where they came from, and that life is much more like a game than we for the most part imagine.[18]

Taking a sizable step back, we can see moral psychology to owe an account of how we arrive at our driving concerns, and also at the constraints—in the familiar language of Harry Frankfurt, the "practical necessities"—that channel the pursuit of our ends, and more generally our responses to those concerns. (If the term is new to you, a mark of a practical necessity is someone telling you that what you are asking them to do is unthinkable, and *simply* out of the question.[19]) We could not have been originally argued into them, and that is not just a belittling remark about the intellectual capacities of children. Traffic in reasons is itself constrained by its own rules, of that game as it were; it is directed sometimes by goals, and more generally by concerns that must themselves be acquired.[20] Our experience of the force of laws of logic is itself a practical necessity that is part of what is to be explained here, and so cannot be appealed to as the basis for that explanation.

Moreover, our society has in the past few centuries become very highly specialized. Inculcation into one of the disciplines that make up its fabric involves internalizing the priorities, standards, ideals, and guidelines that govern activity within it, and we cannot, for the most part, acquire these by being argued into them, either. These areas of expertise develop their own proprietary intellectual tools—concepts, first and foremost, but not only—and so the standards, etc., that must articulated using these tools cannot so much as be expressed by someone who has not gone through the requisite apprenticeship. That makes it hard to see how an argument for those standards, priorities, etc., could even be intelligible to an outsider.[21]

Nguyen is providing us with the ingredients of the sort of explanation we need. (To be clear, I am pressing his view in a direction I am not myself sure he wants to go.) Both in the course of one's upbringing and, subsequently, in the course of the training that makes a specialist out of a layperson, we summon up the dispositions that allow us to inhabit games. We find ourselves assuming the

18 For the methodological imperative, compare the remarks in Nietzsche, *Beyond Good and Evil*, sec. 36.

19 Frankfurt, *The Importance of What We Care About*, chs. 7, 13.

20 What might such concerns be, if not goals? For a start on the alternatives, see Millgram, "On Being Bored out of Your Mind."

21 The problems here are spelled out at greater length in Millgram, *The Great Endarkenment*.

mantle of ends, constraints, and so on that we are offered, with verbal guidance and other prompting; the picture is one in which immersion in a game captures, in sharpened form, the central aspects of immersion in life. And the picture reminds us that seriousness and playfulness are not mutually exclusive—on the contrary.

But if games are to be our model for our engagement with life itself, it is all the more important to keep track of the ways in which games are unrepresentative of agency across the board. Although life and games perhaps share an underlying source of motivation and commitment, they differ in being, respectively, conducted in an in-principle wide-open field of action, and in one that is closed off by its designated borders.

To reiterate, the issues I have been reviewing are not meant as criticism or complaints. In my view, Nguyen has done us a real service by identifying an important theoretical and practical resource. But it is important to be aware of its limitations, and that reminder has especial urgency for philosophers: as we know, when you give philosophers a new tool, it does not take very long for some of them to start insisting that anything you cannot do with that tool does not matter, and often enough they will start to insist that anything else literally does not exist. Put more abstractly, methods get reified into ontologies, and so if you lose track of the limitations of a method, it is all too easy to end up with impaired vision. But keeping the limitations of a new method in mind from the get-go can forestall that outcome. And that is why I have been attempting to supplement Nguyen's eye-opening observation, that we have at hand a library of agency made up of games, with the reminders about shortfalls in coverage assembled here.[22]

University of Utah
elijah.millgram@gmail.com

REFERENCES

Bowman, Margaret. *Are Our Goals Really What We're After?* PhD thesis, University of Utah, 2012.

Ferrero, Luca. "Constitutivism and the Inescapability of Agency." In *Oxford Studies in Metaethics*, vol. 4, edited by Russ Shafer-Landau, 303–33. Oxford: Oxford University Press, 2009.

Frankfurt, Harry. *The Importance of What We Care About*. Cambridge:

22 For helpful conversation, I am grateful to Svantje Guinebert, Gilad Kleinman, Hillel Millgram, and Michael Millgram, and for comments on drafts, to Chrisoula Andreou, Margaret Bowman, and C. Thi Nguyen.

Cambridge University Press, 1988.
Hlobil, Ulf. "Choosing Your Nonmonotonic Logic: A Shopper's Guide." In *The Logica Yearbook 2017*, edited by Parvel Arazim and Tomáš Lávička, 109–23. London: College Publications, 2018.
Horty, John F. *Reasons as Defaults*. Oxford: Oxford University Press, 2012.
Millgram, Elijah. *Ethics Done Right: Practical Reasoning as a Foundation for Moral Theory*. Cambridge: Cambridge University Press, 2005.
———. *The Great Endarkenment*. Oxford: Oxford University Press, 2015.
———. "On Being Bored out of Your Mind." *Proceedings of the Aristotelian Society* 104, no. 2 (June 2004): 163–84.
———. "The Persistence of Moral Skepticism and the Limits of Moral Education." In *Oxford Handbook of Philosophy of Education*, edited by Harvey Siegel, 245–59. New York: Oxford University Press, 2009.
———. *Practical Induction*. Cambridge, MA: Harvard University Press, 1997.
———. "Practical Reason and the Structure of Actions." *Stanford Encyclopedia of Philosophy* (Winter 2020). https://plato.stanford.edu/entries/practical-reason-action/.
———. "Relativism, Coherence, and the Problems of Philosophy." In *What Happened in and to Moral Philosophy in the Twentieth Century?* edited by Fran O'Rourke, 392–422. Notre Dame, IN: Notre Dame University Press, 2013.
———, ed. *Varieties of Practical Reasoning*. Cambridge, MA: MIT Press, 2001.
Nguyen, C. Thi. *Games: Agency as Art*. New York: Oxford University Press, 2020.
———. "Games and the Art of Agency." *Philosophical Review* 128, no. 4 (October 2019): 423–62.
Nietzsche, Friedrich. *Beyond Good and Evil*. In *Basic Writings of Nietzsche*, edited and translated by Walter Kaufmann, 179–436. New York: Random House, 2000.
Reutlinger, Alexander, Gerhard Schurz, and Andreas Hüttemann. "*Ceteris Paribus* Laws." *Stanford Encyclopedia of Philosophy* (2015). https://plato.stanford.edu/entries/ceteris-paribus/.
Suits, Bernard Herbert. *The Grasshopper*. 3rd ed. Peterborough, Ontario: Broadview Press, 2014.
Vogler, Candace. *Reasonably Vicious*. Cambridge, MA: Harvard University Press, 2002.
Watterson, Bill. *It's a Magical World: A Calvin and Hobbes Collection*. Kansas City, MO: Andrews and McMeel, 1996.
Wittgenstein, Ludwig. *Philosophical Investigations*. 1953. 3rd ed. Translated by G. E. M. Anscombe. Oxford: Blackwell, 1998.

Journal of Ethics and Social Philosophy
VOL. 23, NO. 3 · JANUARY 2023
https://doi.org/10.26556/jesp.v23i3.2709

GAMES UNLIKE LIFE

A REPLY TO CAMP AND MILLGRAM

C. Thi Nguyen

AWHILE BACK, I had been struggling to write about games in the established terms of analytic philosophy. Then my friend and longtime philosophical confederate, Jonathan Gingerich, explained the problem to me quite nicely. He said: our contemporary philosophical theories of value, rationality, and agency had been captured by the moralists. Our theories had been designed by ethicists and political philosophers to handle their very specific concerns. As a result, we have inherited a philosophical picture of ourselves as rigid, straight-ahead, and serious agents. And when we in turn try to think about the other kinds of activity—art, beauty, and play—we find them hard to analyze using our inherited theories. So philosophers tend to dismiss art, play, fun, and games as trivial. But that is not the fault of art or play. It is the fault of our inherited theories.[1]

In "Games and the Art of Agency," I tried to push that point—to show that there are complex, vitally important agential phenomena hiding right in front of our faces.[2] I wanted to show that there are elaborate structures of agency hiding in trivial-seeming activities, like party games and drinking games. In their excellent comments, Elizabeth Camp and Elijah Millgram have set out to complicate my story. Camp and Millgram are in accord with me about the main themes of the paper. They have been convinced, they say, that thinking about games does reveal a remarkable complexity of agency. But Camp and Millgram want to push me on the details in two very different directions. Millgram wants to emphasize the *artificiality* of games. In my picture, games are extremely rigid artifacts. They are explicitly formulated activities, where the goals are fixed, the permitted affordances wholly specified, and the space of reasoning precisely delimited. In that case, says Millgram, they are incredibly

1 This is not an exact quote. I believe we were walking from one bar to another at one o'clock in the morning, in New Orleans, between days of an aesthetics conference, when this conversation took place.

2 Nguyen, "Games and the Art of Agency."

distant from ordinary life. In real life, we have to decide on our ends; we have to negotiate and settle on our rules and norms. But in games these features are all set in stone, preestablished by the game. So playing games might be satisfying, fun, and beautiful—but, says Millgram, there are some severe limitations on what we can really *learn*, for use in the real world, from such peculiar and artificial environments.

Camp pushes me in the opposite direction. Perhaps, she says, game life is not really that distinctive or unique. According to my account, our gaming agency is supposed to be wholly quarantined from our ordinary agency. Also, our gaming agency is supposed to be peculiarly fluid and malleable, while our enduring agency is more stable. But, says Camp, things are not actually so neatly divided up as all that. Our gaming agency is not actually so different or isolated from our enduring selves as I make it out to be. For one thing, says Camp, our enduring goals constantly influence our in-game actions. For another, our "real" agency turns out to be fluid and ever changing, rather than some fixed monolith. To put Camp's delicate suggestions into my own, possibly more dramatic, terms: perhaps it is the serious, enduring, neatly coherent self that was the illusion all along. Perhaps we are fluid agencies all the way down.

In what follows, I am going to quibble with some of the details of these challenges. I am not going to address every challenge; I would rather take it slowly through the most interesting points of contention. But let me stress, at the outset, where we agree. We agree that games are incredibly sharp crystallizations. Games are artificial structures that take what is ambiguous, negotiated, and fuzzy in normal life, and force it into an explicit mold. And I think it will turn out that our enduring agency is a lot more game-ish—one might say, a lot more *playful*—than we might have otherwise thought. But my purpose was never to argue that this fluidity of agency was some strange and peculiar capacity, uniquely deployed in games. I think our agency is often fluid. What I wanted to show was that games highlight that particular aspect of basic human agency by formalizing agential fluidity. So, when we study games, we are forced to confront a particularly crystallized version of this essential part of our nature.

In fact, thinking about Camp's and Millgram's comments—and Gingerich's—I am tempted toward an even stronger formulation. In the standard philosophical framing, it turns out that our real selves show most truly in moral and political life. In this framing, games look quite peculiar. Games look like this odd, liminal space, where we step back from our usual mode of quite stable agency and allow a brief moment of fluidity. But perhaps the standard framing gets things the wrong way around. Perhaps we are deeply fluid, ever-changing, and malleable things. Perhaps it is in games and play that our real selves are more deeply exposed. And perhaps it is the enduring, static, committed self that

is more of an illusion. Perhaps this presented stability is an artifact of how are forced to represent ourselves in political negotiation—a fiction generated by the social demand for us to appear as relatively stable, so that our vote may be counted, our desires satisfied, and our wishes represented. The appearance of a stable proxy self might be something we construct so that we may take part in the practices of contracts and negotiation and governance. And games might be especially important to us now—as the institutionalized beings we have been shaped into becoming—because they are a space where we are allowed to let go of those strictures and relax into our more deeply fluid natures.[3]

1. The Artificiality of Game Life

I claim that games can help us learn new forms of agency that can come in handy in real life. In order to be useful, however, the kinds of agency on offer in games must adequately resemble the kinds of practical agency that we use in real life. But, worries Millgram, the essential nature of games—their artificial clarity—makes them crucially unlike real life. So the forms of agency we might learn in games is far less applicable, and so less useful, than we might hope.

In real life, says Millgram, practical reasoning happens against a blurry and dynamic landscape. So many of the key reference points are negotiable, unknown, or in the process of development. In real practical life, our goals are not set in stone. We can deliberate about our ends, deciding what we really care about. We can come to see that a long-cherished goal is actually worthless, or discover something new to value. But in games, our goals are nonnegotiable. At most, we can deliberate about the instrumental value of midlevel goals. In a game of chess, I can think about whether an advantage in material or in position would be the best way to win. But those deliberations over midlevel goals are always conducted against the backdrop of an entirely fixed final goal: winning in the terms specified by the game. In games, we do not deliberate over our deeper ends, only our midlevel, instrumental ends. We do not deliberate over what really matters, only how to achieve it.

3 I am influenced here by James Scott, who suggests that states—large-scale bureaucratic structures—can only process and see those parts of the world that are easily put into the terms that institutions can process—standardized, quantifiable, regular (*Seeing Like a State*). They can only see the parts of the world that are legible to large-scale bureaucracies. He suggests, then, that states have an interest in making the world more legible to them by evening it out. My suggestion here might be put in the following way: that the stable self is itself a useful legibilization of a more strange and fluid thing that we might have been. I am also influenced here by Annette Baier's suggestion that the practice of contracts is a very odd and specific one, optimized for relations between relative strangers who wish to exchange goods ("Trust and Anti-Trust").

Millgram's observation would be very worrisome if we had little to no voluntary control over which games we played. And that might be some folks' experience of some games. In some communities, participation in sports, say, might arise from inescapable social pressures. But that is not the scenario I was imagining when I suggested that games might be able to give us expand our autonomy. According to my account, many of the development advantages of games depend on interacting with a variety of well-bounded games. We play games, we stop playing them, we try out other games. But the precise features that are valuable in such well-bounded games—their value clarity, their explicit rules—can be toxic when instantiated in pervasive or inescapable real-world systems. The gamification of education and work, for example, turns out to undermine agency. For example: Twitter enshrines certain communicative goals in its metrics—likes, retweets, and follows. But those goals are preestablished and nonnegotiable. So when we internalize those goals, we actually undermine our autonomy.[4] But, I want to suggest, there is a very different—and much healthier—relationship we can have with games in which we rehearse the process of deliberating about our deeper ends.

Think about how people often play games for leisure, fun, and aesthetic satisfaction. You read a bunch of reviews of games describing the different experiences you might have. This game is fun, that one absorbing, this one genteel and relaxing, that one a fascinating simulation of how epidemics spread. (Really—*Plague Inc.* is a great little iOS game, in which you can play as a variety of infectious diseases out to kill humanity. As the game progresses, you choose from a variety of "level-up" mutations, which change how you infect, spread, and kill. I eventually figured that if I became too infectious and too deadly, then I would just wipe out a couple countries and burn out before I could kill all of humanity. And if I kill too quickly and dramatically, then those humans will panic and close the borders. You need to be pretty sneaky and slow for optimal lethality.)

Once you read the reviews, you pick a game and play it, and then you find out whether it really is fun, absorbing, or beautiful. And sometimes you will discover that a game is valuable (or terrible) in a way that you did not expect. You might discover that this interesting-looking game actually forces you into a boring exercise in painstaking micro-optimization. Or you might discover that in the seemingly silly party game *Codenames*, you end up having to model the shape of other people's networks of conceptual associations, and this process is far more interesting than you had guessed. And after you play a game and make these discoveries, you make more decisions: whether to play that game

4 Nguyen, *Games*, 189–215, and "How Twitter Gamifies Communication."

again or sell it, whether to froth online about how terrible it is or become an obsessive fan of that game designer.

Let me retell that same story, but cast into more philosophical language. In my analysis of the motivational structure of game playing, there is a crucial distinction between the local goal and the larger purpose. The local goal is the thing we aim at during game play ("collecting gold tokens" or "making baskets"). The larger purpose is the reason we are playing the game: to get exercise, be a winner, have fun, relax, find beauty and thrill in the movement. And in striving play, local goal and larger purpose come apart. During the game, I am trying to win, but winning is not my larger purpose. My larger purpose is, say, to get some exercise and destress.

Notice that the game sets the local goal we will pursue inside the game, but it does not set our larger purpose for playing it. A route setter at a rock-climbing gym creates a climbing problem that emphasizes delicate and painstaking footwork. One climber repeats that problem because it helps them train, refining their footwork. Another climber relishes the graceful movement the climb evokes. Another climber wants to show off their flexibility to their friends. Another one just wants to climb everything in the gym because they are keeping a scorecard. All of these climbers are playing the same game with the same local goal, but for different purposes. And that purpose can shift. Maybe one climber starts climbing the problem to improve their foot technique, but after some teeth-gnashing fumbles, starts to discover something unexpected—that they can be graceful, and that the feeling of gracefulness is its own delight.

The aesthetic practice of trying out different games, then, involves moving between fixed local goals and larger, more open-ended purposes. That is: I adopt a local goal and follow it rigidly for a small amount of time. I then back up and reflect on the value of the activity in an open-ended way. Maybe I dive back in and play the game again, and then step back and reflect again on the value on offer, and whether it is worth it. Notice that two kinds of deliberation about ends are going on at once. First, I can deliberate about the purposes for which I might take up the activity. That deliberation is entirely open ended. The act of aesthetic reflection on striving play emphasizes this form of deliberation. I think about the wide range of *values* available in the activity of game playing: fun, fascination, challenge, exhilaration, catharsis, discovery, improvement, intensity, glory, elegance, comedy. And a player can *discover* new forms of value available through the process of play. Before playing *Galaxy Trucker*, I had not known that there could be a glorious comedy to slapping a machine together and then watching my hastily jury-rigged contraption fall apart. Once I have discovered these new joys in the game, I decide whether it is worth engaging in the activity again. I decide whether that particular value, and that particular

instantiation of that value, is worth the time and effort. The first form of deliberation over ends we can find in game play, then, is in deliberating about the *different purposes for which we play different games.*

Second, I can also deliberate about the local goals in games and how they inspire a particular experience of play. I suspect this second form of deliberation is less common than the first; it involves taking on a game designer's frame of mind. When I aesthetically reflect on the design of a game, I am reflecting on how the fixed features of the design shape the resulting activity and what values might arise in that form of activity. I see, for example, that the goal of Imperial is to manipulate the course of World War I for profit by changing around my investments in the various countries involved, and steering their military encounters. I can see how this goal leads to fascinatingly tangled allegiance structures, and how much less interesting it would be if the goal were simply to guide a particular country to victory. In other words, I can see how the pursuit of a particular specified goal informs the texture of the activity of pursuit. And I can see how pursuing slightly different specified goals might change the activity of pursuit—by trying my hand at some game design, or simply by playing a number of mechanically similarly games with subtly different goals.

Take, for example, Reiner Knizia's beloved series of tile-laying games, especially *Tigris & Euphrates*, which is generally considered a masterpiece of European-style board-game design. As is typical in Eurogames, the player attempts to collect goods from a number of different categories. In many other Eurogames from that era, the player's goal is simply to collect the most goods—with, perhaps, some bonuses for collecting sets of the same category. But in many of Knizia's games, your score is determined by how many goods you have *in the category in which you have the least goods*. That is, you are scored on your weakest category. You cannot make up for having failed to collect any farmer tokens by collecting a large number of war tokens. This scoring structure forces players to maintain diversified portfolios. You do not spend much time thinking about your best categories, but fretting over your weaknesses. Because of that victory condition, the way to attack your opponents is to figure out their weak spots and deny their attempts to shore them up. So play becomes much more about protecting your weakness and exploiting your opponents' than simply about making a lot of points really fast. The weakness-oriented design helps encourage a deeply interactive form of play.

Games let us experience how a slight variation on the game's victory conditions will change the experience of play. Games thus permit a second kind of deliberation about ends: deliberation about the *selection of local goals*, and how the precise articulation of a local goal can inform the texture of the activity of its pursuit. Since the activity of pursuit is the locus of value for striving players,

deliberation about local goals flows into deliberation about larger purposes. That is, we can see both how Knizia's particular selection of the goal inspires the vulnerability-centric activity of playing *Tigris & Euphrates*, and then see how that kind of vulnerability-centric activity gives rise to a particular kind of value—in this case the value of cognitive-absorption interplay of differing player weaknesses. Games help us see how a specification of a local goal can shape the activity of its pursuit, and how that shaped activity can foster distinctive forms of value.

What I am suggesting is that games can model a kind of life deliberation that has been, in fact, best described by Millgram himself. In his wonderful book, *Practical Induction*, Millgram argues that we cannot figure out our values by deducing them from some abstract conception of the good. Rather, we discover which values are good for us to have through practical experience. We choose a value and try living life with it for a while, and see how it goes for us. We discover that a life lived under one value makes us miserable, compressed, annoyed, and that a life lived under another value makes us happy, alive, vivid. In a later paper, Millgram offers a slight variation of this picture. In "On Being Bored out of Your Mind," he argues that we cannot be identified with our desires because we change our desires all the time. We shift desires based on the experiential feedback of how our life goes when we follow these desires. When we pursue a desire and feel interested and engaged, this is a sign that it is a good desire to have. When we feel bored, it a sign that this desire is a bad one for us, and that we should, as he puts it, excrete out this desire and find a new one. This is the psychological dynamic behind changing hobbies or majors and midlife crises.

Let us elide some of the complexities here and treat both values and desires as forms of ends. Millgram is suggesting that we deliberate about our ends through experience. Of course, for Millgram, it is not just the end itself that is under assessment. It is the way the pursuit of that end shapes your life, partially through the roles you assume and activities you undertake in your pursuit of that end. So, it turns out, whether an end is good or bad for you depends on your psychology, your ambient culture, and the roles and positions available to you. The selection of an end interacts with your personality and your environment—the particular practical possibility space you happen to inhabit—and drags you into a specific form of life.

I myself have tried on many different values during my life. I have valued making money, contributing to the advancement of neuroscience, being a successful tech entrepreneur, writing interesting novels, becoming a good food reviewer, being a successful philosopher by the standards of a particular ranking system, attaining more complex yoga poses, writing interesting philosophy, getting better at fly-fishing, getting better at rock climbing, becoming really good

at chess, aesthetically exploring board games, aesthetically exploring perfume, and learning to cook a variety of cuisines. Each of these goals pulled me into a radically different form of life. Trying to make it as a tech entrepreneur involved constantly sussing out business possibilities, constantly scanning the world for unexploited potential. Trying to be a good neurobiologist turned out to involve learning an enormous amount of biochemistry and anatomy and getting good at dissecting mouse brains. Trying to be a good food reviewer involved driving around Los Angeles, getting familiar with the ethnic neighborhoods of the city, eating food, and trying to come up with new ways to describe really delicious fried shit. Trying to climb the ladder of philosophical status by publishing mainstream epistemology in fancy journals meant reading piles and piles and piles of Gettier epicycles and getting into a lot of technical hairsplitting about formal definitions. Getting good at chess involved memorizing openings and practicing sharp look ahead. Learning to fly-fishing involved a lot of wandering around in the silent woods and staring intensely at flowing water—which turned out to be a strangely meditative practice. Getting good at rock climbing involved long road trips with friends, lots of camping, and then intense attention paid to minute details of a rock face—which turned out to heighten my visual sensitivity to nature. Trying to learn to cook Korean food turned out to involve learning a lot about pickling and dried chiles—and, it turned out I could not get certain ingredients because I live in Utah, so the attempt to cook Korean food gave me a reason to grow certain herbs and vegetables, so suddenly I was researching composting techniques and kneeling in my backyard weeding every weekend.

In each of these cases, setting a particular end for a particular person in a particular circumstance drags in all sorts of other changes to their lifestyle and attitude. To deliberate about ends, in Millgram's practical and experiential manner, is to try out living life under a particular end, and then seeing how it goes—how that form of life feels to you— and then asking yourself: Is it worth it?

What I am suggesting is that this complex, open-ended deliberation about ends is modeled in the practice of aesthetic striving play and reflection on that play. We deliberate about ends when we play different games and then ask ourselves if taking up those ends yielded a good, satisfying, beautiful, interesting, or otherwise valuable form of life. In games, we take up specified goals inside particular assemblages of ability and environment. Games show us, in a particularly schematic and crisp form, how different specifications of local goals can generate different forms of activities with radically different textures. And games give us an opportunity to reflect on the value of these different forms of activity. The process of playing *many games*—trying them out, reflecting on

them, and choosing which to play again—is a compressed version of Millgram's practical induction. This exposes one of the truly remarkable and special features of games. It explains why games occupy a special place in the dizzying array of human practices. In what other activity do we so concentrate our gaze upon the relationship between a particular goal and the activity of its pursuit? Where else do we try out so many variations, and where else is it so easy to see precisely how a goal shapes a pursuit, and shapes the ensuing richness or poverty of activity? The reflective game-playing practice is, in fact, practical induction crystallized.

To sum up: Millgram's primary worry is that since each particular game comes with fixed ends, game players do not deliberate about ends. My response is: but *which* games we play are not fixed. And since we have a *choice* of games, we have a choice of ends, on two levels. We do not confront the local goals of games as entirely nonnegotiable givens. While the goals are fixed in any particular game, we do have a choice of which games to play—and thus a choice of goals. Furthermore, we have a choice about which *purposes* we seek in play, in reflecting on the value of playing different games. This offers us the particular experience of deliberating over the larger purposes that are fulfilled by our pursuit of narrower, more tightly specified in-game ends. We get to decide whether we want to play for relaxation, thrills, or intellectual absorption. It also offers us the opportunity to deliberate over which formalized ends we wish to adopt to achieve our larger purposes. Games were never supposed to be a perfect reflection of nongame practical life, but a crystallized, concentrated, controlled model of it—an *art* of agency. Games model both the process of deliberating about larger, more open-ended purposes, and model how the choice of some particular shorter-term, local goals might shape the larger values that emerge.

My claim here is not that any single game can encode this kind of deliberation about ends. If somebody forced on me me to play a particular game, I would, as Millgram worries, never practice deliberation about ends. This kind of lifestyle might, by immersing us in a single hyper-clear value system, plausibly work to undermine our capacity to deliberate about ends. This is exactly why I think the gamification of pervasive real-world systems—like work and education—is actually corrosive to our autonomy.[5] The important question, for me, is: What does access to, and a rich engagement with, the *library* of agencies encourage and foster? My answer is that engagement with a wide diversity of games, conjoined with the proper kind of deliberation about the value of those games, models deliberation about ends. So the reflective opportunity here is not the result of playing *a game*, but from the practice of playing

5 Nguyen, *Games*, 189–226, and "How Twitter Gamifies Communication."

a diversity of games and reflecting on them—for example, as we might find in the practice of exploring and aesthetically evaluating a wide swathe of games.

But this is not a limitation unique to games. Throughout many arts, we see a similar pattern. Insofar as the arts might aid personal development, that development requires not just engagement with one piece of art but diverse consumption and reflection. Suppose you think that a novel can encode a particular emotional perspective. It seems doubtful that exposure to a single novel would bring about any significant moral growth or help develop any significant ability to see the world from many perspectives. But access to a whole library of differing emotional perspectives, along with some complex reflective integration, might plausibly foster perspectival flexibility. Similarly, the library of games is a powerful resource for practicing the deliberation of ends, but only for those users willing to make a substantial investment in exploring that resource.

Millgram offers other criticisms in a similar key. For example, Millgram worries that in real life we renegotiate rules, but in game life, we do not, because the rules in games are fixed. My response comes along similar lines. There are activities in which we do change the rules—including game designing, house rules, and all other sorts of game-hacking activities. Whole communities are devoted to modifying popular games and finding new ways to play existing video games, like speed running. Both practices modify or add rules to existing games. But we do not even need to modify games to reflect on what the rules do. The practice of playing many games involves seeing how different rules lead to different sorts of activities and different forms of life. This is not a practice of negotiating rules, exactly, but it provides a crucial resource for thinking our way through rule negotiations. Playing games lets us quickly explore how different rules give rise to different activities. When we choose to play a particular game, we are choosing to accept a rule set. And since there is such a variety of games—many of which are only subtle variations on other extant games—then, in picking which game to play, we are picking which particular rule set to adopt from a constellation of closely related ones.[6]

I will not take the time here to respond to every flavor of Millgram's criticism; I think the general drift of my take should be clear enough. The larger

6 Camp also offers one worry in this spirit: that real life requires us to be actively flexible, that we know how to apply the right agential mode and know how to tweak it. But how could we learn this when games offer us activities under fixed and highly specified agential modes? My answer will be in a similar key: playing a variety of games, and aesthetically reflecting on them, can contribute some resources toward developing flexibility by giving one a tour of the variety of ends, modes, and practices available. But that is only a resource for the development of flexibility and adaptability; it surely does not guarantee that development.

theme here is that Millgram is worried about what is fixed in games: rules, space of reasons, ends. Since they are fixed, he worries, then the kind of reasoning we do in inside a game will be unlike the more open-ended form of practical deliberation in real life. My response is: Millgram's worry only holds for the reasoning we do *inside* a particular game, once that game is chosen. But the experience of playing lots of games is an experience of variation across those fixed elements. If we play enough games, what we will experience is *what happens when you vary those fixed elements*—when you try on different rules, different goals, and different sorts of reasoning. The act of choosing between games is one in which we deliberate about which collection of rules, reasons, and ends we wish to inhabit for a while. This may not be immediately applicable to our deliberation about real-life ends. But it is, I suggest, a model of such deliberation. To play games, and then reflect (aesthetically or otherwise) on the value of the activity, is to practice a version of practical induction. What it loses in precise fit to real life it can make up for in the speed, rapidity, and wideness of its experimental submersions in differing agencies.

2. Is the Gaming Agency Really Quarantined?

Camp's worries come from the opposite direction: that games are more like, and more integrated with, ordinary life than I suggest. Camp has two distinct worries. My responses will eventually converge into a single picture, but let me start by taking Camp's worries one at a time.

First, Camp worries about the degree of *quarantine* between gaming life and nongaming life. In my account, striving play can involve a kind of agential submersion. I decide to play chess for the purpose of total cognitive absorption in the struggle to win. To get that particular experience, I need to forget my larger purpose for a while and absorb myself in the local goal. If I recalled my larger purpose, then I could not entirely absorb myself in the pursuit of the local goal, because the local goal and larger purpose often suggest opposing actions. My example from *Games*: if a player's larger purpose is to have an interesting struggle and that purpose guides their particular actions in the usual way, then they should pass by opportunities for quick wins, since a win would end their interesting struggle. In that way, being perpetually guided by your larger purpose can undermine your ability to obtain it. To have a certain type of absorption in an interesting struggle, you cannot aim at having absorption in an an interesting struggle—you have to just aim at winning. In these cases, then, our gaming agency must be significantly disconnected from our enduring agency. In games, you shut out the larger reasons from your enduring agency and absorb yourself in a more local interest in winning.

But, Camp responds, in real life, our game-playing agency is not usually so utterly quarantined from our enduring agency. When we play, our enduring reasons actually often penetrate into our in-game decisions. We care about our friends' emotional reactions, the general fun level of the social gathering. We modulate our in-game actions, sometimes abandoning the all-consuming pursuit of the win for enduring social reasons—like preserving a friendship, or making things more fun for a frustrated friend. We see that a friend is getting a little upset and we avoid strategies that might humiliate them. Our larger purpose for play—like having some light social fun and togetherness—often directly informs our choice of in-game actions.[7] So external reasons often leak into the inner, game-playing agent.

Let me offer a handful of fussy qualifications. I am happy to take onboard the observation that there are many cases of less quarantined instances of game playing. My claim was never that *all* striving play has to involve such strict quarantine, but only that strict quarantine is possible. So Camp's observations—that, in many circumstances, we do not invoke the strictest quarantine—are compatible with my own, as she herself points out. My argument in *Games* is an analysis of how striving play should proceed *if your purpose was complete practical absorption in the instrumental struggle*. And, as Camp's examples show, that is not always our purpose. Convivial social play is often oriented toward other goods, and so does not require such complete absorption.

But I do want emphasize here that there are plenty of other contexts of play where players really do seem to want that complete practical absorption, and do not seem to modulate their gaming actions for social considerations. Totally absorbed play is, perhaps, somewhat unusual in casual, social game playing. But total practical absorption is common elsewhere, especially in contexts built to support devoted, intense game play. I am thinking of things like chess tournaments, *Magic: The Gathering* tournaments, and the Olympics. Consider, too, online games like *Dota 2* and *EVE Online*—known for their vicious, no-holds-barred play environments.

In many social play circumstances, we find ourselves in a social group organized along some other axis than the pursuit of the aesthetics of absorbed playing. We are at a family gathering, or hanging out with old friends. Such groups typically consist of people with varying degrees of interest in the joys of practical absorption and varying levels of skill. In those circumstances, we

7 Quill Kukla makes a similar criticism in an *Analysis* symposium discussion of my book, *Games* (Kukla, "Sculpted Agency and the Messiness of the Landscape"). My response here touches on some of the same themes, though I have tried to offer a somewhat different angle for variety's sake.

often play games to aid in convivial socializing—and to achieve that purpose, we often modulate our quest to win for the sake of the larger social purpose.

But in some other, more gaming-centric contexts, people often gather precisely for the sake of absorbed and intense play. And we often build into such contexts systems to ensure that people with similar skill levels are matched against each other so that nobody has to hold back.[8] These environments are ones where it seems reasonable—and desirable—to permit yourself total absorption. And that demonstrates my primary point: that deep quarantine is, at the very least, possible.

But I think we can uncover even more interesting phenomena if we look more closely at cases where players do, in fact, modulate in-game decisions for social considerations and other extra-game reasons. Camp here is interested in how some of our enduring reasons can intrude into the game space. But my original argument was never directed at showing that we excluded *all* enduring reasons. It was directed at how we set aside *specific* enduring reasons, especially those whose inclusion would interfere with successful achievement of our real purposes in play.

Here, we need to distinguish between two different ways in which we break quarantine, so to speak—two ways in which the enduring agent's reasons can directly inform game-playing actions. A game rule can be thought of as directing us to *bracket* a certain set of our reasons by directing us to exclude those reasons from consideration while taking on the game's specified agency. We can look at two separate forms of intrusion:

1. Intrusion by non-bracketed reasons into practical deliberations in a game
2. Intrusion by bracketed reasons into practical deliberation in a game

Let us start with the intrusion by non-bracketed reasons. Think about considerations of style. I am, for the most part, a person who thrives on chaos and improvisation. My friend (and frequent board-gaming companion) Andrew thrives on precise planning and micro-optimization. We typically import our personal sense of style into our game play. I value creative, chaotic, edge-of-the-seat life experiences and creative, slapdash actions. Those values shows up in my play choices. I tend to play wild, big, over-ambitious moves, which often collapse on me—thought I suspect that I play this way precisely because I enjoy the process of desperately improvising my way out of the broken remains. Andrew values controlled, well-planned environments and sequences, and those values show up in his play choices, as he tends to make plans that are uncollapsible

8 Nguyen and Zagal, "Good Violence, Bad Violence."

and uninterruptible, and make the game space more controlled. Partially, he does it because he enjoys the experience of winning by seeing a meticulously laid plan come off like clockwork. In these cases, we certainly import some of our external values into our gaming choices. But, in many games, those external values were never bracketed out in the first place. Games can leave a space open for the player to import their differing interest in, say, chaos versus order.

Note, though, that some games do instruct us to bracket out those very same values and styles. Some games allow no creativity or chaos at all—like *Canabalt*, which is a reflex-based endless runner that only gives you one action and one affordance: perfectly timed jumps. *Canabalt* gets me to bracket my interest in creativity and self-expression and just focus on precisely timing my jumps, which is interesting for me, since it involves setting aside one of my most cherished values. Games can direct us to bracket certain reasons implicitly or explicitly. When a game directs me to help my teammates and hinder the other side, it is explicitly telling me to bracket my usual social relationships. But *Canabalt* does not explicitly tell me to bracket my creative style through a direct specification of a goal or rule; rather, the limited structure of affordances leads me to bracket my interest in creativity.

We can build these observations inside the context of Camp's observations to offer a more refined story than the one I offered in *Games*. A game *specifies certain aspects of agency* and leaves others unspecified. This is what I was gesturing at—but failed to adequately develop—when I said that a game specifies an "agential skeleton."[9] The game supplies a skeleton, and then each player puts their own flesh on those bones. So when we occupy an in-game agency, we take on the goals it specifies and bracket some of our enduring reasons. But, since the specification is skeletal, we can import other parts of our personality and agency into the open spaces, filling out those parts of the playing agency left unspecified by the game.

The rules of basketball specify that I will cooperate with these people against those people. In doing so, it asks me to bracket my usual relationships with certain people. I am to bracket the fact that, in real life, this person is my friend and that one my nemesis. I am to pay attention only to whether or not they are on my team or the other team in deciding whom to help and whom to hinder. But basketball's rules leave unspecified whether I should play flashily or carefully. So I get to choose, and I am free to import my external preferences. Thus, the player and the game together generate an alternate agency in the game. Some importation of our enduring values is quite common in game playing.

9 Nguyen, "Games and the Art of Agency," 423, and *Games*, 17, 52, 158.

But a game can also exclude almost any part of our enduring agency through its specification. *Canabalt* can get me to ignore my love of creative self-expression; soccer can tell me to put aside my love of doing things with my hands. In particular, competitive games direct us to bracket our usual desire to support other people's actions and act selfishly.

This bracketing of sympathy is particularly interesting. In most such games, we are supposed to turn into wholly selfish beings, uninterested in helping others. This selfishness is often not written directly into the rules, but presumed as part of the background of standard gaming practice. This practice is so natural and pervasive that it can be invisible, so we have to do a little work to foreground it. So: I often play games with my spouse and many friends. I am, in ordinary life, partial to my spouse. I usually take her interests to be more important than the interests of other people, especially strangers, and I will often protect her interests when they are threatened by strangers. But I bracket that partiality in many games. Imagine we are playing a standard competitive game where we are all supposed to be playing for our own victory. But then I begin to assist my spouse, taking it easy on her, or giving her resources from my collection. For many game players, this would break the proper spirit of a competitive game. Many games are fragile and fall apart if all the players are not behaving with egalitarian selfishness. In many such games, to have the kind of interesting struggle players are interested in, all the players need to behave as wholly self-interested and equally antagonistic toward all other players—at least, until in-game conditions change that balance. (In such contexts, you are allowed to treat another player partially because they just gave you a sweet deal in the last turn. You are not allowed to treat another player partially because they promised you a back rub after the game.)

So games ask us to bracket certain enduring reasons. The really interesting part, then, is not just that we sometimes import parts of our external agency into games. That is normal and unsurprising when those parts of our agency have not been bracketed out by the game. What is really interesting is that *sometimes we override the game's requested bracketing*.

Suppose that we are playing for collective fun. Because of the sort of gaming experience we are all interested in, and the kind of game we are playing, we bracket our interest in collective fun and put foremost in our minds the desire to win. Yet still, as Camp points out, our external interests can sometimes break the bracketing and change how we act. How is this possible? Importantly, there is no direct logical conflict between my interest in winning and my interest in having a collective good time. The two are logically compatible. In fact, my interest in winning is partly justified by my interest in our collectively having a good time. It is merely that, in some circumstances, I need to *exclude my larger*

purposes from the set of considerations from which I am actively reasoning in order to achieve those larger purposes. But Camp's modulation cases show that sometimes we do break that quarantine and act in light of our larger purposes. Suppose we are playing an intensely competitive game for collective fun. I notice that my friend is profoundly miserable and floundering, and the best path to victory for me would be to deprive them of a crucial resource that would completely undercut their position and leave them without any interesting actions for the rest of the game. I might, very reasonably, avoid that action specifically for enduring social reasons. Here I am acting on social considerations that I was supposed to have bracketed during the game. But how could I do that if I was supposed to have bracketed them and excluded such enduring reasons from my consciousness?

We explain our ability to act on excluded considerations by postulating what we might call a *flickering agent*.[10] When we *flicker* during a game, we occasionally poke our heads up out of the inner gaming agent and return to the enduring agent's perspective. If we see that we are failing in our purposes—like that nobody is having fun—then the enduring agent can change the inner agent's goals, or abandon the inner agent completely. This model fits both my own observations about the need for absorption, and Camp's observations about the frequency of social modulation. And it fits, at least, my own experience of play. I am often absorbed in the intricate calculations of the game, but I also occasionally step back from those calculations and take a second to survey the faces in the room. It is possible to rationally and reasonably flicker between the two perspectives because of the logical compatibility of the enduring purposes and local goals. I exclude my larger purposes from consideration, not because of some logical contradiction between the larger purpose and the smaller goal, but because of a psychological constraint: that I cannot have the particular experience of absorbed focus until I exclude certain larger considerations from my reasoning stream. This psychological constraint means that I cannot simultaneously occupy the absorbed instrumental stance and the stepped-back, enduring stance. But I can get both the goods of absorption and the goods of reflection by quickly snapping between the two stances.[11]

10 I discuss the flickering agent at greater length in Nguyen, "The Opacity of Play."

11 I have also entertained an alternative to the flicker model, what we might call the simultaneous-layers model. Here, our enduring agency runs in the background—something like a computer operating system—while the gaming agency runs in the foreground—something like a program I have open. The gaming agency dominates our awareness, though the enduring agency is still running at the same time and is capable of noticing things and breaking through. Though the simultaneous-layers model is different in its psychological details, it is logically equivalent to the flicker model in the current dialectic.

Finally, I am not arguing that the flickering agent is the only way to play. I think there are all sorts of different possible modes of play. We can play in a fully transparent and unquarantined way—with no absorption in an inner agent, constantly in the light of our larger purposes. (Such a player will be good at tending to the social needs of their friends, but have less absorbed fun in certain types of games.) We can have a deeply absorbed agent who, during play, cuts themselves off entirely from any awareness of their larger purposes. (Such a player will be really good at having that absorbed fun, but sometimes miss the fact that their friends are having a really terrible time.) And we can have an agent who is mostly absorbed, but flickers out of it to check at some rate. (Such a player splits the difference between the two extremes.) Different players and different play contexts support different modes of play.

Everything I have said so far concerns the way that our enduring agency might inform our inner agent. But the real oddity of games lies in how the reasons flow in the other direction—in the limitations on how the inner reasons of our gaming agency might influence our enduring agency. The truly fascinating oddity with games lies in how my interest in, say, winning over my spouse—in cutting off her plans and vanquishing her—are *cancelled entirely* when I leave the game. They have no animating power outside of the specific context of the game. This shows that the interest in winning, for a striving player, is not an enduring end, but something more peculiar. It is a temporary construct. It is here where we see the most potent form of quarantine.

Of course, there are plenty of locally active instrumental reasons. I am trying to fix my torn pants, and so acquire an instrumental interest in finding the right thread. The interest in finding the right thread ends once I am done fixing my pants. Low-level instrumental reasons like this flitter in and out easily. What is interesting is that the interest in winning presents itself with the phenomenology of a final end during the game, but that interest is cancelled the moment the game ends.

So my most important reply to Camp is this: we should not think that the inner agent is wholly quarantined from the outer agent such that no reasons cross between them in any direction. There is, rather, a limited and specific kind of quarantine, which works differently in different directions. On the *inbound* direction, we bracket off some of the enduring agent's reasons and prevent them from showing up for the inner, game-playing agent's deliberation. We do it sometimes so we can achieve certain effects, like absorption. In this way, gaming agency is much like other kinds of practical screening, where we exclude certain reasons from our mind to achieve a certain mental focus. But the more profound form of quarantine happens in the other direction—in the *outbound* direction. The truly odd features of our gaming agency lie in how the gaming

reasons are confined to a particular context. The gaming reasons *do not reach out* to ordinary life in an interesting way. This is not a mere pragmatic firewall, where we exclude relevant reasons for the sake of cognitive finitude. For striving players, in-game reasons—which appear as final, and rule with the power of finality over our in-game agent—simply do not reach out into our nongame life.

There is, of course, some relationship between game life and nongame life, but it is of a very complex kind. My spouse and I have been regularly playing a very nice strategic card game, *Res Arcana*. We are both striving players. Suppose our enduring interest is to have fun. But we adopt a temporary interest in winning in order to have fun. Notice that, in the game, our rational structure is centrally guided by an interest in winning. Our interest in having fun is psychologically bracketed, though it is still central to justifying why we have adopted the interest in winning.

Outside the game, I may take actions that will impact my in-game experience. But, insofar as I am a striving player, I will *take the kinds of actions that serve my enduring interest in fun*, and not the kinds of actions that will serve the in-game interest in winning. Let me elaborate on one of my old examples. Suppose I find a strategy guide for a game. If I read it by myself and conceal it from my usual game-playing partners, I would win more often—but the game would be less fun, because winning would be too easy. I should not read it by myself because my inner agent's interest in winning *does not reach outside the game*. But if we all read the strategy guide, then the game would get more complex and fascinating and enjoyable. Then I have a good reason for all of us to read it—because my interest in collective fun is part of my enduring agency.

This can get quite complicated. When my friends are over, we play a board game for fun. My guiding interest is in, say, making sure we all have a good time. Suppose we are all striving players, and we want the fun of absorbed competition. I temporarily adopt an interest of winning—all my in-game actions are guided by my interest in winning and beating my friends. But at the same time, I take all sorts of out-of-game actions to be nice to my friends. Even while I am trying to totally destroy them in the game, I am also making sure that they have adequate tasty snacks and beverages, joking around with them, and generally doing what I can to sustain a warm and delightful social atmosphere. The gaming agency infects none of this.

What has emerged here is an interesting picture—and, to be clear, one, that I had not adequately articulated in my earlier discussions. The quarantine involved with game playing is interestingly complex and partially asymmetric. On the inbound direction, there is often a pragmatic firewall between my enduring reasons and my in-game reasons that helps me achieve certain goals by excluding from my attention certain enduring considerations. But those

enduring considerations are obviously still justificatorily active. And this pragmatic firewall is highly specific. In some cases, the firewall excludes only their awareness that their larger purpose is to have fun, in order to actually have the fun of absorption. In other cases, the firewall excludes many standard social considerations—like excluding various reasons of sympathy, say, for one's friends and spouse. But this exclusion is fascinatingly precise. I can have no sympathy for my friends' desperate struggles to escape from this in-game trap, even while I carefully attend to their culinary and physical needs. ("Do you need a pillow?" I asked my friend with back problems, then I brought her a selection of lumbar supports, even as I plotted the deadly move that would undercut her entire in-game economy.) So my in-game reasons are deeply quarantined, in the outbound direction, from my enduring self.

We can get a better handle on this curious structure if we approach it from another angle. For that, let us turn to Camp's second criticism.

3. NOT SO SEPARATE

Camp's second worry is that I am exaggerating the difference between the in-game agent and our full, enduring agency. In the Nguyen account, she says, we are fluid with our gaming agencies. But, says Camp, in the Nguyen account, our enduring agency is supposed to be very different: it is a stable and somewhat monolithic form of agency.

Camp asks us to consider a different account of our enduring agency. In the Campian account of agency, the enduring agent, too, is fluid and shifting. An agent, for Camp, is actually a repertoire of different practical modes.

> In place of the enduring, purposeful rational agent, we might embrace a model that construes agency and selfhood in terms of repertoires of interpretation and action, with beliefs and goals as especially stable functional nodes within these repertoires. The locus of agency resides as much in one's choices about which contexts to enter, and so which modes to cultivate, as in one's long-term reflectively endorsed commitments or active, moment-to-moment decisions. We achieve selfhood, not necessarily by subsuming our lives under stable teleological structures, but by integrating our repertoires of engagement into coherent characters: ones whose contextual variations hang together in higher-order, often highly complex, wholes.[12]

12 This precise text is from Camp's comments at the Author Meets Critics session on *Games* at the 2020 American Philosophical Association Eastern Division. She takes this to be a summary of her view in "Perspectives in Imaginative Engagement with Fiction."

Let us take onboard the Campian account of the enduring agent, which strikes me as very close to the truth. Our authentic agency is not some fixed and enduring singular set of values, aims, or commitments. Instead, we shift between repertoires of nodes, where each node includes a cluster of values, beliefs, and goals. I had meant to indicate something very much like this in my discussion of how we use different modes of agency in ordinary, nongame life. I have a particular mindset I use for political machinations in administrative meetings against hostile forces; another mindset I use to teach wary undergraduates forced into my ethics class; another mindset I use when mentoring graduate students; another mindset I use to comfort my wailing children; and another mindset I use when trying to write replies to frustratingly devious critics. One mindset is hyper-careful and fussy, another loves big ideas and broad strokes; another suspicious of possible veiled motivations; another grounded in empathy and love. My argument was that games helped to train up different psychological modes, so that we might better access these different modes in practical life. We are, I said, something like a Swiss Army knife of practical modes.

I suspect, however, that this image might have problematically implied a certain hierarchy: that these practical modes were temporary sub-agencies, chosen by some kind of master agent to fit the moment at hand. The image suggests that there is the Swiss Army knife with many modes—but also suggests that, behind it, there is some singular agent who deploys the knife. Camp is resisting this picture of the hierarchy, and the thought that, somewhere up the rational chain, there must be a stable, committed master agent. Instead, in the Campian account, the Swiss Army knife is all there is. There is no master mode to rule them all—only different modes subject to some very complicated coherence conditions.

Suppose we take on board the Campian account of the enduring agent as a fluid, multifaceted, and non-hierarchical collection of modes. Still, I do not think this collapses the difference between the enduring agent and the enduring agent. Gaming agencies are fluid in a distinctive way. But if both gaming self and enduring self are fluid, what distinguishes them? In biting the Campian fluidity bullet, I owe an account of what makes the enduring agent special and distinct from the gaming agencies.

Consider Millgram's account of what it is to be unified as a practical agent. A practical agent takes into account a variety of considerations, which thereby encompass what matters to the agent. You could almost think of a particular rational agency as a thing that is responsive to some specified set of considerations. What unifies a set of considerations into a singular agent is that, in a chain of practical reasoning, *any one consideration from that set might bear on another consideration from that set*. I have a number of modes: teaching mode,

parenting mode, research mode, cooking mode. When I am engaged in any one of these modes, my attention is usually narrowed to a certain set of considerations. I usually do not think about my students when I am trying to pickle some kimchi, and I usually do not think about my children's dietary needs when I am trying to write a philosophy article. Those narrowings are practical strategies for dealing with my cognitive limitations. They are the strategies of a finite being. I narrow my focus so as to exclude what is *unlikely* to matter, in order to save some cognitive resources in my desperate attempt to actually get something done. But this is just a labor-saving, defeasible, heuristic strategy. Such considerations *could* weigh against each other, and when it becomes apparent that they are relevant, it is completely straightforward for me to weigh considerations against each other. My child wants to go to his school fair, which is at the same time as an important conference I wanted to Zoom into. Now kid reasons are in play against research reasons. I am cooking and suddenly an idea for a paper hits me, and I have to decide whether to prioritize timing this omelet perfectly or writing down that idea. Now cooking reasons are suddenly in play against philosophy reasons. What makes me one unified agent, says Millgram, is precisely the fact that it makes perfect sense to weigh any of these reasons against the other. After all, it is *I* that is involved in parenting, philosophy, and cooking, and all of these things are important to me, so I have to weigh them against each other and decide.[13]

In "Games and the Art of Agency," I argued that my account of games shows the problem in Millgram's account of the unified agent. There are, I argued, aspects of my agency that are not subject to such a unity constraint: specifically, my in-game agencies. This way of putting things now strikes me as a bit crude, so let me offer a more refined version of the point.

Here is what I propose: my *enduring agency* is subject to such a unity constraint, but my temporary gaming agencies are not—at least, not in their most full-blooded form. There are particular ways in which my in-game reasons will not emerge from their context to weigh against, say, my kid-rearing reasons.

To recast this into Camp's terms: though I, as an agent, may be a thing that moves fluidly and non-hierarchically between various modes, insofar as these modes are part of my enduring agency, they are all subject to some kind of coherence conditions. This is not to say that I need to create some master agency, with some explicitly delimited set of values that could be used to deduce all the other values of the various temporary sub-agents. Rather, it is that I need to be

13 See also Carol Rovane's account, in which a particular agency is individuated as a deliberative point of view subject to a demand of rational unity—that is, that the set of considerations that belong to "an agent" are responsive to each other (Rovane, "What Is an Agent?" and "Group Agency and Individualism").

able to find a way of conceiving of my different considerations as at least coherently cohabiting. Something has gone wrong if, in my business life, I ruthlessly destroy people's lives and take away their homes, and then in my home and spiritual life I think of myself as a charitable, kind, empathetic person. To put it another way: even if our enduring agency consists of a number of different personality nodes, each of them is at least *answerable* to the others. I can reflect on my parenting life from my philosophical mindset and see if my parenting decisions hold up from a more philosophical perspective. And I can reflect on my philosophical life from the parenting perspective and wonder whether all those weird abstractions I seem to be committed to can possibly hold up from the perspective of my parenting life. And I can question particular choices and reasons that I have made in one mode from the perspective of another.[14]

I often do not connect perspectives like this, but sometimes I do—because the considerations from my various modes do bear on each other. My standard disconnections are, again, simply a pragmatic strategy to get around my cognitive limitations. When I am not on childcare duty these days, I am usually in my basement office trying to get some work done. During those times, my kids are usually upstairs; I can hear them running around and laughing and crying. Normally, I put that out of mind—not because they do not matter to me, but because I need to focus pretty intensely to actually get anything done. I have created the kind of life where I have certain periods of time (mostly) reserved for work, and certain periods (mostly) reserved for childcare. Having a period in which I devote myself to work is consonant with my goals as a parent. I have a pragmatic reason not to think about my kids: sometimes, I need to focus completely on my work because it takes every inch of my mind to get any forward progress in philosophy. So I set up a temporary firewall between the various nodes of my enduring agency to manage the cognitive overload. There is no deep logical antipathy between considerations from different nodes, only a

14 Notice that Camp worries precisely that she does not like playing *Monopoly* because she does not like to be the kind of person who is capable of entering the mental mindset of *Monopoly*. Notice that this complaint is about a developmental effect of playing *Monopoly*—which is developing a capacity to do so—and not about the particular reasons acted on during a playing of *Monopoly*. I think it is far less plausible to say that one does not like playing *Monopoly* because one takes unkind actions toward one's friends. Or, at least, the latter complaint invites the diagnosis that the complainant does not understand the nature of games, in a way that Camp's original complaint does not. Notice, too, a key difference between the cases. It is very hard to imagine how a person might really offer a coherent explanation of themselves that could accommodate being a destroyer of lives in their business life, but a charitable and kind person on weekends at church. But it is easy to offer a coherent explanation of how I might be a kind person and also play *Monopoly*: that explanation is that I play *Monopoly* only insofar as all the players are, by and large, mostly having fun.

temporary firewall, set up as a practical solution to the problem of limited attention. We can see this by noting how the various considerations from different domains will sometimes break through the pragmatic firewall and come into play with each other. Maybe, instead of the typical tantrum sounds, there is the cry of genuine hurt and pain. The firewall comes down in an instant; I spring upstairs. I can, in fact, easily weigh my kid reasons against my research reasons and come to a quick conclusion.

But, I want to suggest, the barrier between my gaming agency and my enduring agency is not merely pragmatic. Here, I suspect that Camp would take the opposite tack. She might say: notice how similar the parenting/research firewall is to the gaming/nongaming firewall. In a game, I concentrate on playing it—but if my opponent starts crying, I can snap out of playing the game and pay attention to their sorrow. I can concentrate on my research, but if a sufficiently panicked yowl comes from upstairs, I can snap out of it and run upstairs to see what the hell is going on with my kid, and if whoever is on childcare duty needs some help. But, to my mind, the parenting/research divide is different in kind from the game/nongame divide. There are two significant differences: one fussy, one broad.

Let us start with the fussy difference. There is *a particular motivational state of play* in which my gaming reasons and my enduring reasons cannot be brought into a single line of reasoning with each other. This state arises, for example, when my goal in playing a game is to have the enjoyments of total practical absorption in the attempt to win. And this is a very peculiar state. External considerations about *playing for the sake of having fun* cannot be brought to bear on the set of gaming reasons, because those considerations will undermine the gaming agency's ability to be absorbed in the attempt to win. So they must be excluded. This is still a pragmatic reason, but one different in kind. The parenting and research reasons can exist in the same line of reasoning, without undermining each other. It is simply convenient, most of the time, to break them up into their own little cubicles. But the enduring interest in that particular kind of fun and the particular interest in winning cannot be put into the same line of reasoning without undermining the quest for fun. Let me be clear: I do not think this kind of pragmatic exclusion is required for every kind of play. It is a feature of a very specific context, where the enduring agent's interests and reasons in playing the game are self-effacing—that is, where the enduring agent must put themselves out of contact with their larger purpose in order to achieve that purpose. In the parenting/research case, the firewall is merely for the sake of managing cognitive load.

Notice that the enduring agent's reasons can cancel their absorption in the temporary gaming agency—but this is a very different relationship from

directly weighing their enduring reasons against their inner gaming agent's reasons. At no point am I weighing my reason for gaining material advantage against my reasons for having fun. Rather, either I am devoted to the win and reasoning in order to win, or my enduring interest in fun has cancelled my interest in winning entirely. One might reply: Am I not weighing my interest in winning against whether or not we are having fun? But I am a striving player. It is not like I am weighing my real interest in having fun against my equally real interest in winning. I am adopting a temporary interest in winning entirely in service of my enduring end of having fun.

Here is the second, broader difference between the parent/child firewall and the gaming firewall. I am not subject to the same coherence conditions across those agencies. I can entirely understand the question of how you could possibly be the kind of person interested in writing philosophy about games, and raising a happy child. These interests are part of a coherent set of values and interests. But a much stranger question is how you can be the kind of person who carefully and lovingly makes your spouse's favorite dinner and then sets out to vanquish all their plans in *Res Arcana*. This question strikes us strange because it presumes that these two reasons have a similar status, such that it is meaningful to expect that they can be made directly coherent with one another. Of course, there is a way to make them coherent with one another, but it requires referencing the specific logic of games. That is, I can do it because I am only setting out to vanquish their plans in the specific context of a pleasing competitive game. In other words: the reasons I have in a game cannot be brought directly into a chain of reasoning with the reasons I have outside the game, except via an understanding that devoting oneself to the in-game reasons will instrumentally support my enduring reasons. Outside the context of the game, I no longer have any of the reasons to win.

So the coherence conditions here work in a very funky way. My in-game reasons are subject to systematic coherence with my enduring reasons, but my enduring reasons are not subject to systematic coherence with my in-game reasons. That is, whenever I am playing a game, I can subject my game-playing reasons to the demand for coherence with my enduring reasons. Why am I trying so hard to win? Because it was supposed to be fun. Is it not really so fun after all? Then perhaps we should stop trying to win and quit this game. But, once again, that relationship only applies to the inbound reasons. On the outbound side, my enduring reasons are not subject to the demand to be coherent with my in-game reasons. Suppose I often spend several hours trying to best my spouse at a board game and cut off her best attempts at victory. Those attempts can be subject to coherence with my enduring interests. I can say: "Why am I trying to disrupt my spouse's intricate economic plans?" And the answer makes

reference to my enduring interests: "Because it is fun for me and fun for her." But when I am trying to cook dinner for my spouse, I do not need to square that attempt with, say, my interest in disrupting her plans.

What is more: the various modes of my enduring agency are subject to coherence conditions with each other. But the various agencies I adopt for different games are not subject to coherence with one another. There is nothing strange about the fact that I want to collect lots of red tokens in one game, but avoid collecting any red tokens in another. In-game reasons do not cross from one gaming agency to another.

To put it in the language of the earlier tussle with Camp: the coherence demands are not a two-way street, where any reason from one perspective can be put into the mix with any other reason. For one thing, reasons for making in game-moves can be made coherent with my enduring reasons, but only via reference to the context of the game and the striving motivational structure. For another, my enduring reasons are under no constraint to be coherent with my in-game reasons. For yet another, my reasons from one of my gaming agencies are under no constraint to be coherent with any of my other gaming agencies. This is because my in-game reasons are *highly limited in scope to one specific context*. Quarantine, it turns out, is not the right analogy. In a quarantine, there should be no mingling at all: the quarantined should stay in, and everybody else should stay out. The structure here is different: enduring reasons have the logical reach to extend inward (with the proviso that we often want to forget about them), but the inner gaming reasons cannot get out. Perhaps the right analogy is: gaming reasons are in a sort of agential prison.

Let me stay on this point a little longer, because the picture that is emerging, prompted by Camp and Millgram, is more textured than what I have presented in the past. Here is the picture: my motivations for doing certain things only arise in the game context and do not arise outside the game context. This is true even if action outside the game proper would have results inside the game. For example, one of my typical strategies in games is to exhaust my opponent's cognitive resources by making moves that make my opponent's position more complicated while making my situation simpler. This is, I take it, a way of attacking their cognitive resources and driving them to exhaustion. (Interestingly, I think many players of board games instinctively attack their opponent's cognitive resources in this way, but this sort of strategy is rarely explicitly articulated. But such strategies are often explicitly articualted in sports like basketball, soccer, and especially any kind of martial art. In such physical games, a basic strategy is to take actions that are relatively energy-efficient for you but are energy-costly for your opponent to respond to—to make moves that give you the advantage in remaining stamina.)

Notice that we do not think that exhausting our opponent's resources is a good thing to do outside the game, even if the consequence is an in-game victory. That is, it would be strange for a striving player like me to try to ask my friend to do incredibly complex calculations before the game—like casually asking them to explain Hegel's ontology and pretending not to understand their answer, with the goal of exhausting their mental resources for the game itself. The goal of winning by exhausting my friend is local to the game, because the goal of winning over them is local to the game.[15] Notice that the distinction here is not between some artificial in-game agent and the real thing. It is not that I attack my friend's virtual in-game avatar but not their real self. When we play basketball, I am really trying to deplete their real energy reserves and they mine. What matters here is that the reasons we have to exhaust each other are only operative inside the gaming context.

It might seem silly to talk about these phenomena in such an elaborate way, because, zoomed out, the phenomena are so familiar. But that is, for me, the most important part. What I am talking about here sounds arcane when put in the language of philosophy, practical rationality, and agency, but it is a basic fact about game playing. When I play a game, I erect a structure of reasons and considerations. But the gaming structure of considerations only has pull inside the game, and I discard it for extra-game reasoning. Game reasons are highly temporary and highly confined reasons. This is why I think of the gaming agency as a sub-agency, layered within my enduring agency. My gaming reasons are always subject to coherence demands with my enduring reasons (via the logic of striving play), but my enduring reasons are not always subject to coherence demands with my gaming reasons.

To sum up: one's gaming agency is interestingly isolated from one's enduring agency, but that isolation is much more complex than a simple, brute quarantine. That isolation has a different structure depending on the direction of the demand for coherence. I exclude certain enduring considerations from my gaming consciousness for pragmatic reasons—like forgetting that I am climbing to relax, in order to actually relax from my absorption in the climb. The enduring reasons are still live for me, but they will interfere with my absorption if I am aware of them during the game. But the reason I do not try to exhaust my friend before we play the game is a matter of logical structure, and not some practical, psychological trick. It is because the reasons involved in games simply do not extend outside of the game.

15 Of course, I have heard tell of people who do things like this, like tournament poker players who try to psych out their opponents with out-of-game behavior. But, of course, this is easily explained by the fact that these players are achievement players and not striving players. They are enduringly interested in winning the game.

Let us return to the original objection. Camp worried that there was no major difference between the gaming agency and the enduring agency, because our enduring agency involves fluidly shifting between a wide set of perspectives, none of them dominant. I am happy to grant that picture of the fluidity of our enduring agency. But I want to add: there is a key difference between the fluidity of the enduring agent's modes and the fluidity of our gaming agencies. The enduring agency's many modes are subject to a thoroughgoing unity constraint. The gaming agencies are subject to a very different constraint—a one-way constraint. The gaming agencies need to make sense from the perspectives of the enduring agent, but the enduring agent's many modes need not make sense from the perspective of the gaming agency. The gaming agency is a disposable sub-mode, which is not subject to the same thoroughgoing demand for unity and coherence. The gaming agency is answerable to the justificatory perspectives of the enduring self, but the enduring self is not answerable to the justificatory perspective of the gaming agency.

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

REFERENCES

Baier, Annette. "Trust and Antitrust." *Ethics* 96, no. 2 (January 19986): 231–60.

Camp, Elisabeth. 2017. "Perspectives in Imaginative Engagement with Fiction." *Philosophical Perspectives* 31, no. 1 (December 2017): 73–102.

Kukla, Quill. "Sculpted Agency and the Messiness of the Landscape." *Analysis*. 81, no. 2 (April 2021): 296–306.

Millgram, Elijah. "On Being Bored out of Your Mind." *Proceedings of the Aristotelian Society* 104, no. 2 (June 2004): 163–84.

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

Nguyen, C. Thi. *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.

———. "The Opacity of Play: A Reply to Commentators." *Journal of the Philosophy of Sport* 48, no. 3 (2021): 448–75.

Nguyen, C. Thi, and Jose Zagal. "Good Violence, Bad Violence: The Ethics of Competition in Multiplayer Games." *DiGRA/FDG '16 – Proceedings of the First International Joint Conference of DiGRA and FDG* 1, no. 13 (August 2016).

http://www.digra.org/digital-library/publications/good-violence-bad-violence-the-ethics-of-competition-in-multiplayer-games.

Rovane, Carol. "Group Agency and Individualism." *Erkenntnis* 79, no. 9 (October 2014): 1663–84.

———. "What Is an Agent?" *Synthese* 140, no. 1–2 (May 2004): 181–98.

Scott, James. 1998. *Seeing Like a State*. New Haven: Yale University Press.