JOURNAL *of* ETHICS & SOCIAL PHILOSOPHY

VOLUME XXV · NUMBER 3
September 2023

ARTICLES

DISCUSSIONS

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. 25, NO. 3 · SEPTEMBER 2023
https://doi.org/10.26556/jesp.v25i3.2109

THE KIDS AREN'T ALRIGHT

EXPANDING THE ROLE OF THE STATE IN PARENTING

Connor K. Kianpour

> The kids are grown up but their lives are worn.
> —The Offspring

PROPONENTS of *private parenting* believe that "individuals should be able freely to decide whether or not they wish to have and raise children without public regulation" and that the costs of child-rearing are generally to be borne by particular parents without government support.[1] Still, proponents of private parenting are amenable to the state using its coercive power to relocate children who have been abused or neglected by their parents but only *after* they have been abused or neglected, and they are amenable to certain social programs that benefit children and those who rear them but not because the costs of child-rearing should be subsidized by the state. By contrast, proponents of *regulated parenting* believe that the state ought to play a comparatively larger role in regulating *who* may have and raise children, *how* those who have and raise children may do so, and/or the extent of the support parents receive from the state to help raise their children. One of my objectives in this essay is to argue that we should presume the desirability of regulated parenting policies in the absence of compelling reasons to favor private parenting.

There are, however, distinct views about the form regulated parenting should take. Proponents of regulated parenting might advocate for policies of *public parenting support*, *parental monitoring*, *parental licensing*, or some combination of the three. Daniel Engster, for example, defends the idea of public parenting support. For Engster, there are three features of parenting that mark it off from other activities: (1) many of the obligations correlative to parenting fall disproportionately on women, (2) parenting produces social goods (children) necessary for sustaining civil society, and (3) parenting is the mechanism through which the claims children make on others as emergent persons are realized. Because parenting involves considerations relevant to social justice,

1 Engster, "The Place of Parenting within a Liberal Theory of Justice," 234.

social stability, and the rights of children, the costs of parenting should be shared across individuals in liberal society.[2] Public parenting support would, thus, involve a range of social programs, among them being paid parental leave, public childcare, and public subsidies and tax benefits to parents.[3] These programs would allow parents to raise their own children while providing children with the resources necessary for safeguarding their interests.

Jurgen de Wispelaere and Daniel Weinstock, as well as some others, have alternatively defended a policy of parental monitoring.[4] The proposal in its general form requires social workers and healthcare professionals to visit households regularly and to evaluate how the interests of children are being protected and promoted. Perhaps these visits would be more frequent when the child is younger and would become less frequent as the child gets older. Presumably, these visits would incentivize parents who desire to rear their children to do so well enough to pass these evaluations, as well as create opportunities for professionals to intervene relatively quickly when abuse or neglect takes place in a household.

Finally, some have expressed support for a policy of parental licensing.[5] Parental licensing involves the state using evaluative tools to determine whether individuals are competent to be parents before raising children and excluding those who are judged to be incompetent from raising children.[6] The primary benefit of parental licensing as compared to private parenting is that licensing parents would, if efficacious, protect children from abuse and neglect *before it takes place*. Another objective of mine in this essay is to argue that parental licensing, out of the regulated parenting proposals that exist, is best suited to safeguarding the interests of children along one significant dimension. In particular, parental licensing, unlike public parenting support and parental monitoring, can insulate children from being raised by those who are *objectionably intolerant*, such as racists, sexists, and homophobes. The argument I make in this essay can be summarized as follows:

2 Engster, "The Place of Parenting within a Liberal Theory of Justice," 236–42, 254.

3 Engster, "The Place of Parenting within a Liberal Theory of Justice," 255–56.

4 See De Wispelaere and Weinstock, "Licensing Parents to Protect Our Children?"; Archard, "Child Abuse"; LaFollette, "Licensing Parents Revisited," 338; and Engster, "The Place of Parenting within a Liberal Theory of Justice," 257.

5 See LaFollette, "Licensing Parents" and "Licensing Parents Revisited"; Mangel, "Licensing Parents"; Taylor, "Children as Projects and Persons"; and Cohen, "The Harm Principle and Parental Licensing."

6 Some believe that people should be required to obtain licenses to permissibly procreate. See, for example, Taylor, "Children as Projects and Persons"; and McFall, *Licensing Parents*. This is not what I mean by parental licensing. Rather, I mean that people should obtain licenses in order to permissibly rear—*not* give birth to—a child.

P1. Regulated parenting is presumptively justified because child-centered accounts of child-rearing rights are true.
P2. If regulated parenting is presumptively justified because child-centered accounts of child-rearing rights are true, then one of the means by which parenting may be regulated involves ensuring that parents are sufficiently tolerant of people with different backgrounds and ways of life because this is what child-centered accounts of child-rearing rights recommend.
C1. One of the means by which parenting may be regulated involves ensuring that parents are sufficiently tolerant of people with different backgrounds and ways of life because this is what child-centered accounts of child-rearing rights recommend.
P3. If one of the means by which parenting may be regulated involves ensuring that parents are sufficiently tolerant of people with different backgrounds and ways of life because this is what child-centered accounts of child-rearing rights recommend, then we have special reason to institute a policy of parental licensing.
C2. We have special reason to institute a policy of parental licensing.

In section 1, I defend P1 by arguing, against the accepted wisdom in the philosophical literature on child welfare policy, that a special burden of justification falls on proponents of private rather than regulated parenting to justify their preferred position. In section 2, I defend P2 (and C1) by arguing that children have a right against being reared by parents who are objectionably intolerant, and that this suggests that regulated parenting policies may be directed at safeguarding this right. Then, in section 3, I defend P3 (and C2) by explaining how we have reasons to institute a scheme of parental licensing which are not likewise reasons to institute policies of public parenting support or parental monitoring. In particular, parental licensing offers the best solution to the problems that befall children who are victims of a distinctive, insidious form of bad child-rearing—child-rearing by those who are strongly homophobic, racist, sexist, and the like. I ultimately hope to persuade you that some form of regulated parenting is presumptively justified, and that it is harder than one might have initially thought to rule out the implementation of a policy of parental licensing in particular.

1. PRIVATE PARENTING VS. REGULATED PARENTING

Defenders of both private and regulated parenting assume that the burden is on proponents of regulated parenting to justify their position. Since it is assumed that private parenting is what the default position should be and that

regulated parenting requires special justification, it is easy for opponents of regulated parenting to claim that the reasons offered to defend these policies do not meet the special burden of justification they must meet. There are two arguments for why the burden of justification falls on proponents of regulated parenting to justify their view: *the cost argument* and *the risk argument*. The cost argument applies to all forms of regulated parenting policies, whereas the risk argument applies to parental licensing specifically. In this section, I show how the cost argument, when taken seriously, actually grounds a presumption against private parenting and how the risk argument cannot ground a theoretical presumption against parental licensing. To begin, I will lay out the cost and risk arguments in standard form:

The Cost Argument

P1. Regulated parenting policies are costly and interfere with people's pursuit of their preferred ends.

P2. Costly and liberty-constraining policies require a special justification.

C. Therefore, regulated parenting policies require a special justification.[7]

The Risk Argument

P1. If a licensing scheme risks jeopardizing the fundamental rights of a disproportionate number of individuals, a special justification is required for permissibly enforcing the scheme.

P2. Parental licensing schemes risk jeopardizing the fundamental rights of a disproportionate number of individuals to rear children.

C. A special justification is required for permissibly enforcing parental licensing schemes.[8]

In what follows, I make a series of arguments that support my responses to the cost and risk arguments. I will examine different accounts of *child-rearing rights*, or the rights parents have to control or exercise global authority over the lives of their children; offer objections against dual accounts of child-rearing rights; and argue that those accounts remaining—namely, child-centered accounts of child-rearing rights—converge on the conclusions that private parenting risks violating the rights of children to an unacceptably high degree and

7 In his defense of public parenting support, Daniel Engster offers arguments meant to show that public parenting support meets this burden of special justification. See Engster, "The Place of Parenting within a Liberal Theory of Justice." Christopher Freiman and Hugh LaFollette assume that a special burden of justification falls on proponents of parental licensing because it is costly to society and individuals. See Freiman, "Against Parental Licensing," 114; and LaFollette, "Licensing Parents Revisited," 328.

8 See De Wispelaere and Weinstock, "Licensing Parents to Protect Our Children?," 200–201. See also Sandmire and Wald, "Licensing Parents."

that regulated parenting, including parental licensing, does not risk violating the rights of individuals to rear children.[9] By showing that private parenting risks violating the rights of children, I position myself to respond to the cost argument by showing how it leads us to the conclusion that regulated parenting, not private parenting, is presumptively justified. And by showing that regulated parenting, including parental licensing, does not risk violating the rights of parents, I position myself to refute the risk argument by showing how parental-licensing schemes do not risk jeopardizing the fundamental rights of a disproportionate number of individuals to rear children.

1.1. Against a General Right to Rear Children, Biological or Otherwise

One might, as S. Matthew Liao does, argue that adults have a human right to rear their biological children because rearing one's biological child is an activity that enables many human beings, *qua* human beings, to lead good lives.[10] This is because, by rearing one's biological child, "one is (a) creating a new life, (b) a right holder; (c) with one's own genetic material which in part determines the genetic identity of this new individual; (d) and one has the opportunity to see and shape the growth of this new individual."[11] Doing this, for many humans, is integral to their leading good lives as humans.[12] Liao claims that for the right to rear one's biological child to be respected, one must have "the power to exclude others from trying to be the primary providers for [one's] biological

9 It is easy to think that "parental rights" encapsulate reproductive rights, since to become a parent one must generally reproduce. To avoid this confusion, I use the term "child-rearing rights" to signify that the rights in question are specifically rights correlative to *raising* a child, not *creating* a child.

10 Some, like Hillel Steiner and Jan Narveson, maintain that parents are sovereign over the children they create unless the parents relinquish rights of control they have over their children. Such an account is implausible, though, because it is widely accepted that children have at least some rights that can prevent a parent from treating them in some ways. See Steiner, *An Essay on Rights*, 248; Narveson, *The Libertarian Idea* and *Respecting Persons in Theory and Practice*. For a more sustained critique of this kind of position, see Okin, *Justice, Gender, and the Family*, 79–88. Given the problems with the account just mentioned, I take Liao's account of a right to raise one's biological child to be the strongest account of such a right.

11 Liao, "Biological Parenting as a Human Right," 658.

12 Liao assumes that "human rights are grounded in . . . the fundamental conditions for pursuing a good life, where a good life is one spent in pursuing certain valuable, basic activities" and that "basic activities are ones that if a human life does not involve the pursuit of any of them, then that life could not be a good life. In other words, a human being can have a good life by pursuing just some, and not all, of the basic activities" ("Biological Parenting as a Human Rights," 654). Thus, Liao is not committed to the claim that those who do not rear biological children fail to lead good lives, since their lives could be spent pursuing some other basic activity or activities.

child."[13] Without this power, parents would presumably be insecure in their ability to see and shape the growth of their children, which is, according to Liao, part of the interest parents have in rearing their biological children in the first place. For Liao, the human right to rear one's biological child is defeasible, as all human rights are. The right to rear one's biological child may be permissibly restricted, for instance, if one abuses or neglects one's child. I argue, however, that Liao's arguments, if successful, support the conclusion that one is entitled to *some kind* of protected relationship with one's biological child rather than the conclusion that one is entitled specifically to a protected relationship with one's biological child *in which one has global authority over that child's life*.

Liao treats child-rearing rights as though they are in part derivative of parents' interests in seeing and shaping the growth of a rights holder whom they created using their own genetic material. The right that biological parents have to see and shape the growth of a rights holder whom they created using their own genetic material can be protected, however, without also treating them as if they have the right to exercise global authority over that child's life. Consider the following case. Maria gives up her newborn baby boy for adoption and, in so doing, relinquishes her (presumed) right to exercise global authority over that newborn's life. This means Maria cannot, for example, determine where the newborn lives, what his bedtime is, what he regularly eats, and the like. Suppose now that Maria thinks it is important for him to have a relationship with his biological mother, and the child's adoptive parents are kind enough to allow Maria to spend time with him every week. Obviously, Maria would not be able to see and shape the growth of her newborn to the extent that the newborn's adoptive parents can and may, but Maria would nevertheless be able to see and shape the growth of her newborn to a significant extent under these conditions. For a biological parent to see and shape the growth of her child, she need not (and should not) be the only person doing so, nor need she be the person in the child's life who does these things the most. She only needs to see and shape the growth of her biological child to some extent that is meaningfully significant, and what this means is likely subject to change depending on specific features of particular parents and their biological children. For example, the parent whose biological child thrives because the child meets with the parent on a weekly basis would have a stronger claim to a protected relationship with that child than the parent whose biological child finds weekly visits with her parent emotionally distressing. Thus, the conclusion Liao's arguments support is that parents have a right to protected relationships with their biological children, but this conclusion does

13 Liao, "Biological Parenting as a Human Right," 660.

not entail that these protected relationships are ones in which the parents are entitled to exercise global authority over the lives of their children.[14]

I want to clarify what I mean when I say that parents have a right to protected relationships with their biological children. It is important that I do this because I subscribe to a child-centered account of child-rearing rights, as will be made clear by the end of this section. Child-centered accounts of child-rearing rights hold that parents have rights to raise their children because they protect their children's interests in the right sort of way. Contrast these accounts of child-rearing rights with *dual accounts*, which hold that parents have rights to raise children both because they—the parents—have weighty interests in raising children and because they can adequately protect their children's interests.[15] Those who subscribe to a child-centered account of child-rearing rights might believe that protecting a child's interests in the right sort of way means parenting a child in a way that is suitably in the child's *best* interest. Call this *the best available parent account* of child-rearing rights. Though I am not aware of anyone who has defended this alternative position, I could also imagine someone arguing that protecting a child's interests in the right sort of way means parenting a child in a way that protects the child's interests *to a sufficient degree*.[16] Such an account, I think, would qualify as a child-centered account if it was predicated on the view that, as a matter of justice, children are owed no more than an upbringing that is sufficiently in their interest without justifying this view on the grounds that the interests of parents ever count for more than the interests of children when determining who should rear a particular child.

14 My arguments against Liao would also work, with some modification, against an account of child-rearing rights that claims the interests that gestators have in rearing the children they gestate should figure into whether they should rear these children. See Gheaus, "A Right to Parent One's Biological Baby." The interests that gestators have with respect to the children they gestate can only ground presumptive child-*relating* rights, not presumptive child-rearing rights.

15 See, for example, Brighouse and Swift, "Parents' Rights and the Value of the Family"; and Liao, "Biological Parenting as a Human Right."

16 Anca Gheaus argues that those like Liam Shields who are proponents of sufficientarian accounts of child-rearing rights ought to subscribe to *child-centered* sufficientarian accounts of child-rearing rights. Shields himself, however, subscribes to a dual sufficientarian account of child-rearing rights, and Gheaus subscribes to a best available parent account of child-rearing rights that is child-centered. Neither defends that a child centered sufficientarian account of child-rearing rights is true, though. That it is possible for the most plausible version of a sufficientarian account of child-rearing rights to be child-centered, I think, provides us with reason to think that the merits of such an account should be explored in a more sustained manner in the philosophical literature. See Gheaus, "Sufficientarian Parenting Must Be Child-Centered"; Shields, *Just Enough*, 121–62, and "How Bad Can a Good Parent Be?"

Call this *the sufficientarian account* of child-rearing rights.[17] In this essay, I do not take a stand on which of these formulations of child-centered accounts of child-rearing rights is true.

One might worry, since I subscribe to a child-centered account of child-rearing rights, that by claiming that parents have a defeasible right to protected relationships with their biological children, I am forswearing my professed commitment to recognizing the primacy of children's interests in adjudicating matters concerning them. Thus, it is important that I am clear about what it means for parents to have a defeasible right to protected relationships with their biological children. There are two ways to conceive of defeasible parental rights to protected relationships that are consistent with child-centered accounts of child-rearing rights. First, it seems perfectly consistent to claim that both child-centered accounts of *child-rearing* rights and dual accounts of *child-relating* rights are true. One might hold that to rear a child, a parent must rear the child such that she protects her child's interests in the right sort of way because exercising global authority over a child's life imposes significant costs on the child that must be justified by appeal to the child's interests alone. Simultaneously, one might hold that to relate to a child, a parent's interests may figure into whether the relationship is worthy of protection because a relationship in which a parent lacks license to exercise global authority over a child's life does not impose significant enough costs to require justifying the relationship by appeal to the child's interests alone.

Second, even if one subscribes to child-centered accounts of *both* child-rearing and child-relating rights, it is possible to claim that parents have rights to protected relationships with their biological children on grounds consistent with both Liao's arguments and child-centered accounts of child-relating rights. What one might mean when they say that parents have a defeasible right to protected relationships with their biological children is this: a parent, if she has a defeasible right to a protected relationship with her biological child, is at liberty to disregard the limits that the child's adoptive parents attempt to impose on the way the biological parent may relate to her child when these limits prevent the biological parent from relating to her child in a way that is,

17 One might worry that such an account would license, in some cases, changing child custody so that a child is reared by worse parents so long as those parents are sufficient because, after all, children are owed no more than sufficiently good parents. But I can imagine a proponent of such an account arguing that having sufficiently good parents is different from but related to having a sufficiently good upbringing (which is what children are entitled to), and very few upbringings in which a child is relocated to a worse living situation qualify as a sufficiently good upbringing. And those who say this, I imagine, would largely evade the force of such a worry.

depending on which formulation of child-centered accounts of child-rearing rights is true, either in the child's *best* interest or *sufficiently* in the child's interest. By contrast, those who are not situated in protected relationships with children would *not* be at liberty to disregard, in the relevant contexts, the limits that a child's parents impose on the way that child may be related to. And we may presume, in the absence of reason to believe otherwise, that parents have rights to protected relationships with their biological children because the interests Liao claims parents have in these relationships give us some reason to believe that they, more so than others, will relate to their biological children in ways that are conducive to the interests of these children being protected in the right sort of way. We may use these interests to presume that one has child-relating rights without also being forced to presume that one has child-rearing rights because, as I noted before, presuming child-relating rights imposes significantly less costs on children than presuming child-rearing rights. These rights to protected relationships are defeasible because depending on which formulation of child-centered accounts of child-relating rights is true, a biological parent relinquishes her right to relate to her child when she fails to relate to her child in ways that are either in the child's best interest or sufficiently in the child's interest.

So far, I have argued that individuals lack a right to rear their biological children (though they may have a right to some kind of protected relationship with their biological children). Now, I turn to an account of child-rearing rights that takes as its foundation the interests, unrelated to biological relatedness, that adults have in being party to a parent-child relationship.[18] Harry Brighouse and Adam Swift argue that the parent-child relationship "cannot be substituted by other forms of relationship."[19] This is because the parent-child relationship

18 Ferdinand Schoeman has also defended an account of child-rearing rights that highlights the importance of respecting the interest that parents have in parenting. See Schoeman, "Rights of Children, Rights of Parents, and the Moral Basis of the Family." His view, however, has been criticized on the grounds that it fails to take seriously enough the interests of children and that it fails to fully explain what makes the goods of the parent-child relationship distinctive from the goods of other relationships. See Hannan and Vernon, "Parental Rights"; and Brighouse and Swift, "Parents' Rights and the Value of the Family." Loren Lomasky has also defended an account of child-rearing rights that emphasizes the interests parents have in staking "a claim to long-term significance through having and raising a child." Lomasky, *Persons, Rights, and the Moral Community*, 167. Again, there are ways in which Lomasky's account fails to give the interests of children their due consideration, particularly when he mentions how the state should permit parents to determine how (and whether) their children are educated. See Lomasky, *Persons, Rights, and the Moral Community*, 174–75. This is why I take Brighouse and Swift's account to be representative of the strongest version of the view that the interests of parents should factor into whether or not they may permissibly rear a child.

19 Brighouse and Swift, *Family Values*, 86.

generates an asymmetrical, intimate relationship between parent and child in which the child is especially vulnerable to the parent because the child relies on the parent for the protection of the child's interests and because the child cannot exit the relationship. Moreover, the love that a child feels for her parents is spontaneous, unconditional, and outside of the child's rational control, particularly in the early years of childhood, and parents take great satisfaction in being party to such a love. And finally, parents have a nonfiduciary interest in occupying the fiduciary role as a child's guardian, given the virtues and capacities occupying such a role helps parents to develop.[20] Taken together, these distinguishing features of the parent-child relationship generate for individuals a conditional, limited right to rear children.[21] This right is conditional and limited in the sense that it tracks the fiduciary responsibilities that parents have to their children.[22] For Brighouse and Swift, it is enough that an adult is capable of minimally meeting the needs of a child to enjoy the right to rear children. Parents need not be perfect parents to enjoy the right to rear a child since the substantial interest they have in experiencing a parent-child relationship weighs against the substantial interest that a child has in experiencing the same.[23]

I will now present an argument concerning Brighouse and Swift's account of child-rearing rights that is similar to the argument I presented concerning Liao's account. As was the case with Liao's arguments, Brighouse and Swift's arguments in defense of a fundamental right to rear children support the conclusion that individuals are entitled to *some kind* of protected relationship with children—*not* a protected relationship with children *in which they have the authority to rear those children*. Indeed, many who do not rear children are party to the spontaneous, unconditional, and arational love of children who are vulnerable to them: the extended family members of particular children, the caretakers of particular children, etc. And many who do not rear children are able to develop the virtues and capacities associated with doing so by establishing and continuing long-term, caring relationships with particular children. As Anca Gheaus writes,

20 Brighouse and Swift, "Parents' Rights and the Value of the Family," 92–96.

21 Brighouse and Swift, "Parents' Rights and the Value of the Family," 96, and *Family Values*, 86.

22 Brighouse and Swift write that "what children need, above all, is a spontaneous, intimate relationship with an adult who loves them, one who acknowledges the intrinsic goods of childhood while caring about their well-being, and respecting their individuality, enough to give them the huge amounts of attention, and the loving discipline, that are required for them to develop into the adults they are capable of becoming." Brighouse and Swift, *Family Values*, 85.

23 Brighouse and Swift, *Family Values*, 94–95.

> Adults' weighty interest in rearing children can be satisfied by establishing beneficial intimate and caring, although not globally authoritative, relationships with children, relationships which are protected from outside interference. Such relationships can satisfy adults' interest in self-knowledge and self-development: maintaining a long-term intimate and caring relationship with a child comes with great responsibility for how the child's life goes, and for the child's development. Not just parents but all parental figures exert great influence of this kind over children; by dint of being in an intimate relationship with an adult, a child becomes particularly vulnerable to that adult in material and emotional ways when strong attachments are formed. Protected relationships with children are also likely to display the experiential value of the parent-child relationship, since children can love and trust other adults with whom they stand in caring relationships. Most of the interest that Brighouse and Swift ascribe to adults can be satisfied by long-term and secure association with children, although, perhaps, not to the same extent that it is satisfied within the parent-child relationship.[24]

Again, an argument that purports to show that the interests of parents are integral to the child-rearing rights of these parents in fact shows that the interests of individuals are integral to rights to protected relationships with children—relationships in which these individuals do not necessarily possess the global authority to control children's lives. And as was the case with the rights to protected relationships that Liao's arguments lend credence to, the rights to protected relationships with children that Brighouse and Swift's arguments lend credence to can be understood in terms that are consistent with both dual and child-centered accounts of child-relating rights. If it is possible, as I suggested before that it might be, for child-centered accounts of child-rearing rights to be endorsed in tandem with dual accounts of child-relating rights, then it would be possible to recognize that people have child-relating rights consistent with Brighouse and Swift's arguments without conceding that child-centered accounts of child-rearing rights are false. And if one subscribes to a child-centered account of child-relating rights, one might use Brighouse and Swift's arguments to ground the rights that individuals have to protected relationships with children in the following way: Liao's arguments show us that there is some set of interests an individual has antecedent to relating to her biological child that gives us reason to believe that the individual's relationship with that child fits with the child's interests in the right sort of way. By contrast, Brighouse and Swift's arguments show us that there is a set of interests an individual has once

24 Gheaus, "The Best Available Parent," 455.

they have begun relating to a child that gives us reason to believe that the individual's relating to that child fits with the child's interests in the right sort of way. Because the individual in question relates to a child such that she has the strong interests in continuing to relate to that child that Brighouse, Swift, and Gheaus identify, we have some reason to believe that she, more so than others, will relate to the child in question in ways that are conducive to the interests of the child being protected in the right sort of way.

At this point, I have concluded that individuals lack rights to rear their biological children (even if they have rights to protected relationships with their biological children) and that individuals lack rights to rear children (even if they have rights to protected relationships with particular children). All that is left to ground the child-rearing rights of those who have the right to rear children is the interests of the children they rear. As noted before, these accounts of child-rearing rights are called child-centered accounts. Peter Vallentyne and Anca Gheaus defend versions of the best available parent account of child-rearing rights that I described earlier.[25] And again, I could also imagine someone arguing, *contra* Vallentyne and Gheaus, that children have no more than a moral right to be reared by the parent for whom custodial authority over a child is *sufficiently* in the child's interest. In either case, what explains the right that individuals have to control the lives of their children is that their doing so fits with the interests of their children in the right sort of way. These child-centered accounts of child-rearing rights accord with most peoples' intuitions about child-rearing rights. Most people believe, for example, that parents who routinely abuse and neglect their children relinquish their right to exercise global authority over the lives of their children precisely because these parents fail to stand in the right relation to their children's interests by abusing and neglecting their children.

Proponents of dual accounts of child-rearing rights might, nevertheless, resist the conclusion that child-centered accounts of child-rearing rights are true. Denying that the interests of those who rear children figure into determining who should rear children, as proponents of child-centered accounts do, might force us to accept potentially unsavory conclusions. Proponents of the best available parent account of child-rearing rights, for example, have been criticized on the following grounds. The best available parent account licenses changing child custody when a child's current parents provide a minimally good upbringing if doing so would be best in terms of the child's interests. It also licenses "reshuffling custody of babies at birth so that children are reared by those other than birth parents who are at least minimally good . . . if doing

25 See Vallentyne, "The Rights and Duties of Child-rearing"; and Gheaus, "The Best Available Parent."

so would be best in terms of the child's interests or would enhance equality."[26] Since many people find these implications of subscribing to the best available parent account implausible, they might be reluctant to concede that child-centered accounts of child-rearing rights are true. But there are a few things to say in response to this concern.

First, I argued above that the arguments offered in defense of dual accounts of child-rearing rights—such as those offered by Liao, Brighouse, and Swift—do not support the conclusion that the interests of those who rear children figure into determining who rears children. This gives us reason to believe that we should not think the conclusions arrived at by proponents of the best available parent account are as implausible as they at first appear. Presumably, these conclusions appear as implausible as they do because it is assumed that parents have interests strong enough to generate rights to rear children, but parents do not have interests strong enough to generate such rights so we should not think these conclusions are so implausible. Second, while the best available parent account licenses changing child custody and reshuffling custody of babies at birth if doing so is in the best interest of children, it seldom recommends that people should, in fact, change child custody and reshuffle custody of babies at birth. Children have strong interests in continuity of care, which often recommend that they continue being reared by those parents who have been adequately rearing them.[27] Moreover, there are many practical considerations that make it infeasible to change child custody such that it is in the best interest of the children involved and to reshuffle custody of babies at birth such that doing so is in the best interest of the children involved. And third, the criticisms of the best available parent account of child-rearing rights do not apply, at least with the same force, to a sufficientarian account of child-rearing rights. If children are owed, as a matter of justice, no more than an upbringing that is *sufficiently* in their interest, then changing child custody and reshuffling custody of babies at birth would be licensed if doing so helped ensure that children are given sufficiently good upbringings. These conclusions are far less implausible than those rendered in the case of the best available parent account and might actually strike many as quite intuitive. This might provide us with reason to seriously consider and further explore the possibility of a sufficientarian, child-centered account of child-rearing rights.

1.2. In Defense of Regulated Parenting

Now, I can respond to the cost and risk arguments I laid out at the beginning of this section. My responses to these arguments depend on two claims. The first

26 Shields, "Parental Rights and the Importance of Being Parents," 121.

27 Gheaus, "Sufficientarian Parenting Must Be Child-Centered," 192.

is that private parenting risks violating the rights of children. This is entailed by the truth of child-centered accounts of child-rearing rights. If child-centered accounts are true, then leaving socioeconomically disadvantaged parents without the resources to protect their children's interests in the right ways risks violating the right that children have to be reared by parents who can protect their interests in the right ways. While proponents of public parenting support, a form of regulated parenting, are able to offer a direct way to respect the rights of children born to socioeconomically disadvantaged parents, proponents of private parenting are not because they are resistant to state subsidization of the costs of child-rearing for parents. Even if proponents of private parenting support policies aimed at improving the material conditions of parents as adult citizens and in so doing likewise improve the material conditions of their children, the protection of children's rights in this state of affairs would be merely incidental, rather than the policy's aim. And insofar as the state should aim at protecting the rights of individuals, particularly those like children who are distinctively vulnerable and at the mercy of others, we should be critical of private parenting. Moreover, if child-centered accounts of child-rearing rights are true, then allowing children to be reared by unfit parents would violate the right that children have to be reared by parents who protect their interests in the right ways. Proponents of private parenting, unlike proponents of parental monitoring or parental licensing, are unable to offer a direct way to protect the rights of children in this regard. Again, this suggests that private parenting unacceptably risks violating the rights of children.

The second claim my responses to the cost and risk arguments rely on is that regulated parenting, including parental licensing, does not risk violating the rights that individuals have to rear children.[28] This is entailed by recognizing that, as I have argued, individuals have no right to rear their biological children, nor do they have any interests weighty enough to justify a right to rear children

28 Someone like Margaret Somerville might argue that it is *children* who have rights to be raised by their biological parents, and a parental licensing scheme would risk violating *these* rights. Somerville argues that we cannot assume that children would consent to being adopted by another family if they were able to consent given the testimony of those adopted children who feel "a profound sense of loss of genetic identity and connection" upon finding out that they are adopted. See Somerville, "Children's Human Rights to Natural Biological Origins and Family Structure," 42. Kimberly Leighton points out, however, that this sense of loss is likely to be less distressing, if not nonexistent, if we did not privilege the importance of biological relatedness in the family as we do. See Leighton, "Addressing the Harms of Not Knowing One's Heredity." So rather than privileging child-rearing within the biological family, we should challenge our preconceptions of what families are and should be at the sociopolitical level to protect children from feelings of loss of genetic identity and connection.

generally. Since these rights do not exist, regulated parenting policies cannot be said to jeopardize them. Even under a parental licensing scheme in which some who would, in fact, make good parents to a child are denied the opportunity to rear that child because they did not pass the licensing test, the prospective parents adversely affected by the scheme would not have their rights to rear children violated because no such rights exist. So long as these prospective parents are permitted to maintain long-term caring relationships with particular children (in which they do not exercise global authority over these children's lives), these prospective parents would have their rights to protected relationships with children respected, which are the only rights they can plausibly be claimed to have. And if one were to insist that individuals are entitled to protected relationships with their biological children specifically, then we may instate a policy that grants individuals rights to regularly visit with their biological children in tandem with a parental licensing scheme. Of course, these rights would only be enforced on the condition that doing so is consistent with the interests of the children when the parents in question are deemed unfit to raise their biological children.

With these two claims in mind, I now offer my responses to the cost and risk arguments. If private parenting unacceptably risks violating the rights of children and regulated parenting does not risk violating the rights of individuals to rear children, then it is private parenting and not regulated parenting which is costly and liberty constraining in the ways that are relevant to claiming that a presumption exists in favor of one of these views. Private parenting risking the rights of children imposes significant costs and constraints on children's liberties. Regulated parenting, by contrast, is aimed at mitigating these costs and constraints. The costs and constraints on liberty that regulated parenting imposes on people are those that are justified because rights are protected by imposing these costs and constraints. Moreover, liberties constrained by regulated parenting policies are not liberties that individuals are *entitled* to. Public parenting support would limit how much of one's income one could keep for oneself, but we routinely recognize that people are under an obligation to forgo this liberty when doing so helps protect the rights of others. Parental monitoring would limit the liberty of parents to exclude others from associating with their child, but parents are not *entitled* to this liberty when another's association with their child is important for the child (more on this in the following section). And parental licensing would limit the liberty of individuals to rear children (biological or otherwise), but, as I have argued, there is no fundamental right to rear a child. Thus, my response to the cost argument is that taking the principles motivating the argument seriously requires that the special burden of justification fall on proponents of private parenting—not proponents of regulated parenting—to defend their preferred position.

To refresh, the risk argument is an argument specifically about parental licensing. It claims that because parental licensing risks jeopardizing individuals' fundamental rights, we should think a special burden of justification falls on proponents of parental licensing to defend their preferred position. Jurgen de Wispelaere and Daniel Weinstock call attention to the fact that parental licensing schemes would inevitably produce many false positives.[29] That is, those enforcing the scheme will sometimes prohibit people who would be perfectly fine parents from rearing children simply because no licensing scheme is accurate all of the time. De Wispelaere and Weinstock take it for granted that individuals have a fundamental right to rear children, so they interpret the existence of these false positives as evidence that parental licensing risks jeopardizing individuals' fundamental rights. However, as I have argued, individuals have no fundamental right to rear children. And if they have no such right, then it cannot be claimed that such a right is in jeopardy when a parental licensing scheme is instituted. Thus, no special justification needs to be offered to permissibly enforce a parental licensing scheme.

Taking all of this in stride, I submit that proponents of regulated parenting need not offer some special justification to defend their preferred policies against a presumption in favor of private parenting. Indeed, if anything, proponents of private parenting must offer some special justification to defend their preferred policies against the presumption in favor of regulated parenting. Regulated parenting, in other words, is presumptively justified, whereas private parenting is not. At this point, you might wonder which regulated parenting policy proposals are best to adopt. For the remainder of this essay, I will argue that we have reasons to institute a scheme of parental licensing which are not likewise reasons we have to institute policies of public parenting support or parental monitoring. To do this, I must explain why a certain kind of person is unfit to rear children. This is now what I turn to.

2. THE PROBLEM OF OBJECTIONABLY INTOLERANT PARENTS

Regulated parenting is presumptively justified in my view because it, unlike private parenting, aims at helping parents treat their children in ways consistent with child-centered accounts of child-rearing rights. Child-centered accounts of child-rearing rights tell us that those who are entitled to exercise global authority over a particular child's life are those who are able to protect their child's interests in the right sort of way. In order to know, then, who is entitled to exercise global authority over a particular child's life, we must know if they

29 De Wispelaere and Weinstock, "Licensing Parents to Protect Our Children?" 200–201.

are capable of protecting their child's interests in the right sort of way. The aim of this section is to show that individuals who are *objectionably intolerant*—that is, they subscribe to prejudicial dogmas such as racism, sexism, and homophobia to such an extent that their ability to direct caring attitudes toward, for example, Black people, women, and/or gay people is significantly impaired—are unable to protect their children's interests in the right sort of way. I am concerned with the child-rearing rights of *objectionably* intolerant individuals because the forthcoming arguments draw issue with racists, sexists, and homophobes being *significantly* impaired in their ability to direct caring attitudes toward Black people, women, and gay people, respectively. It seems conceivable, especially if the child-centered account of child-rearing rights we subscribe to is sufficientarian in character, that those whose ability to direct caring attitudes toward members of the aforementioned groups is only slightly impaired (weakly intolerant individuals) or moderately impaired (moderately intolerant individuals) would not threaten the interests of children so much that the rights of children would be violated, whereas those whose ability to direct caring attitudes toward members of these groups was *significantly* impaired *would* so threaten the interests of children. I will return to this point later on, after having presented the arguments against individuals who are objectionably intolerant having a right to rear children.

First, I lay out Samantha Brennan and Colin Macleod's argument about how "strongly homophobic" individuals, specifically, are unfit to rear children. Then, I spell out a problem that Riccardo Spotorno identifies with Brennan and Macleod's argument and argue that this problem is only apparent, not actual. Even if it was actual, the solution Spotorno proposes is not the only available solution. After providing an overview of Spotorno's solution, I offer an alternative. The following discussion will produce three arguments, each of which may be consistently endorsed with the others, in defense of the claim that certain individuals are unfit to rear children because they are objectionably intolerant of certain backgrounds and ways of life.

Samantha Brennan and Colin Macleod offer a *precautionary argument* in defense of the claim that what they call "strongly homophobic" individuals are unfit to rear children. Brennan and Macleod argue that (1) children have an interest-based right to being provided with affective care, (2) those who rear children have a corresponding duty to provide their children with affective care, (3) strongly homophobic individuals cannot provide gay children with affective care, (4) because there is a nontrivial chance that the child of a strongly homophobic individual could be gay, strongly homophobic individuals are unreliable providers of affective care to children, and (5) 1–4 entail that strongly homophobic individuals are unfit to rear children.

According to Brennan and Macleod, affective care "involves manifesting love, affection, and emotional support to children; being attentive to their emotions, concerns, and enthusiasms; and being moved and concerned by threats to their well-being in ways that are transparent to children themselves."[30] Brennan and Macleod offer three reasons to think that children have an interest-based right to affective care from those who rear them. First, affective care promotes the welfare of children; children who are not loved by their parents fare worse than those who are. Second, affective care is one of the social bases of self-respect, meaning that it is in significant part through being loved by our parents that we see ourselves as valuable and meriting respect. And third, affective care facilitates intrinsic goods of childhood, such as innocence, trust, and intimacy. If children are denied affective care, then they will lose out on many intrinsically valuable goods. Taken together, these interests are arguably weighty enough to generate a right on the part of children to the affective care of their parents and a duty on the part of parents to provide their children with affective care.[31]

Strong homophobia, on Brennan and Macleod's understanding, "consists in belief in the moral wickedness or depravity of gay sexuality and identity" which "gives rise to attitudes of contempt, disgust, disrespect toward gay people."[32] The reason that strong homophobes—henceforth, homophobes—are unfit to rear children is that they would be unlikely to provide affective care to gay children, and there is a nontrivial chance that a homophobe's child could be gay. If homophobes are contemptuous of gay people, they are not in a position to manifest love to their gay children or to be moved by the distinctive threats to well-being that gay children face. Indeed, many gay children with homophobic parents do not complete high school, end up homeless, develop substance abuse problems, and take their own lives precisely because their homophobic parents are inadequate affective caregivers.[33] And if there is a nontrivial chance that any child could be gay, homophobes run the risk of violating the right their children have to affective care since any of their children could be gay. But by the time a homophobe learns that a child of theirs is gay, the child will have already developed significant attachments to them, and it will be exceedingly difficult for other adults who are not homophobes to step in and situate themselves in a loving relationship with the child, which would make the homophobic parent's withdrawal of affective care particularly troublesome for her child.

30 Brennan and Macleod, "Fundamentally Incompetent," 236.

31 Brennan and Macleod, "Fundamentally Incompetent," 236–37.

32 Brennan and Macleod, "Fundamentally Incompetent," 237.

33 Brennan and Macleod, "Fundamentally Incompetent," 238.

Thus, by exposing their children (who could grow up to be gay) to the risk of having affective care withdrawn from them at a crucial stage in their development, homophobic parents threaten the rights their children have to affective care and are therefore unfit to rear children.

Riccardo Spotorno argues that Brennan and Macleod's position renders an incomplete conclusion. Homophobes and racists alike commit a moral wrong by regarding others as morally inferior to them because they possess certain arbitrary characteristics, so we should expect that homophobes and racists face comparable consequences in terms of their claims and liberties for committing a comparable moral wrong.[34] While Brennan and Macleod's argument "rules out a moral right for homophobes to parent because there is always a nontrivial probability that their children will be gay, it fails to rule out a moral right for racists to parent because they can ensure that they have white children and it is virtually impossible that white children will become black."[35] So Spotorno takes it upon himself to construct an alternative account of why homophobes are unfit to rear children, which likewise renders the conclusion that racists are unfit to rear children.

I do not think the "problem" Spotorno identifies with Brennan and Macleod's argument is even a problem at all. We can see this by considering the following: suppose that Adam and Eve, two homophobic adults, want to adopt a 15-year-old boy named Straight. Straight is unequivocally, unquestionably a heterosexual. Brennan and Macleod's position would not rule out a moral right for Adam and Eve to be Straight's parents, even though Adam and Eve are homophobic, because there is no chance that Straight could turn out to be gay. This suggests that, on Brennan and Macleod's view, homophobes and racists do, in fact, face comparable consequences in terms of their claims and liberties for committing a comparable moral wrong. The wrong Brennan and Macleod are zeroing in on is not the wrong that homophobes and racists commit simply because they are homophobes and racists, but the wrong that homophobes and racists commit because they fail to direct affective care to their children in virtue of being homophobes and racists. And for *this* wrong, racists and homophobes face comparable consequences. Indeed, if there were a nontrivial chance that white children could grow up to be Black, Brennan and Macleod's position would indict racist child-rearing for the same reasons

34 Just as I use the term "homophobe" to designate "strong homophobe," my use of the term "racist" should be interpreted as designating "*strong* racist," or a racist who believes in the moral wickedness or depravity of members of a certain racial group that gives rise to attitudes of contempt, disgust, or disrespect toward members of that racial group.

35 Spotorno, "Homophobes, Racists, and the Child's Right to Be Loved Unconditionally," 7.

it indicts homophobic child-rearing.[36] Thus, the problem Spotorno identifies with Brennan and Macleod's position is merely apparent.

Nevertheless, I will lay out the alternative position Spotorno offers as a remedy to this apparent problem. Spotorno claims that children have a right to be loved unconditionally. That is, children have a right that the affective care directed to them by those who rear them is not conditioned on their possessing certain characteristics. The racist parent who provides an abundance of affective care to her white child fails to love her child unconditionally, since if the child were Black the parent would not provide that same affective care. Spotorno grounds the right to be loved unconditionally in the value of self-respect: those who are loved unconditionally by their parents are better able to grasp that they are valuable irrespective of certain contingent features they possess, whereas those who are not are more likely to mistakenly believe that their value is shaped by these features. Even if a white racist is capable of providing her white child with an abundance of affective care, the child would have her interest in self-respect threatened to the extent that she is aware her parent would not provide her this affective care were she Black. The interest that children have in recognizing their own value, according to Spotorno, is so weighty that it grounds a right on the part of children to be loved unconditionally and a duty on the part of parents to unconditionally love the children they rear. Thus, racist and homophobic parents alike are unfit to rear children because they are incapable of loving, or at least unlikely to love, their children unconditionally.[37] It is worth noting that while Spotorno himself only addresses how racists and homophobes wrong the children they rear, his account would also indict sexists, ableists, xenophobes, and the like for the very same reasons.

So far, I have surveyed two accounts of why certain kinds of objectionably intolerant individuals are unfit to rear children: one explaining why homophobes specifically are generally unfit to rear children, and another explaining why the gamut of objectionably intolerant individuals is unfit to rear children. Now, I develop a third account—an account explaining why the gamut of objectionably intolerant individuals is generally unfit to rear children. This account, while it does not indict all objectionably intolerant individuals as unfit to rear children in every imaginable circumstance, does indict those objectionably intolerant individuals who live in most parts of most multicultural societies as unfit to rear children. Some may take an interest in this third account because

36 For defenses of transracialism, which might someday make such a state of affairs seem less implausible, see Overall, "Transsexualism and 'Transracialism'"; and Tuvel, "In Defense of Transracialism."

37 Spotorno, "Homophobes, Racists, and the Child's Right to Be Loved Unconditionally," 10–15.

they are unconvinced by Spotorno's and desire an account like his that, unlike Brennan and Macleod's account, explains why we may, not only in principle but often in fact, object to racists rearing children. Others may take an interest because they think that in addition to Brennan, Macleod, and Spotorno's arguments, this third account gives due consideration to an interest children have in being raised by sufficiently tolerant parents, an interest that Brennan, Macleod, and Spotorno overlook.[38] Either way, this third account offers a novel contribution to the burgeoning literature on children's rights.

It is clear to me that children have weighty interests in being able to continue associating with particular individuals, adults and children alike.[39] These interests are especially weighty when a child's continued association with another is crucial to the child's well-being. For example, many believe that children of divorce should still associate with the parent who lost custody of them and not just because this association is beneficial to the parent. Such an association is also presumably beneficial to the child, both because the association facilitates valuable goods to the child (e.g., quality time with a loving adult) and because the association is valuable in itself. Call those whose association with children is crucial to their well-being *important associates*. Children often have many important associates in childhood: their parents and siblings, teachers and mentors, neighbors, family friends, the family members of their peers, and, of course, friends. And while, no doubt, parents have the moral authority to impose reasonable time, place, and manner restrictions on the way their children may associate with important associates, I hold that parents lack the moral authority to determine whether their children may continue to associate with

38 An anonymous reviewer suggests that the third account I provide is not meaningfully distinct from Brennan and Macleod's account. Brennan and Macleod argue that parents are under duties to provide affective care to their children. And I argue that parents are under duties to respect the associational rights of children. But plausibly, providing affective care to one's child requires that one respect the associational rights of one's child. This does not pose a problem for my argument. Plausibly, providing affective care to one's child requires that one feed one's child. But the parent does not *merely* wrong her child by failing to provide the child affective care when she does not feed her child. She *also* wrongs the child by violating the child's right against neglect. Similarly, a parent does not *merely* wrong her child by failing to provide the child affective care when she prohibits a child from associating with someone the child is entitled to associate with. She *also* wrongs the child by violating the child's associational rights.

39 Anca Gheaus argues that those who would be beneficial associates to children have rights to associate with those children, and that the children's parents are under an obligation not to prohibit such associations because they are beneficial to children. Gheaus, "The Best Available Parent," 456. In a similar vein, David Meyer draws on United States case law to sketch the beginnings of a theory of children's associational rights. Meyer, "The Modest Promise of Children's Relationship Rights."

these individuals *at all*. To forbid a child from continuing to associate with an important associate is, I claim, to violate an interest-based right that child has to continue associating with such individuals.

I suspect many will accept that there are cases in which an adopted child may have parents who divorce and that the child should be able to continue associating with both parents even if only one is awarded full custody. If you believe an adopted child whose parents divorce should be able to continue associating with the parent who lost custody of them, and you can imagine some extraparental figure (i.e., an adult who associates with a child but is not the child's parent) who associates with a child in a manner that is relevantly similar to the way that the non-custodial parent associates with her child, then you should think that the child should be able to continue associating with the extraparental figure. One might be inclined to argue, at this point, that it is impossible to imagine an extraparental figure whose association with a child is relevantly similar to the non-custodial parent's association with her child because the extraparental figure is not the child's parent, and she must be to be considered one of the child's important associates. But to claim this is to likewise claim that a neighbor who provides a child with refuge from abuse and neglect by her parents is not an important associate of that child, which is absurd. And if you think the parent who won custody of the child in the divorce, by forbidding the child from ever again associating with the parent who lost custody, would violate not only the rights of the non-custodial parent to continue associating with the child but also the rights of the child to continue associating with the non-custodial parent, then you should think that a parent forbidding a child from ever again associating with the extraparental figure in question would violate the rights of the child to continue associating with the extraparental figure.

If we accept that children have certain associational rights, the argument for why the objectionably intolerant are unfit to rear children is straightforward. There is a nontrivial chance that a child will associate with an individual who is gay, or Black, or what have you, and have an extremely weighty interest in continuing to associate with that individual.[40] The objectionably intolerant are at high risk of preventing these important associates from continuing to associate with their children. If someone is at high risk of arbitrarily prohibiting her child from continuing to associate with important associates, then she reveals herself as unfit to rear children, given her willingness to deprive her child of

40 An anonymous reviewer notes that in a case where a strongly anti-Semitic family lives where no Jews live, the parents in that family would potentially be considered fit to raise children on my account. While this is true, it is also the case that in most parts of most multicultural societies, no such comparable conditions obtain, so most parents would be exposing most children to intolerable risks when the parents are objectionably intolerant.

important social, emotional, and relational goods. Therefore, objectionably intolerant individuals are generally unfit to rear children.

Someone might point out that, in some cases, the problem with objectionably intolerant parents rearing children is not that they would prevent their children from associating with certain important associates *altogether*, but that they would prevent their children from associating with certain important associates *as equals*. Consider the following case: Elizabeth is a white woman with a young daughter, Mae Mobley. One of Mae Mobley's important associates is her Black caretaker, Aibileen, whom Elizabeth employed to look after Mae Mobley. Elizabeth, however, is a racist, and while she does not prevent Mae Mobley from associating with Aibileen altogether, she encourages Mae Mobley to regard Aibileen as morally inferior because Aibileen is Black and Mae Mobley is white. Thus, Elizabeth prevents Mae Mobley from associating with Aibileen as *an equal*, and even though she is not barring Mae Mobley from associating with Aibileen *at all*, this nonetheless seems to speak against Elizabeth's fitness to rear Mae Mobley because she potentially deprives her child of important social, emotional, and relational goods by preventing her child from associating with an important associate as her equal.[41]

To evaluate whether preventing a child from associating with an important associate as her equal violates that child's rights, it will be useful to revisit and modify our case concerning the adoptee whose parents divorce. Suppose that one of the parents of the child is awarded full custody and that the other is granted visitation rights, but the parent who is awarded full custody encourages their child to view the parent who is granted visitation rights as morally inferior. Perhaps the reason the divorce precipitated was that the parent who is granted visitation rights cheated on the parent who is awarded full custody, and this is why the parent who is awarded full custody encourages their child to view the parent who is granted visitation rights as morally inferior. If you have the intuition that the parent who is awarded full custody violates their child's rights when preventing her from associating with her parent who is granted visitation rights as her equal, then you should likewise think that Elizabeth violates Mae Mobley's rights when she prevents Mae Mobley from associating with Aibileen as her equal. But I suspect fewer people would have this intuition than those who intuit that preventing a child from associating with an important associate *at all* violates the child's rights. This might be because preventing a child from associating with an important associate deprives the child of *any* of the goods bound up in that relationship, whereas preventing a child from

41 This case is inspired by Elizabeth, Aibileen, and Mae Mobley of Kathryn Stockett's *The Help*.

associating with an important associate *as her equal* deprives the child of perhaps many but not all of the goods bound up in that relationship. And if you subscribe to a child-centered account of child-rearing rights that is sufficientarian in character, it may (but need not) strike you as plausible to say that a child may be prevented from associating with certain important associates as her equals so long as the way she associates with them sufficiently benefits her. Because of this complication, I tentatively suggest that the associational rights of children *might* be violated when they are prevented from associating with their important associates as their equals because I suspect proponents of child-centered accounts of child-rearing rights may reasonably disagree about whether a rights violation occurs. Nevertheless, I more confidently assert that the associational rights of children *are* violated when they are prevented from associating with their important associates altogether because proponents of different child-centered accounts of child-rearing rights would agree that this constitutes a rights violation, given how most recognize that there are at least some cases in which an adoptee would be entitled to associate with both of her parents *somehow* after they divorce even if only one of her parents has full custody of her.

The foregoing discussion has produced three accounts of why objectionably intolerant individuals are unfit to rear children. A strongly homophobic parent, for example, may be said to be unfit to rear a child because (a) her child may grow up to be gay and will need affective care from her that she is unwilling or unable to provide, (b) she is incapable of loving or unlikely to love her child unconditionally, and/or (c) her child may have an interest in a continued association with someone who is gay at some point, and she is likely to put an end to this association. Notice how these arguments need not also indict weakly or moderately homophobic parents as unfit to rear children, especially if it turns out that a child-centered sufficientarian account of child-rearing rights is true. If such an account is true, then children are entitled to be reared by parents who protect their interests to a sufficient degree. If children are entitled to be reared by parents who protect their interests to a sufficient degree, and if it is possible that those who rear children could be weakly or moderately homophobic without (a) being unwilling or unable to provide affective care to their children, (b) being incapable of loving or unlikely to love their children unconditionally or (c) being likely to put an end to their children's association with important associates who are gay, then it is possible that weakly or moderately homophobic parents could nonetheless be entitled to rear their children, provided that their doing so is consistent with their (a) providing a sufficient amount of affective care to their children, (b) being capable of loving and likely to love their children unconditionally, and (c) being unlikely to put an end to their children's

association with important associates who are gay. And if this is true, then the most we can say, at least without further argument, is that *objectionably* intolerant parents lack rights to rear children consistent with child-centered accounts of child-rearing rights.

If regulated parenting is presumptively justified on the grounds that it meets the demands of child-centered accounts of child-rearing rights, and if child-centered accounts of child-rearing rights support the conclusion that objectionably intolerant parents are unfit to rear children, then it follows that whatever forms of regulated parenting policies we institute, at the very least *may*— if not *must*—be designed to protect children from objectionably intolerant child-rearing when feasible. Now, I argue that we have special reason to institute a scheme of parental licensing, given that certain objectionably intolerant individuals are unfit to rear children.

3. THE PROMISE OF PARENTAL LICENSING?

If private parenting were presumptively justified, proponents of regulated parenting would have to explain why the costs imposed on individuals by regulated parenting overcome this presumption. Regulated parenting is presumptively justified, though, so no such an explanation is necessary. And if there were a special presumption against parental licensing, then proponents of regulated parenting who support a policy of parental licensing would have to explain why the costs imposed on individuals by parental licensing overcome this presumption. However, no such presumption exists against parental licensing, so no such explanation needs to be given. In order to argue, then, that a presumption exists in favor of a child welfare policy regime that includes a parental licensing scheme, it would suffice to show that parental licensing is better suited than public parenting support and parental monitoring at protecting the rights of children, at least along a particular dimension.[42]

42 Robert S. Taylor argues that licensing parents as a way to ensure that parents are capable of raising children without governmental assistance enshrines the value of liberal neutrality. By subsidizing the costs of child-rearing through social programs, the liberal state shows favoritism toward those who value a certain life project: parenting. The liberal state does not subsidize the costs of the childless pursuing their preferred life projects. Therefore, the liberal state fails to treat different ways of life neutrally by subsidizing the costs of child-rearing and ought to institute a scheme of parental licensure aimed at ensuring parents are able to support their children financially to meet the demands of liberal neutrality. Taylor, "Children as Projects and Persons." If Taylor is right, then we have more than just the reasons I have given to think that a regulated parenting policy regime should include a policy of parental licensing, in addition to some reason to think that there might be a special presumption against public parenting support.

To be clear, I am under no delusion that a philosopher, from his armchair, can authoritatively prescribe the specifics of a policy regime. The considerations that go into determining whether a particular policy is worth implementing are extremely complicated. So I should not be thought of as arguing that child welfare policy regimes that do not include a scheme of parental licensure are necessarily unjustified, morally speaking. Rather, I am arguing that it is far harder than opponents of parental licensing have thought up until this point to rule out parental licensing as a policy that the state may permissibly implement. This is because there exists no special presumption against parental licensing and because parental licensing is best suited to protecting the right that children have to be reared by sufficiently tolerant parents.

In the preceding section, I argued that children have an interest-based right to be reared by individuals who are not objectionably intolerant. And before I explain how parental licensing is the regulated parenting policy best suited to protecting this right, I want to explain why public parenting support and parental monitoring would, on their own, likely be deficient policies in this regard. Let us start with public parenting support. To refresh, proponents of public parenting support endorse policies like paid parental leave, public childcare, and public subsidies and tax benefits to parents.[43] It is hard to see how policies like the ones just mentioned protect the right that children have not to be raised by the objectionably intolerant. If anything, the public parenting support policies just mentioned would provide objectionably intolerant parents with more resources to provide for those children, if any, who are not adversely affected by their objectionable intolerance, but would fail to insulate those children who *are* adversely affected by their objectionable intolerance from its ill effects.

Samantha Brennan and Colin Macleod, as well as Riccardo Spotorno, suggest that the state may mount advertising campaigns directed at parents on the importance of, in Brennan and Macleod's case, accepting one's child if they come out as gay and, in Spotorno's case, unconditionally loving one's child.[44] Such a policy, I think, can be classified as a policy of public parenting support because the state, by mounting these campaigns, is subsidizing a service that encourages parents to be better child-rearers. I suspect such a policy is likely to be effective in the sense that, over time, children of objectionably intolerant parents would be treated better by their parents than they otherwise would have been. Such a policy would not, however, be effective in the sense that it would protect children from being reared by objectionably intolerant parents

43 Engster, "The Place of Parenting within a Liberal Theory of Justice."

44 Brennan and Macleod, "Fundamentally Incompetent," 239–40; Spotorno, "Homophobes, Racists, and the Child's Right to Be Loved Unconditionally," 17.

who lack the moral authority to rear them in the first place. And if there is a policy that is effective in this *latter* sense, as I will later argue parental licensing is likely to be, then we would have reason to implement such a policy rather than, or in addition to, the kind of policy Brennan, Macleod, and Spotorno advocate for, given that it more directly faces and remedies the problems associated with children being reared by the objectionably intolerant.

Parental monitoring, I argue, is also ill equipped at protecting the right children have to not be reared by objectionably intolerant parents. As a reminder, parental monitoring involves social workers and healthcare professionals visiting households with some degree of regularity and assessing how the interests of the children of the household are protected or promoted by the parents of the household.[45] While parental monitoring may be effective at identifying objectionably intolerant parents once children are old enough and feel secure enough reporting information to social workers about their parents that is relevant to determining whether or not their parents are objectionably intolerant, I suspect it would not be effective at identifying objectionably intolerant parents in the early years of childhood. Parents would have ample opportunity to conceal things about themselves that might lead a social worker to think they are objectionably intolerant when making a home visit. They would also have ample opportunity to coach their young, impressionable children into giving answers that paint the parents in a favorable light to questions that a social worker may ask. In the vast majority of cases, the best it seems parental monitoring could do in terms of protecting children's rights to not be reared by objectionably intolerant parents is to identify the objectionably intolerant after their children have developed significant attachments to them, and place the children of these parents under the care of others who have the moral authority to rear the children. At that point, however, the child would have to suffer not only the harm of being reared by an objectionably intolerant parent but also the harm of being separated from parents to whom she has already, for better or for worse, developed significant attachments. And if a policy can avoid both of these harms, as I will now argue parental licensing can, then we would have reason to implement such a policy rather than or in addition to a policy of parental monitoring because it is able to avoid these harms.[46]

Proponents of parental licensing advocate for public officials to determine standards for parental competency, evaluate whether particular individuals meet these standards, and prevent those who do not meet these standards

45 De Wispelaere and Weinstock, "Licensing Parents to Protect Our Children?"

46 Andrew Jason Cohen criticizes parental monitoring policies and favors parental licensing policies on similar grounds. See Cohen, "The Harm Principle and Parental Licensing," 834 (esp. n20).

from rearing children. We typically think that an activity may be licensed when it is potentially harmful to innocent others, requires a certain level of competence to engage in safely, and when the competence necessary to safely engage in the activity can be determined through a moderately reliable procedure.[47] This is why, for example, the state may license individuals who want to operate a motor vehicle: driving is potentially harmful to innocent others, it requires a certain level of competence to drive safely, and there exist moderately reliable procedures through which we can determine whether individuals are competent to drive. It is uncontroversial, I think, to claim that parenting is a hazardous activity that is potentially harmful to innocent others—that is, children. And, as I established in the previous section, part of being minimally competent with respect to permissibly rearing children is not being objectionably intolerant.

Now, do moderately reliable procedures exist to determine whether prospective parents are objectionably intolerant and therefore unfit to rear children? I think so, and such procedures are largely part and parcel of parental licensing proposals that have been made in the past. Hugh LaFollette defends a parental licensing scheme that denies licenses to prospective parents who are evaluated psychologically and determined to be significantly more likely than not to abuse or abandon their children.[48] Similarly, Andrew Jason Cohen suggests that psychological examinations can be used to determine whether parents have the mental fortitude to deal with the pressures of parenting, and to deny parental licenses to those who lack it.[49] It strikes me that it is most likely during a psychological evaluation that one could determine whether an individual displays objectionably intolerant attitudes toward, e.g., gay people, especially if but not only if the evaluator makes use of an instrument used to measure homophobia in psychological subjects. One such instrument is the Index of Homophobia (IHP), a twenty-five-item questionnaire comprising statements (e.g., "I would feel uncomfortable if my neighbor was homosexual") with which psychological subjects are meant to strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree. A subject's responses to the statements correspond to values that are inputted into an equation that generates a score falling between 0 and 100, and those who receive an IHP score between 75 and 100 are classified as "high grade homophobics."[50] In addition

47 LaFollette, "Licensing Parents," 183.

48 LaFollette, "Licensing Parents," 190–92.

49 Cohen, "The Harm Principle and Parental Licensing," 835.

50 Hudson and Ricketts, "A Strategy for the Measurement of Homophobia," 361. Hudson and Ricketts also write that "on the average, an individual's IHP score will fall within a range of plus or minus 9.5 points of their true score about 95% of the time." Hudson and Ricketts, "A Strategy for the Measurement of Homophobia," 363. Moreover, Costa, Bandeira, and

to psychological evaluations for prospective parents, Robert S. Taylor has suggested that public officials consult "records of past criminal activity, institutionalization for mental health problems, and so on" when determining whether to grant someone a license to parent.[51] In this same spirit, I suggest public officials conduct background checks to determine whether prospective parents have been, to give two examples, convicted of a hate crime or successfully sued for employment discrimination. These background checks might also be used to determine if prospective parents are affiliated with organizations that would give us reason to believe they are objectionably intolerant. An example of a proverbial "red flag" in this regard would be prospective parents who are active members of the Westboro Baptist Church.[52]

So, I submit that there exist moderately reliable procedures—not all too different from those that have been advocated for by past proponents of parental licensing—for determining whether parents are objectionably intolerant and therefore unfit to rear children. Even if these procedures would fail to catch many prospective parents who are objectionably intolerant, they would still catch some, and that would be enough to justify implementing a parental licensing scheme that uses these procedures since I showed earlier that there are neither presumptions in favor of private parenting nor against parental licensing to be overcome. Even if we can only protect some children from having their right to be reared by sufficiently tolerant parents violated through the use of these procedures, that is still much better than not preventing any from having that right violated. And if some individuals who would make sufficiently tolerant parents are inadvertently deemed by the licensing scheme to be unfit to rear children, that will not violate their rights since, as I argued, people do not have a right to rear children in the first place.

Public parenting support, on its own, can only mitigate the ill effects of children being reared by objectionably intolerant parents, whereas parental licensing can protect children from being reared by objectionably intolerant

---

Nardi rate the IHP very highly among existing instruments used to measure homophobia because it reliably predicts levels of homophobia in psychological subjects in diverse populations, contexts, and cultures. See Costa, Bandeira, and Nardi, "Systematic Review of Instruments Measuring Homophobia and Related Constructs," 1329.

51 Taylor, "Children as Projects and Persons," 570.

52 The Westboro Baptist Church is an unaffiliated Primitive Baptist Church in Topeka, Kansas, that is classified by the Anti-Defamation League and the Southern Poverty Law Center as a hate group, in large part because of the many homophobic pickets the church's members participate in across the United States. See "Westboro Baptist Church," Anti-Defamation League, February 8, 2017, https://www.adl.org/resources/profiles/westboro-baptist-church; and "Westboro Baptist Church," Southern Poverty Law Center, https://www.splcenter.org/fighting-hate/extremist-files/group/westboro-baptist-church.

parents *at all*. And parental monitoring would inevitably subject the children of objectionably intolerant parents to the harms of both being reared by objectionably intolerant parents and being separated from parents, no matter how unfit, to whom the child already developed significant attachments, whereas parental licensing could protect children from *both* of these harms. With respect to protecting the right that children have to be reared by sufficiently tolerant parents, parental licensing is best suited out of the available regulated parenting policies to achieve this goal. This provides us with special reason to think that a regulated parenting policy regime should include a policy of parental licensure. And for those who think that enforcing a parental licensing scheme would threaten the rights that individuals have to protected relationships with their biological children, we could simply amend the policy proposal such that individuals who are deemed unfit to rear children and are subsequently denied the opportunity to rear their biological children are granted visitation rights with respect to their biological children on the condition that granting these rights is consistent with their children's interests.

My arguments have significant implications for how philosophical debates concerning child welfare policy should be conducted. Those who oppose parental licensing who are proponents of private parenting will have to reconceive their arguments to account for the fact that regulated parenting is presumptively justified, whereas private parenting is not. And those who oppose parental licensing who are proponents of different forms of regulated parenting must revise their arguments in the following ways. They will have to either account for the fact that there exists no presumption against parental licensing because it does not run the risk of violating peoples' rights to rear children or they will have to offer a defense of the right to rear children that does not suffer the problems I have pointed out in this essay or some other right and explain how this right is jeopardized by a parental licensing scheme and not by policies of public parenting support or parental monitoring. Or if they are unable to, they will have to either show that children have no right against being reared by objectionably intolerant individuals or they will have to show how either or both public parenting support and parental monitoring are best suited to protecting rights undergirded by weightier interests than the interests undergirding the right against being reared by objectionably intolerant individuals and that parental licensing jeopardizes these rights. In other words, my arguments make it considerably more difficult for opponents of parental licensing to establish a successful case against it.

The reason I think it is so important to shine light on how difficult it is to make the case against parental licensing is that I cannot shake the feeling that parental licensing could redound to the benefit of a great many children, and

it is imperative for this reason that we remain open to it as a policy possibility. It is no secret that many are unfit to rear children. And it is no secret that many who are demonstrably unfit to rear children nevertheless are permitted to do so. If parental licensing has the potential to protect children from the havoc these people could wreak on their lives, we should certainly remain open to it in the absence of reasons to think such a policy is objectionable in principle. Just as I think it would be hasty to conclude from the armchair that a scheme of parental licensing must now be instituted given the arguments I have made, I think it is hasty for opponents of parental licensing to conclude from the armchair that a scheme of parental licensing must *never* be instituted given the practical difficulties we anticipate facing when implementing such a scheme.[53] At the very least, I hope to have made the case for parental licensing seem far less implausible than critics of the policy seem to have thought it was up until this point.[54]

University of Colorado Boulder
connor.kianpour@colorado.edu

REFERENCES

Archard, David. "Child Abuse: Parental Rights and the Interests of the Child." *Journal of Applied Philosophy* 7, no. 2 (October 1990): 183–94.

Brennan, Samantha, and Colin Macleod. "Fundamentally Incompetent: Homophobia, Religion, and the Right to Parent." In *Procreation, Parenthood, and Educational Rights: Ethical and Philosophical Issues*, edited by Jaime Ahlberg and Michael Cholbi, 230–45. New York: Routledge, 2017.

53 Christopher Freiman offers a battery of practical objections against parental licensing. Freiman, "Against Parental Licensing," 118–21. And while I agree with him that his objections would rule out the possibility of parental licensing if there were a special presumption against parental licensing, they fail to rule out the possibility of parental licensing given that there is actually a presumption *in favor* of its inclusion in a regulated parenting policy regime.

54 I am grateful to Chris Freiman, both for writing the essay that inspired my own and for offering me feedback on an earlier version of this piece. I also owe a huge intellectual debt to Bob Taylor and Andrew Jason Cohen, both of whom mentored me at different stages of my academic career and played a significant role in shaping my views about parental licensing. Thanks also to David Boonin, who read and provided feedback on all too many earlier versions of this piece. Finally, I want to thank Andrew Jason Cohen, Tom Crean, three anonymous reviewers, and an associate editor at *JESP* for reading and offering feedback on earlier drafts of this essay.

Brighouse, Harry, and Adam Swift. *Family Values: The Ethics of Parent-Child Relationships.* Princeton: Princeton University Press, 2014.

———. "Parents' Rights and the Value of the Family." *Ethics* 117, no. 1 (October 2006): 80–108.

Cohen, Andrew Jason. "The Harm Principle and Parental Licensing." *Social Theory and Practice* 43, no. 4 (October 2017): 825–49.

Costa, Angelo Brandelli, Denise Ruschel Bandeira, and Henrique Caetano Nardi. "Systematic Review of Instruments Measuring Homophobia and Related Constructs." *Journal of Applied Psychology* 43, no. 6 (June 2013): 1324–32.

De Wispelaere, Jurgen, and Daniel Weinstock. "Licensing Parents to Protect Our Children?" *Ethics and Social Welfare* 6, no. 2 (May 2012): 195–205.

Engster, Daniel. "The Place of Parenting within a Liberal Theory of Justice: The Private Parenting Model, Parental Licenses, or Public Parenting Support?" *Social Theory and Practice* 36, no. 2 (April 2010): 233–62.

Freiman, Christopher. "Against Parental Licensing." *Journal of Social Philosophy* 53, no. 1 (Spring 2022): 113–26.

Gheaus, Anca. "The Best Available Parent." *Ethics* 131, no. 3 (April 2021): 431–59.

———. "The Right to Parent One's Biological Baby." *Journal of Political Philosophy* 20, no. 4 (December 2012): 432–55.

———. "Sufficientarian Parenting Must Be Child-Centered." *Law, Ethics and Philosophy* 5, no. 5 (June 2018): 189–97.

Hannan, Sarah, and Richard Vernon. "Parental Rights: A Role-Based Approach." *Theory and Research in Education* 6, no. 2 (July 2008): 173–89.

Hudson, Walter W., and Wendall A. Ricketts. "A Strategy for the Measurement of Homophobia." *Journal of Homosexuality* 5, no. 4 (October 2010): 357–72.

LaFollette, Hugh. "Licensing Parents." *Philosophy and Public Affairs* 9, no. 2 (Winter 1980): 182–97.

———. "Licensing Parents Revisited." *Journal of Applied Philosophy* 27, no. 4 (November 2010): 327–43.

Leighton, Kimberly. "Addressing the Harms of Not Knowing One's Heredity: Lessons from Genealogical Bewilderment." *Adoption and Culture* 3 (2012): 63–107.

Liao, S. Matthew. "Biological Parenting as a Human Right." *Journal of Moral Philosophy* 13, no. 6 (November 2016): 652–68.

Lomasky, Loren E. *Persons, Rights, and the Moral Community*. Oxford: Oxford University Press, 1990.

Mangel, Claudia Pap. "Licensing Parents: How Feasible?" *Family Law Quarterly* 22, no. 1 (1988): 17–39.

McFall, Michael T. *Licensing Parents*. Lanham, MD: Lexington Books, 2010.

Meyer, David D. "The Modest Promise of Children's Relationship Rights." *William and Mary Bill of Rights Journal* 11, no. 3 (April 2003): 1117–37.

Narveson, Jan. *The Libertarian Idea*. Philadelphia, PA: Temple University Press, 1989.

———. *Respecting Persons in Theory and Practice: Essays on Moral and Political Philosophy*. Lanham, MD: Rowman and Littlefield, 2002.

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

Okin, Susan M. *Justice, Gender, and the Family*. New York: Basic Books, 1989.

Overall, Christine. "Transsexualism and 'Transracialism.'" *Social Philosophy Today* 20 (2004): 183–93.

Sandmire, Michael J., and Michael S. Wald. "Licensing Parents: A Response to Claudia Mangel's Proposal." *Family Law Quarterly* 24, no. 1 (Spring 1990): 53–76.

Schoeman, Ferdinand. "Rights of Children, Rights of Parents, and the Moral Basis of the Family." *Ethics* 91, no. 1 (October 1980): 6–19.

Shields, Liam. "How Bad Can a Good Parent Be?" *Canadian Journal of Philosophy* 46, no. 2 (April 2016): 163–82.

———. *Just Enough: Sufficiency as a Demand of Justice*. Edinburgh: Edinburgh University Press, 2016.

———. "Parental Rights and the Importance of Being Parents." *Critical Review of International Social and Political Philosophy* 22, no. 2 (2019): 119–33.

Somerville, Margaret A. "Children's Human Rights to Natural Biological Origins and Family Structure." *International Journal of the Jurisprudence of the Family* 1 (2010): 35–54.

Spotorno, Riccardo. "Homophobes, Racists, and the Child's Right to Be Loved Unconditionally." *Critical Review of International Social and Political Philosophy* (forthcoming). Published online ahead of print, March 26, 2021. https://www.tandfonline.com/doi/full/10.1080/13698230.2021.1905467.

Steiner, Hillel. *An Essay on Rights*. Oxford: Blackwell, 1994.

Stockett, Kathryn. *The Help*. New York: Penguin Books, 2009.

Taylor, Robert S. "Children as Projects and Persons: A Liberal Antinomy." *Social Theory and Practice* 35, no. 4 (October 2009): 555–76.

Tuvel, Rebecca. "In Defense of Transracialism." *Hypatia* 32, no. 2 (Spring 2017): 263–78.

Vallentyne, Peter. "The Rights and Duties of Child-rearing." *William and Mary Bill of Rights Journal* 11 (2003): 991–1009.

Journal of Ethics and Social Philosophy
VOL. 25, NO. 3 · SEPTEMBER 2023
https://doi.org/10.26556/jesp.v25i3.2301

INSTITUTIONAL CORRUPTION

THE TELEOLOGICAL AND NONNORMATIVE ACCOUNT

Armin W. Schulz

CORRUPTION is widely recognized to be a major social problem, but its characterization continues to be very controversial. So, while it is frequently noted that corruption is "the abuse of power by a public official for private gain," not all corruption needs to involve public officials (doctors need not be public officials but can be corrupt if they prescribe medicine in accordance with who pays them to do so, rather than with what is best for the patient) or involve a private gain (when a county clerk grants wedding licenses in line with their personal moral or religious convictions and not the law, it can be a case of corruption but need not involve any private gain whatsoever).[1]

Indeed, it is now commonly noted that what is being corrupted need not be an individual person at all but can be an entire social institution.[2] This kind of institutional corruption has, especially in the last few years, come to be seen as ever more central and important.[3] Many of the major contemporary social problems appear to center on the undermining of institutions like voting, the free press, policing, or health care: instead of every citizen being equally able to influence political decision-making, to be informed about what is going on in the wider society, to be secure, or to be healthy, the institutions meant to provide these goods often seem to fail in their task.[4] This form of corruption thus deserves—and has seen—significant amounts of scrutiny in the last few years.

However, it continues to be a challenge to specify exactly what makes something a case of institutional corruption (IC).[5] Exactly which actions subvert the

1 Nye, "Corruption and Political Development."

2 Thompson, *Ethics in Congress*; Lessig, "'Institutional Corruption' Defined"; Miller, *Institutional Corruption*; Ferretti, "A Taxonomy of Institutional Corruption."

3 Thompson, "Theories of Institutional Corruption"; Miller, *Institutional Corruption*.

4 Thompson, *Ethics in Congress*; Lessig, "'Institutional Corruption' Defined"; Miller, *Institutional Corruption*.

5 Thompson, "Theories of Institutional Corruption"; Ferretti, "A Taxonomy of Institutional Corruption."

relevant institution, and exactly why is it the case that these actions subvert the institution? What, specifically, is an institution's purpose? This paper seeks to further the debate surrounding IC by answering these questions. After all, without a clear characterization of the nature of IC, fighting or avoiding it is difficult—for it is then not clear precisely what is to be fought or avoided.[6]

The paper, therefore, presents a general, philosophically and social scientifically well-grounded theory of IC that is centered on the idea that institutions have a social and not inherently normative function that is being subverted in cases of IC. While this theory shares some superficial components with some of the existing ones in the literature—especially those of Lessig and Miller—it is, in fact, quite different from the latter.[7] In particular, by being built on the most compelling form of social functionalism, the theory presented here has a solid theoretical foundation, does justice to the complex ethical nature of IC, and is in line with work in the social sciences more generally. Moreover, this theory is shown to have several important novel features: it is graded (institutions can be more or less corrupted), general (it can be applied to political contexts, but also many other social phenomena, from social media to private corporations and nongovernmental organizations like the International Federation of Association Football [FIFA]), and unifying (it makes clear why highly corrupt societies tend to become unstable, whatever exactly the cause or moral status is of the corruption).

The paper is structured as follows. Section 1 lays out the nature of IC and develops desiderata for its characterization. In order to provide a grounding for the functional ascription at the heart of IC, section 2 presents the currently most compelling form of social functionalism. Section 3 uses this account of social functionalism to develop a new non-normative teleological theory of IC that satisfies the desiderata of section 1. Section 4 concludes.

1. INSTITUTIONAL CORRUPTION

Human social living centers around social institutions: the "rules of the game" that structure human interactions and which set out the kinds of behaviors that, in a given type of situation, members of the society are expected to—and expect others to—engage in.[8] Social institutions, in this standard social scientific sense, comprise a vast array of familiar aspects of contemporary social

6 Rothstein and Varraich, *Making Sense of Corruption*.

7 See Lessig, "'Institutional Corruption' Defined"; Miller, *Institutional Corruption*.

8 Parsons, *The Social System*; North, *Institutions, Institutional Change, and Economic Performance*.

living, from the structure of the government (e.g., representative democracy) and the economy (e.g., free enterprise) to that of the family (e.g., polyandry) and religion (e.g., Hinduism). Note that it need not be obvious *why* social institution *N* prescribes behavior *B* in situation *S*—i.e., what the function of the institution is. Similarly, it is not presumed that the behavior prescribed by the institution is morally obligatory: institutionally based norms are not necessarily moral norms.[9] All that matters is that institutions dictate the norms of behavior for a given society.

Given this, IC concerns cases where people engage in actions that undermine a particular social institution. These actions need not involve a private gain or *quid pro quo* exchanges of favor; indeed, these actions need not be inherently immoral or illegal. However, these actions still prevent the institution from operating as it is meant to. Such cases have come to be seen as being of major importance when it comes to ensuring that societies function in ways that benefit all their members.[10]

For example, in a given democracy, elections might be won only if candidates can obtain vast amounts of funding from major sponsors: only this ensures that they get heard or seen by voters. In that case, though, the only candidates who have a chance of obtaining office are those able to attract the necessary funds to finance their campaigns. This gives big political donors (businesses or wealthy individuals) an outsize influence on the running of the democracy. In turn, this can cause ordinary voters to feel like their voices do not matter, so they cease to participate in the political process. Thus, decisions are made in line with who can pay for access to these lawmakers, not who voted for them. At its extreme, this can spell the end of the relevant democracy. Similar points can be made about other examples, such as the privatization of prisons—which incentivizes incarceration rates and can thus decrease public security, in opposition to what prisons are for—and the mass dissemination of misleading or false information—which can undermine belief in public information of any kind.[11]

9 Miller employs a morally loaded notion of social institution that is furthermore restricted to *organizations* (roughly, complex structures of organized sets of norms). See Miller, *Institutional Corruption*. However, as noted in the text, this is not the standard notion used in the social scientific literature.

10 Miller, *Institutional Corruptiony*; Lessig, "'Institutional Corruption' Defined"; Thompson, "Theories of Institutional Corruption"; Ferretti, "A Taxonomy of Institutional Corruption"; Ceva and Ferretti, "Political Corruption, Individual Behaviour and the Quality of Institutions"; Della Porta and Vannucci, *The Hidden Order of Corruption*.

11 Satz, "Markets, Privatization, and Corruption"; Tsfati, et al., "Causes and Consequences of Mainstream Media Dissemination of Fake News."

Cases like these have come to be seen to be of major importance: they are at the heart of some of the most widely discussed issues afflicting many contemporary societies.[12] A number of theoretical proposals have been put forward to make the nature of institutional undermining that underlies them more precise.[13]

So, Thompson argues that IC concerns cases where public officials—especially legislators—receive political gains for providing services that are "procedurally improper" and that have a tendency to damage the political process.[14] Services are procedurally improper when they are not determined on the merits of the case, and/or they fail to follow the rules that ensure the political process is fair. If done systematically, such services can erode the public confidence in the political process—i.e., corrupt political institutions.

Not unrelatedly, Warren characterizes IC as instances where public officials *claim* to respect the egalitarian idea that all individuals affected by the collective decisions of the public officials should be able to influence these decisions, but where these officials *in fact* make their decisions so as to favor those who have provided benefits to these officials, and thus have privileged access to them.[15] In other words, according to Warren's account, IC is at heart about duplicitous violations of democratic egalitarian ideals: public officials pretend to uphold these ideals, but do not actually do so, and that in a way that is in fact harmful to some members of the public.

There is no question that both of these characterizations of IC have allowed for many useful insights and advances. Most obviously, the problems caused by some forms of campaign finance for contemporary US democratic processes are well illuminated by both of these accounts: such forms of campaign finance can be procedurally improper and in violation of egalitarian ideals of democratic political decision-making. Beyond this, the abstractness, especially of Warren's account, also makes clear what is wrong with other ills afflicting contemporary (representative) democracies, such as gerrymandering and voter suppression. These are cases that violate the egalitarian ideals at the heart of a genuine democracy—and they do that in a way that is surreptitious and thus hard to notice, avoid, and combat.

12 Satz, "Markets, Privatization, and Corruption"; Miller, *Institutional Corruption*; Lessig, "'Institutional Corruption' Defined"; Thompson, "Theories of Institutional Corruption."

13 For helpful surveys, see, e.g., Thompson, "Theories of Institutional Corruption"; Ferretti, "A Taxonomy of Institutional Corruption"; Brock, "Institutional Integrity, Corruption, and Taxation."

14 Thompson, *Ethics in Congress*. See also Philp, "Defining Political Corruption."

15 Warren, "What Does Corruption Mean in a Democracy?," "Political Corruption as Duplicitous Exclusion," and "The Meaning of Corruption in Democracies."

However, both of these proposals also struggle to go beyond this sociopolitical context and analyze IC more generally. It is not clear that these two proposals can be used to understand the IC of, say, prisons, the press, corporations, and not just that of political decision-making in representative democracies (and the US specifically). For example, the privatization of prisons is not obviously procedurally improper or done in a way that is democratically duplicitous. The issue with this privatization is not how it came about, which may have been entirely proper, or that it is inegalitarian, which it need not be, but that it undermines the institution it concerns. Much the same is true when it comes to the mass dissemination of misleading or false information (the source of which need not even be a public official at all). What matters is just that it concerns an undermining of the public press, not how it was decided on. In short: since IC is widely seen to comprise cases other than those of campaign donation in representative democracies, the proposals of Thompson and Warren appear insufficiently general—whatever other virtues they have.[16]

The account of Lessig is, therefore, a step in the right direction.[17] According to Lessig, "institutional corruption is manifest when there is a systemic and strategic influence … that undermines the institution's effectiveness by diverting it from its purpose or weakening its ability to achieve its purpose."[18] This influence need not be illegal, immoral, or procedurally improper; the key is just that it thwarts the *function* of the relevant institution. In this way, this account is significantly more general than the ones of Thompson and Warren. While it remains the case that the account in Lessig also tends to focus on the kind of ("dependence") corruption of the democratic political process that Thompson and Warren focus on, there is no reason that it cannot be easily extended to cover the corruption of the prison system, the press, and other public or even private institutions; indeed, it has been applied to the pharmaceutical industry with much success.[19]

The main challenge the account faces is that it leaves open exactly what the function of a social institution is. What are prisons, or the press, or the National Collegiate Athletic Association (NCAA) for? Because of this, it also remains somewhat unclear exactly how this function can be undermined. Is the rise of social media undermining the press? Why? Without spelling this out,

16 See also Miller, *Institutional Corruption*, 300–4. For a historical study of political corruption, see also Sparling, *Political Corruptiony*.

17 See Lessig, "'Institutional Corruption' Defined"; this is further developed in Lessig, *America, Compromised*.

18 Lessig, "'Institutional Corruption' Defined," 553.

19 See especially Lessig, *America, Compromised*; and Fields, "Parallel Problems."

the account lacks a thorough theoretical grounding.[20] Now, given that Lessig's focus also is the IC of the US political system—whose function may be relatively clear—this need not be greatly problematic for many of the uses Lessing has put his account to.[21] However, as a full account of IC, Lessig's account falls short; while it has a sufficiently general overarching structure, this structure is not spelled out in enough detail to be able to make sense of IC in all of its different facets.

The account of Miller attempts to fill this lacuna.[22] Like Lessig's, the account is teleological and general in nature; however, unlike that of Lessig, it is more fully spelled out.

According to Miller, social institutions are organizations—i.e., sets of structurally related functional roles—that provide "collective goods by means of joint action."[23] That is, on this account, the purpose of a social institution is the provision, through the joint activity of the members of the institution, of objectively moral goods that are made available to all members of the relevant society.[24] These goods comprise aggregated (needs-based) moral rights, freedoms, or well-being.[25] Note that it is not sufficient that an organization provides collective goods that are *thought* to be moral goods; only organizations that provide collective goods that are *in fact* moral goods qualify as genuine social institutions.[26] In this way, the account of Miller makes it possible to provide a precise and systematic statement of what makes it the case that a given social institution has whatever function it has: namely, the fact that the collective intentions and actions of the members of the relevant society create institutions whose end is the obtaining of a collective, objectively moral human good. In turn, this also allows for a clear and general account of IC. IC occurs when members of an institution intentionally engage in actions that tend to have the foreseeable and/or avoidable effect of undermining the function—spelled out as above—of the relevant institution (without, though, destroying that institution).[27]

The account of Miller—like that of Lessig—is appropriately general. Since it makes the teleological nature of IC central to its characterization, it is not

20 Amit, et al., "Institutional Corruption Revisited"; Thompson, "Theories of Institutional Corruption."

21 Thompson, "Theories of Institutional Corruption."

22 Miller, *Institutional Corruption*.

23 Miller, *Institutional Corruption*, 23, 26.

24 Miller, *Institutional Corruption*, 106.

25 Miller, *Institutional Corruption*, 23.

26 Miller, *Institutional Corruption*, 23, 28, 34–45.

27 Miller, *Institutional Corruption*, 82–88.

restricted or tied to the corruption of the democratic political process (important as that may be). Instead, it can be straightforwardly extended to other phenomena—such as the corruption of prisons or the press—for these two are instances where the provision of the collective moral goods (security and transparency) is thwarted.[28] Furthermore, the account of Miller improves on the one of Lessig, as it spells out this teleological nature of IC in detail.[29] The function of social institutions is not left open as something to be filled in by the whims of the relevant researchers, but it is underwritten by a philosophically well-grounded treatment. However, Miller's account also faces three key drawbacks.

First, the theory does not speak to (what Miller calls) institutional corrosion (where actions are done that happen to slightly undermine the function of an institution but which fail the conditions for IC set out above), institutional destruction (where the institution is fully destroyed), or externally perpetrated IC.[30] However, this restrictive focus of the analysis is not greatly compelling. Institutional corrosion, destruction, and external IC all lead to the same kind of failure of the provision of the relevant collective good as IC in its proper sense according to Miller. While the source and exact nature of the prevention of the provision of the relevant collective moral good are different, the fact that there is this prevention is not. In this way, the account is overly limited. This is an important point to which I return below.

Second, the account of Miller needs to make strong commitments to highly contentious philosophical doctrines, such as a strong moral realism and methodological individualism. However, it is far from clear that these commitments are justified. For example, it is not obvious that the existence of objective moral facts—such as which collective goods are *in fact* morally ood—can be made plausible.[31] Similarly, there are some good reasons to think that a strong methodological individualism is not compelling in the social sciences in general.[32]

Third and most importantly, Miller's account is made problematic by the fact that it is fundamentally normative. On this account, IC *must* be morally bad (at least *pro tanto*): the moral appraisal of IC (and of social institutions in general) is *built into* the nature of IC (and institutions in general).[33] However,

28 Miller, *Institutional Corruption*, 217–18.

29 A point also noted by Thompson, "Theories of Institutional Corruption."

30 Miller, *Institutional Corruption*, 66, 70.

31 Mackie, *Ethics*; Joyce, *The Myth of Morality*; Street, "A Darwinian Dilemma for Realist Theories of Value."

32 Ruiz and Schulz, "Microfoundations and Methodology."

33 Miller allows for the existence of "noble cause corruption," but this would be the case where corruption is engaged in for a (*pro tanto*) morally defensible reason (*Institutional Corruption*). However, this does not affect the main point in the text.

this fails to do justice to the moral complexity of IC.[34] When it comes to the moral status of IC, everything depends on the details of the case and should not be built into the characterization of the nature of IC. People can engage in actions that lead to or constitute IC, but these actions can be morally neutral or even morally good (e.g., when the relevant social institutions are morally problematic).[35]

Put differently, the normative focus of Miller's account makes this account arbitrarily limited. From the point of view of the underlying causal mechanisms—i.e., from the perspective of what is happening to the relevant institutions—the IC of the Mafia or the Nazi Party may be identical to that of US representative democracy or the press. While our normative evaluation of the former two cases may be different from the latter two, the social phenomenon underlying the four cases is the same: they share the crucial feature of thwarting the purpose of a social institution. They should thus be treated in the same way, too.[36] This is an important point to which I return in section 3.

The point is further strengthened by the fact that the normative focus of Miller's account does not fit the long tradition of functional ascription in the social sciences more generally.[37] The next section lays this out in more detail, but for now, the key point to note is just that, according to the most compelling accounts, functional ascription in the social science is *not* fundamentally normative in the way that Miller's account is. Rather, in the social sciences, it is common to ascribe *nonnormative* functions to social institutions. Hence, Miller's picture of functional ascription does not match that of the social sciences

34 Lessig, "'Institutional Corruption' Defined"; Thompson, "Theories of Institutional Corruption."

35 This makes this different from some other related phenomena. For example, arguably, abusing one's power is always (*pro tanto*) morally bad: it concerns cases where a person acts against the reasons why they are in a position of power. It may be that a person aims at morally defensible outcomes by abusing their power, but the fact that they achieve these outcomes by abusing their power is one (moral) reason that speaks against doing so. However, this is different from cases of institutional corruption: the latter does not directly refer to ways of acting, but to the status of a social institution, *viz*., whether it is well-functioning. Put differently, an abuse of power *can result in* the corruption of an institution—but the latter can also result from behavior that is not an instance of the abuse of power. Importantly, also, since institutions can be morally good or bad, the well-functioning of these institutions can be morally good or bad as well. I thank Dale Dorsey for useful discussion of this issue.

36 Miller considers the latter a case of "organizational corruption," and thus excludes it from the analysis (*Institutional Corruption*, 28).

37 See, e.g., Durkheim, *The Elementary Forms of the Religious Life*; Malinowski, *Argonauts of the Western Pacific*; Elster, *Ulysses and the Sirens*; Pettit, "Functional Explanation and Virtual Selection"; Bigelow, "Functionalism in Social Science."

more generally—which is problematic, as the investigation of IC is a part of the social sciences.

Putting all of this together, it becomes clear that what is still needed is an account of IC that has the following three features:

1. *General*: The account needs to focus on the teleological nature of IC generally and not be restricted to the undermining of the (US) political process only.
2. *Spelled out*: The account needs to ground the function of social institutions in a plausible theoretical treatment and not leave it open to the intuitions of a researcher.
3. *Nonnormative*: The account needs to spell out the function of social institutions in a way that does not presuppose that this function aims at some human good; rather, the moral valence of the social institution needs to be assessed depending on the details of the case.

An account that satisfies these three desiderata is able to combine the best features of the existing characterizations of IC while avoiding their drawbacks.

To make headway in developing such an account, the next section outlines the currently most compelling theoretical framework for functional ascription in the social sciences. On this basis, section 3 lays out a novel account of IC that satisfies desiderata 1–3 and that has some further useful implications.

Before doing this, though, it is important to note that implicit in desiderata 1–3 is the idea that IC is, in its nature, quite different from individual corruption. As just noted, this is a common assumption in many views of IC (notably those of Thompson and Lessig), but it is not without controversy. For example, some authors argue that IC reduces to individual corruption and that strongly separating out individual from IC obfuscates the mechanisms by which corruption spreads from one institutional context to another.[38] Relatedly, it is implicit in desiderata 1–3 that IC is to be analyzed teleologically (in terms of what the purpose is of the relevant social institutions) and not, say, deontologically (in terms of what it is our duty to do as members of a certain institution) or in terms of virtue (in terms of what virtuous members of the social institution are like).[39]

Without question, there is a lot that could be said about these alternative, individualistic treatments of IC. However, instead of engaging in these debates directly, the approach here is the reverse. The paper shows that adopting a teleological and non-individualist perspective on IC is coherent and has several

38 Ceva and Ferretti, "Political Corruption, Individual Behaviour and the Quality of Institutions"; Ferretti, "A Taxonomy of Institutional Corruption." See also Philp, "Defining Political Corruption."

39 See, e.g., Rothstein and Varraich, *Making Sense of Corruption*.

advantages. In turn, this provides a reason for adopting this kind of view of IC. Of course, no pretense here is made that this has settled all the issues surrounding this issue (or, indeed, IC in general). Rather, the aim is more modest: it is just to show that a compelling, teleological, and non-individualist perspective on IC is available. If an alternative treatment is to be adopted, it would have to be shown to be superior while taking these benefits into account. (I return to these points below.)

With this in mind, consider ways of spelling out the function of a social institution. This is important, as the very nature of institutional purpose is sometimes seen as incoherent—which would thus make it a highly problematic basis for an account of IC. As the next section makes clear, though, this impression is misleading.

2. SOCIAL FUNCTIONALISM

There is a long tradition in the social sciences that sees value in analyzing social institutions in terms of their function.[40] By understanding what a social institution is for, it is thought that we can better grasp what the institution is, how it relates to other social institutions, how stable it is, and how to best alter it. However, this functionalist approach toward social science has also been faced with some major criticisms; in particular, it is thought that it cannot be made empirically plausible.[41] As it turns out, though, recent advances in this area make clear that social functionalism is, in fact, a compelling and well-grounded research program in the social sciences.

To see this, begin by noting that, according to the traditional version of social functionalism, what grounds the function of a social institution is some form of biocultural evolution.[42] This account of functional ascription can be related to a parallel development in the biological and cognitive sciences.[43]

40 See, e.g., Durkheim, *The Elementary Forms of the Religious Life*; Malinowski, *Argonauts of the Western Pacific*; Merton, *Social Theory and Social Structure*; Elster, *Ulysses and the Sirens*; Pettit, "Functional Explanation and Virtual Selection"; Bigelow, "Functionalism in Social Science." Of course, functionalist accounts of various phenomena go back at least to Aristotle. However, as far as the discussion in the social sciences is concerned, the classic, "traditional" sources are the ones cited in the text.

41 Elster, *Ulysses and the Sirens*; Pettit, "Functional Explanation and Virtual Selection."

42 Bigelow, "Functionalism in Social Science"; Rosenberg, *Philosophy of Social Science*; Kincaid, "Assessing Functional Explanations in the Social Sciences"; Elster, *Ulysses and the Sirens*.

43 Millikan, *Language, Thought, and Other Biological Categories*; Millikan, *Varieties of Meaning*; Papineau, *Reality and Representation*; Neander, "Content for Cognitive Science"; Garson, "Function, Selection, and Construction in the Brain"; Papineau and Garson, "Teleosemantics, Selection and Novel Contents." Note that these views differ in numerous particulars,

On these models, the function of the human heart is to pump blood because pumping blood is what the heart was selected for. Humans with hearts that pumped blood (or whose hearts pumped blood more reliably or efficiently) had a greater expected reproductive success than those whose hearts did not pump blood (or as reliably or efficiently). Other features of the heart—such as the noise they make—did not contribute to their expected reproductive success. Hence, it is the fact that hearts pump blood (reliably or efficiently)—not that they make a certain kind of sound—that should be taken for their function for this supported their spread in the population.[44]

Transposing this to the social realm, a number of authors have argued that a given social institution *N* has the function *F* if past tokens of *N* were biologically or culturally selected to do *F*.[45] If past tokens of *N* that did *F* had a higher chance to reproduce *N* than those tokens of *N* that did not have *F*, then *N* (now) has the function to do *F*. In short, functional ascription is about identifying the selective reasons for the spread of an institution or trait.[46]

However, this way of grounding functional ascription in the social sciences faces what has become known as the "missing mechanisms argument."[47] At the heart of this argument is the claim that few social institutions have the kind of selective history needed for them to have a function of the above sort. Hence, they either need to be seen to have no function—thus undercutting the

---

but for present purposes, these differences do not matter. Note also that there is an alternative view of functional ascription in the biological and cognitive sciences according to which the latter is to be grounded in the causal roles a given trait or component plays in a larger causal system: see, e.g., Cummins, "Functional Analysis." Interestingly, some classic works in the structuralist functionalist tradition in the social sciences follow this line, too: they see social institutions as akin to elements in a large social system and ground their functions accordingly (see, e.g., Parsons, *The Social System*). However, views like this face the problem that it makes the nature of a function observer-dependent: depending on what causal system is chosen, the causal role function of a trait or institution will differ. This is particularly problematic in the social sciences where one of the major reasons why a functionalist approach was sought in the first place lies in the fact that it allows for an analysis of social dynamics that is not purely observer dependent. For this reason, the etiological accounts have come to dominate the discussion there.

44 Millikan, *Language, Thought, and Other Biological Categories* and *Varieties of Meaning*.

45 Bigelow, "Functionalism in Social Science"; Rosenberg, *Philosophy of Social Science*; Kincaid, "Assessing Functional Explanations in the Social Sciences"; Elster, *Ulysses and the Sirens*; Rappaport, *Pigs for the Ancestors* and *Ecology, Meaning and Religion*.

46 Millikan, *Language, Thought, and Other Biological Categories* and *Varieties of Meaning*; Papineau, *Reality and Representation*; Neander, "Content for Cognitive Science"; Garson, "Function, Selection, and Construction in the Brain."

47 Elster, *Ulysses and the Sirens*; Pettit, "Functional Explanation and Virtual Selection"; Bigelow, "Functionalism in Social Science."

motivation for the entire functionalist approach—or their function cannot be grounded in their selective history. In a bit more detail, the missing mechanism argument can be seen to rest on three pillars.

First, a genuine selective process requires variation.[48] However, it is not clear that actual social institutions, in fact, display this kind of variation. Instead, there is often only ever one version of an institution that was present. Hence, this institution cannot have been selected *from* a background population: there was no such background to select from.

Second, even in cases where there *was* the relevant kind of variation, this variation often does not appear to have greatly impacted the evolution of the relevant social institution. Instead, this evolution appears to have been heavily driven by chance alone.[49] A familiar example of this is the adoption of the "qwerty" keyboard, which, for largely fortuitous reasons, ended up the prevalent keyboard design despite its inherent *disadvantages* compared to rival designs.[50] Hence, there was no genuine *selection* of these institutions.

Third and finally, genuine selection requires reproduction.[51] However, social institutions generally differ only in their propensity to survive or grow (assuming the relevant variation even exists) but not in their propensity to have offspring social institutions.[52] The qwerty keyboard design did not give birth to a second generation of qwerty keyboard designs; rather, it simply persisted at the expense of rival designs. Hence, this is not a case of genuine selection.[53]

Now, it does need to be acknowledged that there are limits to the scope of this "missing mechanism argument." In particular, more recent analyses suggest

48 Godfrey-Smith, *Darwinian Populations and Natural Selection*; Brandon, *Adaptation and Environment.* See also Schulz, *Structure, Evidence, and Heuristic*.

49 Godfrey-Smith, *Darwinian Populations and Natural Selection*.

50 David, "Understanding the Economics of QWERTY"; Lewin, "The Market Process and the Economics of QWERTY."

51 Brandon, *Adaptation and Environment*; but see also Godfrey-Smith, *Darwinian Populations and Natural Selection*. Note, though, that the exact nature of the inheritance processes can differ across different cases. See Boyd and Richerson, *The Origin and Evolution of Cultures*; Godfrey-Smith, *Darwinian Populations and Natural Selection*; Sober, "Evolutionary Theory, Causal Completeness, and Theism." Note also that evolutionary processes do not need to involve replication in a narrow sense, but merely reproduction with some resemblance. See Godfrey-Smith, *Darwinian Populations and Natural Selection*; Sober, *Philosophy of Biology* and *The Nature of Selection*.

52 Hodgson and Knudsen use the labels "successor selection" and "subset selection" (derived from Price) for this distinction (*Darwin's Conjecture*, 94–104).

53 Brandon, *Adaptation and Environment*; Vrba, "What Is Species Selection?"; Schulz, *Structure, Evidence, and Heuristic*. But see also Godfrey-Smith, *Darwinian Populations and Natural Selection*.

that, for at least some social institutions, the needed biocultural selection processes may well have been present.[54] For example, some moral frameworks and political systems may have existed in different versions which competed for copying success in novel settings.[55]

However, this point ultimately does not greatly affect the strength of the "missing mechanism" argument. To be a truly compelling approach toward social analysis, functionalism needs to be widely applicable.[56] If there are only a handful of cases to which it can be usefully applied, social functionalism becomes a mere methodological footnote and will not allow for major progress in the social sciences. Therefore, even if it turns out that the historically focused version of social functionalism works in some cases, it remains true that it is not general enough. As a general approach toward the social sciences, it cannot do the kind of work we ask it to do.

Fortunately, more recent treatments of social functionalism are available that improve on the traditional historical account. One of the most influential of these is the account of Pettit.[57] According to the latter, functional ascription in the social sciences should not be seen to rest on an institution's *actual* biocultural selective history but on whether and why that institution would be *virtually* selected. More specifically, according to Pettit, a social institution *N* has function *F* if, in cases where the existence of *N were* threatened by some external factor, *N*'s having *F would* ensure that *N* continued to exist. How *N* actually evolved—whether its existence ever actually got threatened—is not relevant to its function. In this way, Pettit sees social functions as counterfactually grounded: what matters is how the institution would respond if its continued survival were called into question.[58]

This new form of social functionalism certainly has much to recommend it. By shifting the focus away from the actual (selective) history of a social institution, Pettit's account sidesteps all of the above problems concerning the absence of such a history for a large number of social institutions. On top of this, by focusing on what ensures that a given institution is buffered from threats to

54 See, e.g., Boyd and Richerson, *The Origin and Evolution of Cultures*; Henrich and McElreath, "Dual-Inheritance Theory"; Wilson and Gowdy, "Evolution as a General Theoretical Framework for Economics and Public Policy"; Heyes, *Cognitive Gadgets*.

55 Henrich, *The Secret of Our Success*; Nichols, *Sentimental Rules*; Kumar and Campbell, *A Better Ape*.

56 Pettit, "Functional Explanation and Virtual Selection"; Elster, *Ulysses and the Sirens*.

57 Pettit, "Functional Explanation and Virtual Selection." See also Merton, *Social Theory and Social Structure*.

58 Pettit, "Functional Explanation and Virtual Selection."

its existence, Pettit's virtual selectionist account fits well to the major motivation behind social functionalism. As Pettit notes,

> The tradition of thinking associated with the likes of Durkheim in the last century and Parsons in this is shot through with the desire to separate out the necessary and the reliable from the contingent and the ephemeral. The idea in every case is to look for the core features of a society and to distinguish them from the marginal and peripheral. Functionalist method is cast throughout the tradition as a means of providing "a basis—albeit an assumptive basis—for sorting out 'important' from unimportant social processes" (Turner and Maryanski [1979], p. 135).[59]

However, that said, the account of Pettit also faces several problems that prevent it from being fully compelling as it stands.

First, the truth-functional evaluation of counterfactuals is generally very difficult. Would *Y* happen if *X* were to happen? There is no clear method known for assessing these sorts of claims.[60] This is problematic as Pettit's account requires us to know which of the relevant counterfactuals are true. Assuming the NCAA allows young athletes to obtain a college education they could not otherwise afford, what would happen if the number of NCAA scholarships became severely restricted (e.g., due to falling revenue at NCAA games)? Would people still attempt to join the NCAA—and thus, would the NCAA persist—or would they seek other career paths? Would alternative institutions (such as expanded minor leagues) arise that have similar benefits? How do we know?

Note that the issue here is again not that we *never* know how to evaluate counterfactuals.[61] Rather, the point here is just that there are very *many* counterfactuals that we do not know how to evaluate. This matters, as it introduces a parallel problem to the "missing mechanism argument" for the historical versions of social functionalism: it makes Pettit's account too narrow to be useful. We would only rarely be able to say what the function of a social institution is. This does not make for a robust social scientific methodology.

Second, it is not clear which shocks a social institution needs to be protected from for it to have a given function. Requiring that an institution would be able to persist in the face of *all* shocks is too strong. If a new social institution—a professional second division sports league, say—appeared that also gave young athletes the funds and time to obtain a college education, it is not implausible that the NCAA might cease to exist. This, though, might not be

59 Pettit, "Functional Explanation and Virtual Selection," 300.

60 Stalnaker, "A Theory of Conditionals"; Lewis, *On the Plurality of Worlds*.

61 Fodor, *The Theory of Content*.

seen to speak against the NCAA having the function to help young athletes obtain a college education they could not otherwise afford; after all, it may be precisely *because* the professional second division sports league co-opts this feature that it can push the NCAA out of existence in this counterfactual scenario. However, what determines the limits of the counterfactual circumstances to be considered when determining the function of *N*? Every answer to this question seems arbitrary. In turn, this would make functional ascription in the social sciences arbitrary too, and thus violate another key motivation behind social functionalism.[62]

However, this does not mean that it is impossible to provide a compelling version of social functionalism. To do this, the function of *N* should be seen to be dependent on those features of *N*—if any—that increase the expected survival or reproductive success of *N* in its *current* sociocultural environment. [63] That is, the key idea of the account to be defended in what follows is that a social institution *N* has function *F* if it is *now* selected or sorted for *F*. More precisely:

> *Presentist Social Functionalism*: Feature *F* of social institution *N* is (part of) the function of *N* if *F* makes it more likely that *N* will survive or reproduce in the current sociocultural environment.

To put this slightly differently, unlike Pettit's account, functional ascription is seen to lie in actual, not virtual, selection pressures. However, unlike in the historical version of social functionalism (derived from biofunctional accounts like those developed by Millikan), the focus here is on which traits are *adaptive*, not which are *adaptations*.[64] To understand this better, consider the key features of the account in more detail.

First, Presentist Social Functionalism groups together genuine selection (i.e., the heritable differential reproduction of social institutions) and mere "sorting" (i.e., the differential growth or persistence of social institutions). This is useful since (as noted earlier) it is not generally plausible to see social institutions as reproducing, but it *is* plausible to see social institutions as growing or surviving at different rates. So when it comes to social functionalism, the focus *should* be on the latter kind of process (though, as noted earlier, the former

62 See also note 43 above.

63 Pettit at times hints at the importance of the current adaptive pressures on a given social institution ("Functional Explanation and Virtual Selection"). However, these hints are not developed as they are here.

64 For functionalism in the biological and cognitive sciences, see, e.g., Millikan, *Language, Thought, and Other Biological Categories* and *Varieties of Meaning*. For more on the adaptation/adaptive distinction, see Sterelny and Griffiths, *Sex and Death*; see also Nanay, "Teleosemantics without Etiology."

need not be ruled out *a priori* either). Hence, the fact that social functions can be grounded in either "sorting" or genuine selection is made explicit on the present account.

Second, according to Presentist Social Functionalism, saying that *N* has function *F* is making a claim about what is true about *N now*. It is not making a claim about why *N* came to have feature *F*. Like Pettit's account, though, this matches a key motivation behind social functionalism: to express what parts of society are its reliable, core parts.[65] Presentist Social Functionalism allows us to home in on those features of social institutions that make their survival or reproduction more likely—and thus are better able to identify the institutions that are the stable parts of society.

Third, according to Presentist Social Functionalism, the only counterfactual that matters to the evaluation of the function of a social institution *N* is this one: would *N*'s expected reproductive or persistence success decrease if it did not have *F*?[66] This is a much more restricted use of counterfactuals than what is found in Pettit's account. In particular, we do not need to assess whether *N* with *F* would continue to exist in all (relevant) possible worlds. In this way, Presentist Social Functionalism can sidestep the major problems that befall Pettit's account.

Fourth, since Presentist Social Functionalism does not use the history of a social institution to ground its function, it avoids the problems of the historically focused versions of social functionalism. On the one hand, Presentist Social Functionalism can allow the actual biocultural evolution of *N* to have been heavily influenced by chance. It just implies that *N*'s having *F increases* the expected survival or reproductive success of *N in the current environment*. It does not even require that *N*'s having *F* fully determines the survival or reproductive success of *N*: only that it is made *more likely*. On the other hand, the past existence of a population of varying institutions of the same type is not required here either. In fact, Presentist Social Functionalism does not even require the *current* existence of a population of different institutions of the same type. The question is just whether *N*'s having *F* increases its expected survival or reproductive success of institution relative to a (possibly) counterfactual version of *N* that lacks *F*.

Fifth and finally, the present account can still allow for malfunction. It is not like anything that *N* does is part of its function. Rather, only those features that contribute to its expected reproductive or survival success are part of this

65 See also Bigelow, "Functionalism in Social Science."

66 Indeed, precisely the latter is at the heart of Nanay's "Teleosemantics without Etiology" as well.

function. So, to see a social institution *N* (the NCAA) as having function *G* (to provide a space for the twenty-five members of the board of governors to get to know each other better and deepen their professional networks) might turn out to just be *wrong*; while *G* may indeed be a feature of *N*, unless *G* increases *N*'s expected reproductive or survival success (which is plausibly not the case when it comes to the NCAA), it is not its function. Indeed, *G* might *lower N*'s expected reproductive or survival success: networking among the members of the board of governors could make it harder for the NCAA to fulfill its true function (say, enabling college students to stay fit and healthy). If this is so, then if board of governors meetings are used for networking purposes rather than for finding ways to keep students fit and healthy (say), the NCAA is *malfunctioning*.

Now, it is important to note that in the background here—and of Presentist Social Functionalism in general—is the need for an individuation schema that determines what the relevant social institutions and their features are. Many things can impact the likelihood with which a social institution survives or reproduces, including the presence of other social institutions and various external features of the biosocial environment.[67] However, these do not necessarily become part of the function of a given social institution. Only if they are features of the institution could they be part of its function. This point also extends diachronically: it needs to be determined when a social institution remains the same social institution and when it becomes a new one. If institution *N* has feature *F* at time t_1 and a different feature *G* at time t_2, is it still the same institution or a new one (e.g., if a company that solely produced consumer technology at t_1 also starts to provide consumer lending services at t_2, does it become a bank)?

These, though, are familiar issues for all the relevant accounts of functional ascription (throughout the social, cognitive, and biological sciences)—and, indeed, the nature of evolution by natural selection in general.[68] Fortunately, for present purposes, it is not necessary to determine the right social institutional individuation schema; any reasonable approach can be used in conjunction with Presentist Social Functionalism.[69] That is to say, Presentist Social

67 For example, the appearance of the institution of fantasy football leagues can make the institution of the NFL more likely to spread and persist. I thank an anonymous referee for useful discussion of this issue.

68 Bertrand, "Proper Environment and the SEP Account of Biological Function"; Laland, et al., "On the Breadth and Significance of Niche Construction"; Odling-Smee, et al., *Niche Construction*; Dawkins, "Extended Phenotype—but Not Too Extended" and *The Extended Phenotype*; Griffiths, "Review of 'Niche Construction'"; Griffiths and Gray, "Developmental Systems and Evolutionary Explanation."

69 Bertrand, "Proper Environment and the SEP Account of Biological Function"; Griffiths and Gray, "Developmental Systems and Evolutionary Explanation."

Functionalism should be seen to be built on an existing theoretical foundation that individuates society into different institutions with various features.[70]

In this way, it becomes clear that functionalism is an important theoretical approach to the social sciences that can be given a compelling gloss: Presentist Social Functionalism. Importantly also, this gloss is nonnormative. It is not the case that the function of a social institution needs to be focused on a human good. Rather, anything that contributes to its expected survival, growth, or reproductive success can be part of this function. (Indeed, the fact that the function of a social institution is not tied to a human good is something that all of the major versions of social functionalism—historical, counterfactual, or presentist—have in common.) With this in mind, it is possible to return to the question of the characterization of the nature of IC.

3. INSTITUTIONAL CORRUPTION: A PRESENTIST SOCIAL FUNCTIONALIST ACCOUNT

With the presentist theory of social functionalism in the background, a novel characterization of IC can be developed that satisfies all of the desiderata laid out in section 2 and that has several further useful beneficial implications. In particular, given the plausibility of Presentist Social Functionalism, IC can be characterized as follows:

> *Institutional corruption*: The extent to which the actions of a set of agents prevent a social institution *N* from fulfilling its function *F*, where *F* is the set of features of *N* that increase *N*'s expected survival or reproductive success.

Several aspects of this characterization are important to note.

First, it is worthwhile making explicit how this characterization satisfies all of the desiderata laid out in section 1.

> It is *general*: The present characterization of IC applies to all kinds of social institutions and is not restricted to the context of representative democracy (in the US or more broadly). This is due to the fact that the characterization is teleological and sees IC as the thwarting of the purpose of a social institution. Hence, it applies to any social institution

70 Of course, as is standard in non-foundationalist sciences, Presentist Social Functionalism can also be used to help bootstrap such an individuation schema. The point here is just that such a schema is separate from Presentist Social Functionalism—though the latter also brings out the importance of determining such an individuation schema for the study of institutional corruption.

with a function—which includes the prison system, the press, as well as the NCAA, corporations, or even such social institutions as the Mafia (among many others).

It is *spelled out*: The present characterization of IC is based on a well-grounded theory of the function of social institutions. Indeed, this is one of the two reasons why the defense of Presentist Social Functionalism in the last section is important here. This defense ensures that the characterization of the functional ascription of social institutions underlying IC has a strong theoretical basis and is not left to the intuitions of the relevant researchers.

It is *nonnormative*: The present characterization of IC does not inherently see the purposes of social institutions as moral and, therefore, does not see IC as inherently normative. In this way, the present account of IC avoids the challenges faced by Miller's "Institutional Corruption" account: by making the ethical status of IC dependent on the details of the relevant institution, it can do justice to the ethical complexity of IC.

The fact that the above characterization of IC satisfies all of these desiderata further matters, as it shows that the notion of institutional purpose can be spelled out in a coherent manner and thus form the basis of a compelling account of IC. In this way, the present account can respond to some of the worries that have been levied against teleological accounts of this phenomenon more generally: namely, that its core notion—institutional purpose—cannot carry the weight it needs to.[71] As the defense of Presentist Social Functionalism makes clear, it *is* possible to provide a cogent grounding to the notion of institutional purpose and thus to use the latter as a foundation for a plausible account of IC.

This leads directly to the second important point to note about the present characterization of IC, which has also already been hinted at but deserves to be spelled out in more detail. This point concerns the fact that this characterization fits a general theoretical framework in the social sciences. This is the second major reason why the defense of Presentist Social Functionalism from the previous section is important here. Unlike the account of Miller—which is also spelled out in detail—the present account is not disconnected from functionalism in the social sciences more generally.[72] On the contrary, the present theory of IC is a natural extension of this general account of functionalism in the social sciences.

71 See, e.g., Rothstein and Varraich, *Making Sense of Corruption*.

72 See Miller, *Institutional Corruption*.

This not only gives this theory of IC a solid theoretical backing, but it also allows the easy extension of existing findings from the social sciences to the further investigation of IC. In particular, we do not need to establish the function of social institutions anew but can rely on the work already being done in the social sciences. For example, we can rely on whatever theory of the function of corporations ends up being the most plausible one (whether it is the shareholder-benefit one or the stakeholder-benefit focused one), and we do not need to derive this function from scratch in the context of the investigation of potential IC. This way, we may also find instances of IC that we would have otherwise overlooked (for example, concerning the IC of corporations).

The third point to emphasize about this characterization of IC is that it does not require that the cause of the corruption is a systematic, intentional, immoral, or illegal action. Institutions can get accidentally corrupted, and they can get corrupted for moral or legal reasons. On the present account, IC is like the corruption of (electronic) data. If a flash drive (or printed out spreadsheet) falls into a river, it is likely that the data on it will become unusable and functionless. This is so whether the flash drive (or printed out spreadsheet) was intentionally, legally, or morally—or not—thrown into the river, and whether or not the data on the drive (or table) were moral or legal in content.

This is thus another way in which the present account does justice to the complexity of IC: it may sometimes require censure, it may be ethically problematic but excusable, it may be ethically neutral, and it may even be ethically permissible or even required. In this way, the present account can bring out what is common to all cases of the undermining of institutions (including corrosion, rebellion, and accidental prevention of function) without being forced to morally evaluate all of them in the same way. In turn, this places the normative and moral considerations squarely where they can do the most good: in the details of the relevant case.

For example, if someone acted in ways that undermined the function of the Nazi Party, then that may have been morally required. Indeed, even if this undermining of the Nazi Party is the result of mere laziness on the part of the relevant agent, it is still IC, and it is still (*pro tanto*) morally good—though the person engaging in it need not deserve praise.[73] For the same reason, the source of the corruption need not be systematic: just one action—such as the distribution of fliers in front of the University of Munich—can (partly) undermine the Nazi Party and can thus count as IC.[74] (In a similar way, the data on a flash

73 Fricker, "What Is the Point of Blame?"; Friedman, "How to Blame People Responsibly."

74 For this reason, Sophie and Hans Scholl can be praised for corrupting the Nazi Party. (We can also praise someone for sabotaging—corrupting—a bomb so that it fails to go off and cause harm.)

drive or printout can be corrupted with one-off behaviors—throwing it into a river—as well as with systematic actions, such as the careless treatment of the drive or piece of paper that, over time, leads to it getting dirty and unreadable.)

It is important to emphasize that the generality of the present account is one of its features, not one of its bugs. Of course, it is possible to make finer distinctions and focus particularly on certain forms of IC—say, ones that are internally, intentionally, and systematically caused and that target immoral institutions.[75] However, this does not mean that there is not also value in providing a general account of the phenomenon. On the contrary, the generality of the present account is one of its key novel benefits.

In particular, by not using the sources and consequences of the undermining of an institution to characterize the nature of IC, it becomes possible to bring together what many superficially different social phenomena have in common.[76] For example, the Russia-based social media manipulation in the run-up to the 2016 US presidential election and the Trump administration's allegation of wide-scale voter fraud in the aftermath of the 2020 election differ in numerous particulars. The former is driven by sources external to US democratic institutions, the latter by sources internal to these institutions. The cases may also differ in intention and systematicity. However, there is also something important that is shared by these cases: they both (partially) prevented US democratic institutions from functioning properly, and they did so in similar ways—by increasing polarization and spreading propaganda.[77] This is theoretically valuable to bring out when studying democratic resiliency and the ways to improve it. For example, it suggests that similar responses may be useful in both cases, such as ensuring that the electorate is as well informed about the facts as possible. The fact that the present account of IC can bring out these communalities is thus one of its theoretical advantages.

Similarly, it is a major benefit of the present account of IC that it brings out clearly that societies with many instances of IC are less likely to be stable. These are societies many of whose institutions are made less likely to survive

75 As is done by, e.g., Ferretti, "A Taxonomy of Institutional Corruption"; Miller, *The Moral Foundations of Social Institutions*; Ceva and Ferretti, "Political Corruption, Individual Behaviour and the Quality of Institutions."

76 This is similar to other generalizing accounts. There are good reasons to often distinguish viral from bacterial diseases. However, there are also good reasons to often treat these subsets of the same overarching phenomenon: an infectious disease. This allows us to find common causes (e.g., the presence of other infected individuals) or common treatments (isolation, hydration, etc.).

77 In fact, this is shared with other cases, such as attempts to weaken the dictatorship in North Korea.

or reproduce. Importantly, this is so independently of whether the corruption is systematic, intentional, or moral. On the present account, people living in highly corrupt societies—whatever distinguishing details there may be between these societies—have in common the fact that they need to deal with highly unstable institutions (i.e., institutions that face major barriers to their survival and reproduction). This brings out a key common feature of highly corrupt societies that other accounts would miss: whatever the details of their causes, a conglomeration of IC leads to institutional instability.

Importantly also, this is not a trivial inference. Rather, the present approach ties IC to the prevention of an institution fulfilling its function (and not to, say, duplicitous violations of democratic egalitarian ideals) and then spells out the function of an institution to those of its features that give it a current biocultural selective advantage. In this way, the present account can *explain* why societies with much IC are less likely to be stable—this follows from the present characterization of IC. Furthermore, this is not something that is, at least on the face of it, the case for the characterizations of Thompson, Warren, Lessig, or Miller, which would not lead us to expect much IC to go with much social instability: undemocratic and immoral societies can be stable.[78]

Here, it is also noteworthy that not every crime or misdemeanor will count as an instance of IC on the present account. For example, ordinary theft need not block the function of an institution, and neither need all cases of nepotism: the stealing of a bike need not have any implications for the institution of private property to survive or persist.[79] The present theory thus provides a general, encompassing account of the phenomenon without being either trivial or forced to accept contentious moral or metaethical propositions, as is true of other theories in the literature, such as that of Miller.[80] The present account allows us to separate the analysis of the presence and consequences of IC from its causes and moral status. This gives us more degrees of freedom in tackling this phenomenon in a way that is both feasible and compelling.

The fourth point to note about the above characterization of IC is that the source of the corruption need not be an individual human being but can also be a collective agent, like a corporation or foreign government. In particular, the characterization recognizes that an institution can be prevented from fulfilling

78 Thompson, *Ethics in Congress*; Warren, "Political Corruption as Duplicitous Exclusion"; Miller, *Institutional Corruption*; Lessig, *America, Compromised*.

79 Miller, *Institutional Corruption*, 110–15. However, it is important to note that this will depend on the details of the case. If theft becomes sufficiently common, every additional theft could well make it harder for an institution of private property to persist. See also the discussion of graded IC below.

80 Miller, *Institutional Corruption*.

its function by the concerted effort of a number of human beings.[81] For example, if a social network eases the spread of political misinformation, this can prevent the public press from fulfilling its function.[82] Importantly, this is so even if no individual can be seen as the source of this IC: owners and employees of the social network may not have been responsible themselves for furthering the spread of the misinformation—and may even have attempted to block this spread. Indeed, no individual user need have had any kind of significant impact on this spread. However, with sufficient numbers of users and sources of misinformation, misinformation can spread far and quickly, merely as the result of the structure of the institution of the social network.[83]

In this way, the present account diverges from those presented, e.g., by Ceva and Ferretti: IC need not reduce to the corruption of an individual agent.[84] To begin with—and as noted earlier—the IC need not be immoral, and even where it is, it need not result from the actions of a morally culpable individual. More importantly, though, the corruption need not even be analyzable into the intentions, ends, and behaviors of individual humans, as is assumed by Miller.[85] Rather, it can be the upshot of a genuinely collective agent.[86] This matters, as it opens up a wider class of sources of IC and can thus help the study and prevention of the latter.[87] In particular, the present characterization does not need to get involved in debates about the plausibility of individualism in the social sciences but can work with whatever is the upshot of these debates.[88] This is especially important due to the fact—noted earlier—that there is good reason to think that the holism/individualism debate may call for a pluralist solution that allows for both individualism and holism to sometimes be the best approach to a given social scientific issue.[89] In this way, the present account's openness

81 This is a point also stressed by Miller in *Institutional Corruption*—though, as noted below, the latter is committed to spelling out this kind of collective agency in individualist terms. See also Vergara, *Systemic Corruption*.

82 See also Miller, *Institutional Corruption*, 304–9.

83 O'Connor and Weatherall, "Modeling How False Beliefs Spread."

84 Ferretti, "A Taxonomy of Institutional Corruption"; Ceva and Ferretti, "Political Corruption, Individual Behaviour and the Quality of Institutions"; Ferretti and Ceva, *Political Corruption*.

85 Miller, *Institutional Corruption*.

86 List and Pettit, "Group Agency and Supervenience."

87 See also Vergara, *Systemic Corruption*.

88 See, e.g., Elster, "The Case for Methodological Individualism"; Kincaid, "Open Empirical and Methodological Issues in the Individualism-Holism Debate"; Jones, "Methodological Individualism in Proper Perspective"; Epstein, *The Ant Trap* and "Why Macroeconomics Does Not Supervene on Microeconomics."

89 Ruiz and Schulz, "Microfoundations and Methodology."

to collective agency and social holism frees it from the constraints imposed by the individualistic commitments of Miller, Ceva, and Ferretti.[90]

This deepens a point that was mentioned in section 1 already. Without a doubt, there is much complexity in the debate surrounding the question of whether all cases of IC reduce to cases of individual corruption. The same is true for the debate as to whether instead of a teleological account of the phenomenon, a deontological one (say) should be provided. The present treatment cannot be seen to address (or even to attempt to address) all the issues here. However, the point to note is that the present, teleological and non-individualistic account has several key benefits. In particular, it is coherent, it fits well to research in the social sciences elsewhere, and it brings out novel social patterns (such as the greater likelihood of instability in countries with many social institutions whose purposes are undermined). In turn, this means that good reasons need to be provided for giving up these benefits. If an individualistic, moral, and non-teleological account of IC is to be shown to be superior, it would have to be made clear that it has benefits, the sum of which is greater than that of the present account.

The final point to note about the present account of IC is that it is the first one in the literature that explicitly makes IC a matter of degree. This is important, as actions can prevent *some*, but not all, aspects of the function of a given social institution, and they can merely make the fulfillment of that function *harder*. For example, if one particular postal worker, out of tiredness, delivers mail a little late one day, then while this technically is a form of IC, it is a very weak one: the function of the postal service is undermined, but only negligibly so. By contrast, if postal workers are being so overworked—e.g., because of employment cuts—that they *all always* deliver mail late, then this is a more serious case of IC: the function of the postal service is seriously undermined. Finally, if the postmaster general orders the employees *not* to deliver mail, then that would be a very strong case of IC: the function of the postal service is fully undermined.[91]

The present account can easily handle this. It allows for IC to occur on a bigger or smaller scale: the greater the corruption, the more functions of an institution are undermined, and the more strongly they are undermined. The present account thus provides the right kind of framework with which to handle the complexity of the phenomenon. There is no need to make a call

90 Miller, *Institutional Corruption*; Ferretti, "A Taxonomy of Institutional Corruption"; Ceva and Ferretti, "Political Corruption, Individual Behaviour and the Quality of Institutions"; Ferretti and Ceva, *Political Corruption*. Of course, this then raises a host of further questions concerning the ways in which collective agents can be morally responsible for their actions, etc. However, these questions can be left for a future occasion.

91 Note also that these cases span different sources—individual actors and collective actors—as well as different degrees of systematicity and culpability.

as to whether something definitely is or is not a case of IC; instead, we can allow something to be more or less of a case of IC. This is helpful, as existing accounts have tried to handle this fact by requiring genuine IC to be the result of actions that have the "tendency" to undermine the function of an institution.[92] This, though, then requires an account of what such a tendency consists of and when it exists. In turn, this is not easy to do and may be somewhat arbitrary. It is clearer to describe the phenomenon as it is: namely, as leading to more or less of an undermining of the function of the relevant social institution. This is exactly what the present account does.

An example may make this clearer. Consider FIFA. This association may have a number of functions, including growing the sport of football internationally, advocating for fair play, and ensuring it is accessible to everyone. It has, however, been alleged that various actions have led to some of these functions being undermined; for example, its ability to advocate for fair play may have been hindered by some of its officials taking bribes for sponsorship contracts or the awarding of tournaments.[93] However, others of its functions—such as its ability to grow football internationally—may not have been so undermined. In this case, FIFA can now more clearly be stated to be *partially* institutionally corrupted, rather than us having to decide whether the actions of FIFA officials have, or have not, *fully* corrupted the organization.[94]

All in all, therefore, the present theory of IC sees it as the outcome of actions that partly or fully prevent a social institution from fulling its function—i.e., which partially or fully negate those features of the institutions that increase its expected reproductive or survival success. This theory is theoretically well-grounded in a general account of social functionalism and has several further benefits—especially in doing justice to the inherent complexity of the phenomenon.[95]

4. CONCLUSION

The characterization of and response to IC has come to be recognized as a major task of the social sciences (broadly understood). In this paper, I advocate for a

92 See, e.g., Miller, *Institutional Corruption*; Thompson, *Ethics in Congress*.

93 See, e.g., Jennings, *Foul!*

94 Of course, these actions may also have been *individually* corrupt.

95 Another benefit of the account is that it allows for a novel take on US campaign finance laws: instead of just considering individual corruption as a limitation on free speech and campaign finance, it becomes possible to consider some forms of campaign finance as being limited due to their systemically corrupting character (e.g., of the voting process). Further analysis of this goes beyond the bounds of this paper, though. I thank an anonymous referee for useful discussion of this issue.

novel theory of this phenomenon. According to this theory, IC is the result of an individual or collective agent acting in ways that prevent a social institution from partially or fully fulfilling its function. In turn, the function of a social institution is spelled out in line with the currently most well-developed account of social functionalism in the literature: Presentist Social Functionalism. Presentist Social Functionalism sees the function of a social institution as those of its features that increase its expected reproductive or survival success in the current sociocultural environment.

This theory of IC is a useful addition to the literature. It is teleological and thus general, fully spelled out, and non-normative. In particular, it ties IC to the thwarting of the purpose of a social institution and provides a solid theoretical grounding to these purposes, but it does not require them to be based on normative considerations. In this way, it situates the study of IC in a wider functionalist approach toward the social sciences and does justice to the complexity of IC—both when it comes to its inherent nature and its moral evaluation.

University of Kansas
awschulz@ku.edu

REFERENCES

Amit, Elinor, Jonathan Koralnik, Ann-Christin Posten, Miriam Muethel, and Lawrence Lessig. "Institutional Corruption Revisited: Exploring Open Questions within the Institutional Corruption Literature." *Southern California Interdisciplinary Law Journal* 26, no. 23 (Summer 2017): 447–68.

Bertrand, Michael. "Proper Environment and the SEP Account of Biological Function." *Synthese* 190, no. 9 (June 2013): 1503–17.

Bigelow, John C. "Functionalism in Social Science." In *Routledge Encyclopedia of Philosophy*, edited by Edward Craig. Taylor and Francis, 1998. https://www.rep.routledge.com/articles/thematic/functionalism-in-social-science/v-1.

Boyd, Robert, and Peter J. Richerson. *The Origin and Evolution of Cultures*. Oxford: Oxford University Press, 2005.

Brandon, Robert. *Adaptation and Environment*. Princeton: Princeton University Press, 1990.

Brock, Gillian. "Institutional Integrity, Corruption, and Taxation." *Edmond J. Safra Working Papers* 39 (March 2014). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2408183.

Ceva, Emanuela, and Maria Paola Ferretti. "Political Corruption, Individual Behaviour and the Quality of Institutions." *Politics, Philosophy and Economics*

17, no. 2 (May 2018): 216–31.
Cummins, Robert. "Functional Analysis." *Journal of Philosophy* 72, no. 20 (November 1975): 741–65.
David, Paul. "Understanding the Economics of QWERTY: The Necessity of History." In *Economic History and the Modern Economist*, edited by William N. Parker, 30–49. London: Basil Blackwell, 1986.
Dawkins, Richard. *The Extended Phenotype: The Long Reach of the Gene*. Oxford: Oxford University Press, 1982.
———. "Extended Phenotype—but Not Too Extended: A Reply to Laland, Turner and Jablonka." *Biology and Philosophy* 19, no. 3 (June 2004): 377–96.
Della Porta, Donatella, and Alberto Vannucci. *The Hidden Order of Corruption: An Institutional Approach*. London: Routledge, 2012.
Durkheim, Émile. *The Elementary Forms of the Religious Life*. Translated by J. W. Swain. London: Allen and Unwin, 1915.
Elster, Jon. "The Case for Methodological Individualism." *Theory and Society* 11, no. 4 (July 1982): 453–82.
———. *Ulysses and the Sirens: Studies in Rationality and Irrationality*. Cambridge: Cambridge University Press, 1979.
Epstein, Brian. *The Ant Trap: Rebuilding the Foundations of the Social Sciences*. Oxford: Oxford University Press, 2015.
———. "Why Macroeconomics Does Not Supervene on Microeconomics." *Journal of Economic Methodology* 21, no. 1 (2014): 3–18.
Ferretti, Maria Paola. "A Taxonomy of Institutional Corruption." *Social Philosophy and Politics* 35, no. 2 (Winter 2018): 242–63.
Ferretti, Maria Paola, and Emanuela Ceva. *Political Corruption: The Internal Enemy of Public Institutions*. Oxford: Oxford University Press, 2021.
Fields, Gregg. "Parallel Problems: Applying Institutional Corruption Analysis of Congress to Big Pharma." *Journal of Law, Medicine and Ethics* 41, no. 3 (Fall 2013): 556–60.
Fodor, Jerry. *The Theory of Content and Other Essays*. Cambridge, MA: MIT Press, 1990.
Fricker, Miranda. "What's the Point of Blame? A Paradigm Based Explanation." *Noûs* 50, no. 1 (March 2016): 165–83.
Friedman, Marilyn. "How to Blame People Responsibly." *Journal of Value Inquiry* 47, no. 3 (June 2013): 271–84.
Garson, Justin. "Function, Selection, and Construction in the Brain." *Synthese* 189, no. 3 (December 2012): 451–81.
Godfrey-Smith, Peter. *Darwinian Populations and Natural Selection*. Oxford: Oxford University Press, 2009.
Griffiths, Paul E. "Review of 'Niche Construction." *Biology and Philosophy* 20,

no. 1 (January 2005): 11–20.
Griffiths, P. E., and R. D. Gray. "Developmental Systems and Evolutionary Explanation." *Journal of Philosophy* 91, no. 6 (June 1994): 277–304.
Henrich, Joseph. *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*. Princeton: Princeton University Press, 2015.
Henrich, Joseph, and Richard McElreath. "Dual Inheritance Theory: The Evolution of Human Cultural Capacities and Cultural Evolution." In *The Oxford Handbook of Evolutionary Psychology*, edited by Robin Dunbar and Louise Barrett, 555–70. Oxford: Oxford University Press, 2007.
Heyes, Cecilia. *Cognitive Gadgets: The Cultural Evolution of Thinking*. Cambridge, MA: Harvard University Press, 2018.
Hodgson, Geoffrey, and Thorbjørn Knudsen. *Darwin's Conjecture: The Search for General Principles of Social and Economic Evolution*. Chicago: University of Chicago Press, 2010.
Jennings, Andrew. *Foul! The Secret World of FIFA: Bribes, Vote Rigging and Ticket Scandals*. London: Harper Sport, 2006.
Jones, Todd. "Methodological Individualism in Proper Perspective." *Behavior and Philosophy* 24, no. 2 (Fall 1996): 119–28.
Joyce, Richard. *The Myth of Morality*. Cambridge: Cambridge University Press, 2001.
Kincaid, Harold. "Assessing Functional Explanations in the Social Sciences." *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1990, no. 1 (1990): 341–54.
———. "Open Empirical and Methodological Issues in the Individualism-Holism Debate." *Philosophy of Science* 82, no. 5 (December 2015): 1127–38.
Kumar, Victor, and Richmond Campbell. *A Better Ape: The Evolution of the Moral Mind and How It Made Us Human*. Oxford: Oxford University Press, 2022.
Laland, Kevin N., John Odling-Smee, and Marcus W. Feldman. "On the Breadth and Significance of Niche Construction: A Reply to Griffiths, Okasha and Sterelny." *Biology and Philosophy* 20, no. 1 (January 2005): 37–55.
Lessig, Lawrence. *America, Compromised*. Chicago: University of Chicago Press, 2018.
———. "'Institutional Corruption' Defined." *Journal of Law, Medicine and Ethics* 41, no. 3 (October 2013): 553–55.
Lewin, Peter. "The Market Process and the Economics of QWERTY: Two Views." *Review of Austrian Economics* 14, no. 1 (March 2001): 65–96.
Lewis, David. *On the Plurality of Worlds*. Oxford: Blackwell, 1986.
List, Christian, and Philip Pettit. "Group Agency and Supervenience." *Southern*

Journal of Philosophy 44, no. S1 (Spring 2006): 85–105.
Mackie, J. L. *Ethics: Inventing Right and Wrong*. Harmondsworth, Middlesex: Penguin, 1977.
Malinowski, Bronislaw. *Argonauts of the Western Pacific: An Account of Native Enterprise and Adventure in the Archipelagoes of Melanesian New Guinea*. London: Routledge, 1922.
Merton, Robert. *Social Theory and Social Structure*. New York: Free Press, 1968.
Miller, Seumas. *Institutional Corruption: A Study in Applied Philosophy*. Cambridge: Cambridge University Press, 2017.
———. *The Moral Foundations of Social Institutions: A Philosophical Study*. Cambridge: Cambridge University Press, 2009.
Millikan, Ruth Garrett. *Language, Thought, and Other Biological Categories: New Foundations for Realism*. Cambridge, MA: MIT Press, 1984.
———. *Varieties of Meaning: The 2002 Jean Nicod Lectures*. Cambridge, MA: MIT Press, 2004.
Nanay, Bence. "Teleosemantics without Etiology." *Philosophy of Science* 81, no. 5 (December 2014): 798–810.
Neander, Karen. "Content for Cognitive Science." In *Teleosemantics*, edited by Graham Macdonald and David Papineau, 167–94. Oxford: Oxford University Press, 2006.
Nichols, Shaun. *Sentimental Rules: On the Natural Foundations of Moral Judgment*. Oxford: Oxford University Press, 2004.
North, Douglas C. *Institutions, Institutional Change and Economic Performance*. Cambridge: Cambridge University Press, 1990.
Nye, Joseph S. "Corruption and Political Development: A Cost-Benefit Analysis." *American Political Science Review* 61, no. 2 (1967): 417–27.
O'Connor, Cailin, and James Owen Weatherall. "Modeling How False Beliefs Spread." In *The Routledge Handbook of Political Epistemology*, edited by Michael Hannon and Jeroen de Ridder, 203–13. London: Routledge, 2021.
Odling-Smee, F. John, Kevin N. Laland, and Marcus W. Feldman. *Niche Construction: The Neglected Process in Evolution*. Princeton: Princeton University Press, 2003.
Papineau, David. *Reality and Representation*. Oxford: Blackwell, 1987.
Papineau, David, and Justin Garson. "Teleosemantics, Selection and Novel Contents." *Biology and Philosophy* 34, no. 3 (June 2019).
Parsons, Talcott. *The Social System*. London: Routledge, 1951.
Pettit, Philip. "Functional Explanation and Virtual Selection." *British Journal for the Philosophy of Science* 47, no. 2 (June 1996): 291–302.
Philp, Mark. "Defining Political Corruption." *Political Studies* 45, no. 3 (August 1997): 436–62.

Rappaport, Roy A. *Ecology, Meaning, and Religion*. Richmond, CA: North Atlantic Books, 1979.

———. *Pigs for the Ancestors*. New Haven: Yale University Press, 1968.

Rosenberg, Alexander. *Philosophy of Social Science*. 4th ed. Boulder, CO: Westview Press, 2012.

Rothstein, Bo, and Aiysha Varraich. *Making Sense of Corruption*. Cambridge: Cambridge University Press, 2017.

Ruiz, Nadia, and Armin Schulz. "Microfoundations and Methodology: A Complexity-Based Reconceptualization of the Debate." *British Journal for the Philosophy of Science* 74, no. 2 (June 2023): 359–79.

Satz, Debra. "Markets, Privatization, and Corruption." *Social Research* 80, no. 4 (Winter 2013): 993–1008.

Schulz, Armin W. *Structure, Evidence, and Heuristic: Evolutionary Biology, Economics, and the Philosophy of Their Relationship*. New York: Routledge, 2020.

Sober, Elliott. "Evolutionary Theory, Causal Completeness, and Theism: The Case of "Guided" Mutation." In *Evolutionary Biology: Conceptual, Ethical, and Religious Issues*, edited by R. Paul Thompson and Denis Walsh, 31–44. Cambridge: Cambridge University Press, 2014.

———. *The Nature of Selection: Evolutionary Theory in Philosophical Focus*. Cambridge: Cambridge University Press, 1984.

———. *Philosophy of Biology*. 2nd ed. Boulder, CO: Westview Press, 2000.

Sparling, Robert Alan. *Political Corruption: The Underside of Civic Morality*. Philadelphia: University of Pennsylvania Press, 2019.

Stalnaker, Robert. "A Theory of Conditionals." In *Studies in Logical Theory*, edited by Nicholas Rescher, 98–112. Oxford: Basil Blackwell, 1968.

Sterelny, Kim, and Paul E. Griffiths. *Sex and Death: An Introduction to Philosophy of Biology*. Chicago: University of Chicago Press, 1999.

Street, Sharon. "A Darwinian Dilemma for Realist Theories of Value." *Philosophical Studies* 127, no. 1 (January 2006): 109–166.

Thompson, Dennis F. *Ethics in Congress: From Individual to Institutional Corruption*. Washington, DC: Brookings Institution, 1995.

———. "Theories of Institutional Corruption." *Annual Review of Political Science* 21 (2018): 495–513.

Tsfati, Yariv, H. G. Boomgaarden, J. Strömbäck, R. Vliegenthart, A. Damstra, and E. Lindgren. "Causes and Consequences of Mainstream Media Dissemination of Fake News: Literature Review and Synthesis." *Annals of the International Communication Association* 44, no. 2 (April 2020): 157–73.

Vergara, Camila. *Systemic Corruption: Constitutional Ideas for an Anti-Oligarchic Republic*. Princeton: Princeton University Press, 2020.

Vrba, E. "What Is Species Selection?" *Systematic Biology* 33, no. 3 (September

1984): 318–28.
Warren, Mark E. "The Meaning of Corruption in Democracies." In *The Routledge Handbook of Political Corruption*, edited by Paul M. Heywood, 42–55. London: Routledge, 2014.
———. "Political Corruption as Duplicitous Exclusion." *PS: Political Science and Politics* 39, no. 4 (October 2006): 803–7.
———. "What Does Corruption Mean in a Democracy?" *American Journal of Political Science* 48, no. 2 (April 2004): 328–43.
Wilson, David Sloan, and John M. Gowdy. "Evolution as a General Theoretical Framework for Economics and Public Policy." In "Evolution as a General Theoretical Framework for Economics and Public Policy," edited by David Sloan Wilson, John M. Gowdy, and J. Barkley Rosser. *Journal of Economic Behavior and Organization* 90, suppl. (June 2013): S3–S10.

Journal of Ethics and Social Philosophy
VOL. 25, NO. 3 · SEPTEMBER 2023
https://doi.org/10.26556/jesp.v25i3.2590

ATTRIBUTIONIST GROUP AGENT RESPONSIBILITY

Adam Piovarchy

Much work has been carried out showing that group agents exist as distinct entities not merely constituted by aggregating the set of individuals who make them up. Many philosophers believe we can also talk meaningfully about their possessing duties and carrying out actions. Paradigmatic examples of genuinely group agents include nation-states and corporations. But once we have established that group agents exist, that they can have duties, and that they can perform *bona fide* actions, an important question arises: Can group agents be *morally responsible* for violating said duties through their actions (or omissions)? Here, I am concerned with moral responsibility in a backward-looking sense, where an agent is responsible in virtue of what they have done independent of forward-looking considerations such as whether holding them responsible will produce good effects. To say that an agent is morally responsible for a wrong (right) action or omission is to say that they are an appropriate target of blame (praise) for that action or omission. As is standard, I will focus on blame for wrongdoing, given that the risks of incorrectly blaming are typically much higher than the risks of incorrectly praising. For now, as a first pass to help home in on our target phenomenon and to be ecumenical with respect to existing theories of blame, I will understand blame as a negative reactive attitude, which is generally unpleasant to be targeted with, and which communicates disapproval of the agent's conduct.

Intuitively, it seems like group agents can be blameworthy. We blame Volkswagen for its widespread intentional violation of emissions laws. We blame governments for failing to pass laws that reduce the damage caused by climate change. We call on such group agents to exhibit certain kinds of responses, such as apologizing and compensating victims, and we blame them even further if they do not. But demonstrating that group agents can be morally responsible requires that we spell out what features group agents must possess in order to be blameworthy or praiseworthy. This is standardly provided by taking group agents to be constituted by—and responsible in virtue of—certain well-ordered decision-making structures that are responsive to reasons. Since group

agents are the kinds of things that can appreciate moral reasons, and since they can control themselves in response to those reasons, they exhibit the kind of control that is emblematic of moral agency and which therefore makes them an appropriate target of blame and praise.

Though this standard line of argument gets many things right, parties to these debates may not be locating the blameworthy-making features of group agents in the right place. This becomes particularly salient when we notice that some group agents seem to lack the capacity to respond to certain kinds of considerations—and so cannot act on those considerations—and, rather than being excused, seem to be blameworthy precisely in virtue of this fact. The existence of such agents calls for a revised understanding of what it is that makes group agents responsible.

This paper proceeds as follows. First, I will outline a standard way of accounting for group agency and responsibility. Though particular accounts of group agency differ on the details, these will not be relevant for my argument. I will then present two objections. One is from Thompson, arguing that group agents cannot be responsible because they cannot take people as the objects of their attitudes.[1] The other is a new objection from myself: it is that group agents who consistently do the wrong thing due to their decision-making structure seem intuitively blameworthy, but current ways of understanding responsibility are committed to excusing such agents. I will then argue that avoiding these objections requires us to adopt an *attributionist* theory of group responsibility. On this account, group agents are responsible when their actions are attributable to them in such a way that reflects their evaluative judgments. Group agents are blameworthy when this evaluative judgment is objectionable, and importantly, evaluative judgments can be objectionable even if the agent lacks the ability to avoid wrongdoing or recognize some moral considerations.

1. GROUP AGENCY AND GROUP AGENT RESPONSIBILITY

The standard story for how there can be group agents takes a functionalist approach. To be an agent, one must be capable of having representational states and motivational states, and be capable of acting on the basis of these states. One must also meet some minimum standards of rationality, such as having a certain degree of consistency among one's beliefs and motivations. Group agents can possess these states and meet these standards by having certain well-ordered procedures or decision mechanisms in place, such as

1 Thompson, "The Moral Agency of Group Agents."

majority-rule voting.[2] The results of these procedures determine what the group agent "believes" or "decides." That there really is an agent existing over and above the decisions of individual group members is strongly evidenced by the fact that group agents can believe and decide things that none of the individual group members believe or decide.

If we grant that group agents exist and can perform actions, the next question regards which kinds of actions they are morally responsible for performing. In the most thorough treatment of this question to date, List and Pettit propose that group agents are morally responsible when the following criteria are met:[3]

> *Normative Significance*: The agent faces a normatively significant choice, involving the possibility of doing something good or bad, right or wrong.
>
> *Judgmental Capacity*: The agent has the understanding and access to evidence required for making normative judgments about the options.
>
> *Relevant Control*: The agent has the control required for choosing between the options.[4]

To give an example of how this works in practice, we can easily see that Volkswagen qualifies as responsible for its widespread violation of emissions laws. The decision to violate emissions laws is normatively significant because increasing emissions imposes nontrivial costs on others, which Volkswagen does not have a *prima facie* right to impose. Volkswagen has an understanding of the costs of increasing emissions and breaking the law, and access to evidence required for making normative judgments about its options. It also possessed control over its actions—it could have freely chosen to comply with the law, and it freely chose to violate the law, without any compulsion or coercion. Since it meets the above criteria, Volkswagen seems blameworthy for violating emissions laws.

2 List and Pettit, *Group Agency*; French, *Collective and Corporate Responsibility*.

3 List and Pettit, *Group Agency,* 155.

4 List and Pettit, *Group Agency,* 155. Similar sentiments are endorsed by Gilbert ("Collective Guilt"). Ways to adopt something like the standard story of agency with a different conception of responsibility can be found in Baddorf, "Phenomenal Conciousness" and Tanguay-Renaud, "To Fill or Not to Fill." Some philosophers widen the scope of group agency to include some kinds of collectives, such as Gilbert and Pilchman, "Belief, Acceptance, and What Happens in Groups"; Tuomela, Hakli, and Mäkelä, *Social Ontology in the Making*; and Tuomela, *Social Ontology*. An anonymous reviewer asks whether this paper's argument is relevant for such accounts. While I do not have the space to canvass the similarities and differences of these approaches to List and Pettit's, so long as these accounts allow that group agents can experience local structural deficits (explained below) in ways that do not undermine their agency altogether, and remain capable of expressing objectionable evaluative judgments, then the argument will apply to these accounts too.

2. PROBLEMS FOR THE STANDARD STORY

There are a number of ways to object to accounts of group agency, and this has generated a number of defenses in turn. For considerations of space, my focus will be limited to a select few; however, these particular objections are important. While other objections to group agency are typically responded to by finessing our account of group agency or identifying properties possessed by both individual agents and group agents, handling the objections raised in this paper instead requires that we reconsider the nature of moral responsibility and what it is that ultimately makes group agents blameworthy for wrongdoing.

2.1. Group Agents and Persons as Intentional Objects

The first objection is that group agents simply do not possess certain kinds of properties putatively necessary for moral responsibility. Thompson argues that group agents are unable to have certain kinds of emotions that Strawson takes to be essential to moral responsibility, namely, reactive attitudes such as guilt and resentment.[5] Thompson takes guilt and resentment to view the same wrong from different perspectives, with resentment being the second-personal perspective of the agent wronged and guilt being the first-person perspective of the wrongdoer. He believes that resentment *qua* blame has the function of bringing a perpetrator's moral understanding of their actions into alignment with the blamer's and the rest of the moral community by generating guilt and remorse.

Group agents are capable of functional equivalents of the epistemic and motivational components of guilt and remorse, in that group agents can have beliefs like "I have culpably violated a norm," and they can engage in apologies.[6] But Thompson argues that this is not enough. When we blame, we do not simply want the functional equivalents of guilt and remorse; it will not be sufficient for our target to go through the motions, acting *as if* they feel guilt and remorse. It is not enough that perpetrators simply believe they are blameworthy and desire to make amends. Rather, we want them to *care*, and this is something which group agents are unable to do.

Thompson's key argument is that moral emotions require certain intentional objects, and the objects of some reactive attitudes are *people*. Guilt that is not directed at one's self is not truly guilt, for instance.[7] The problem here is not

5 Thompson, "Moral Agency of Group Agents"; Strawson, "Freedom and Resentment."

6 Bjornsson and Hess, "Corporate Crocodile Tears?"

7 For alternative accounts of collective guilt, see Gilbert, *Joint Commitment*; cf. Ziv, "Collective Guilt Feeling Revisited." Hindriks also develops a noteworthy account of the moral emotions of group agents ("Collective Agency"), though for objections, see de Haan, "Collective Moral Agency and Self-Induced Moral Incapacity."

simply that group agents are not moral agents because moral agency requires phenomenal experiences, and group agents are incapable of these;[8] such an argument would beg the question against functionalist approaches to agency. Rather, the problem lies in group agents being able to only take propositions as their intentional objects. If group agents cannot take people as their intentional objects, they cannot care about people, and so they are not genuinely capable of experiencing guilt. Because they cannot genuinely care, they are psychologically abnormal, outside the bounds of our moral community, and therefore we can only respond with Strawson's "objective attitude" toward them.[9] They are a thing to be managed, rather than participants in our moral practices. Group agents seem more analogous to psychopaths, whom various philosophers take to be excused.[10]

2.2. Group Agents and Local Deficits

A second challenge to the responsibility of group agents concerns the existence of group agents who are intuitively blameworthy but whom the standard story will excuse. Though a lot of attention has been given to whether group agents possess the general capacity to deliberate, consider reasons, and act on the basis of those reasons, a problem which has not yet been considered concerns the possibility of what I shall call *local structural deficits* in decision-making capacity. Even if we grant that group agents can be morally responsible in general, there may be group agents who, due to the structure of their decision-making process, lack the ability to make certain kinds of decisions or act on certain kinds of reasons, and so cannot be responsible for failing to make certain kinds of decisions.[11]

Suppose, for instance, that when forming a company, group members design a decision-making structure that specifically precludes the group agent

8 Hindriks, "How Autonomous Are Collective Agents?"; Tollefsen, "Participant Reactive Attitudes."

9 Strawson, "Freedom and Resentment." Two anonymous reviewers helpfully suggest that we might avoid this objection by formulating caring in another way, such as valuing something and acting appropriately in light of that valuing. For instance, a corporation can value its employees and buy safety equipment for them to keep them safe, even if this reduces the company's long-term profits. Alternatively, we might think of caring as patterns of attention that cluster around certain kinds of issues in their deliberation.

10 Nelkin, "Psychopaths, Incorrigible Racists, and the Faces of Responsibility"; Watson, "The Trouble with Psychopaths." This summary is quite quick, and some aspects, as written, call for more clarification. For reasons that will become clear, I return to this argument and elaborate on the details below.

11 List and Pettit hint at such a possibility but think it is unlikely to happen in practice (*Group Agency*, 159).

from considering reasons that favor the environment or performing actions that would sacrifice profits for the environment. If we like, perhaps the group members' constitution has a clause stipulating that such reasons are simply inadmissible in the decision-making procedure, or that the group agent will self-destruct if such an action is performed, or that the group must now fulfill some kind of chain of steps that continues in an infinite recursion until the group members remove that reason from consideration. While we can certainly argue that the group members are blameworthy for this in virtue of their control over their actions and their awareness of those actions' consequences in creating the group agent, it is still tempting to think that the resulting group agent is itself also blameworthy.

Such structures would prevent group agents from meeting List and Pettit's criteria for many actions they perform. This could occur in two ways. The first is that group agents might, in virtue of certain local structural deficits, lack control over some kinds of actions and so cannot make certain choices. This would limit what options the group agent has available, and thus, according to List and Pettit's third criterion, the agent cannot be responsible for failing to avoid wrongdoing as no alternative was available. Alternatively, group agents with local structural deficits could fall afoul of List and Pettit's second criterion, in that its restricted options prevent the agent from forming certain kinds of normative judgments about other, unavailable options.[12] Even if one does not endorse List and Pettit's particular account, group agents would also fail to be responsible according to *control* accounts of moral responsibility, which hold that agents cannot be blameworthy for wrongdoing when they lack a certain kind of control over their actions, typically the capacity to avoid wrongdoing.[13]

The problem is that there appear to be group agents with local structural deficits that prevent them from making certain kinds of judgments or performing certain kinds of actions, and who yet, to many people, still seem blameworthy. For example, when a villager in the Amazon rainforest sees ACME Co. destroying the trees around them, polluting groundwater, and bribing officials to get away with this, it seems very appropriate for said villagers (and us) to blame ACME Co. even once they understand how this group agent truly lacks the ability to do otherwise. For many of us, finding out that the group agent has the kind of decision-making procedure that prevents it from reducing profits to

12 In a similar theme, Albertzart argues that group agents cannot be responsible because they do not truly have autonomy ("Monsters and Their Makers"). Though they can act freely, they cannot deliberate about which ends to adopt in the same way that human persons can.

13 See, e.g., Brink and Nelkin, "Fairness and the Architecture of Responsibility"; Fischer and Ravizza, *Responsibility and Control*.

save the environment does not quite seem like the kind of fact that now makes our blame inappropriate or unfair.[14] Such agents do not seem analogous to a company that, say, cannot avoid wrongdoing because the law prohibits it from doing so or because it lacks necessary resources. And yet, if we think that group agents need to possess a certain kind of control over their actions in order to be responsible, namely the ability to do the right thing for the right reasons, then we will have to accept that such agents are, in fact, excused whenever their wrong stems from such local structural deficits (or exempted from our responsibility practices altogether if those deficits are large enough).[15]

One might be tempted to try to explain away our intuition at this point. Perhaps our blame is misfiring and ought to be reserved for the group members for intentionally creating such an insensitive agent. Perhaps we are responding to the wrongness or badness of our imagined group agent's actions, which are not in question and which have been shown to affect people's judgments of culpability.[16] However, such options might have less pull on us when we are aware of an alternative account with which we can offer a principled justification for holding such group agents responsible while also preserving the intuition that agents who lack control over their actions are typically excused.

14 I appreciate that not all readers will share this intuition, particularly consistent control theorists. But it is a hard sell to argue that our blaming responses toward ACME Co. (and psychopaths) ought to be the same as they are toward Wonka Co. (and young children) who are more clearly excused of wrongdoing. These kinds of results at least motivate reconsidering the ultimate bases of blameworthiness. A common reply in cases like these is to argue that the group agent *could* have avoided their lack of capacity at some prior point, which we can "trace" their blameworthiness back to. De Haan, for instance, uses a tracing approach to argue that group agents can be blameworthy for "self-induced moral incapacity," such as gradually sliding into incapacity as a result of group members' corporate greed ("Collective Moral Agency and Self-Induced Moral Incapacity"). There are significant objections to such moves; see Shabo, "More Trouble with Tracing"; Smith, "Attitudes, Tracing, and Control"; Agule, "Resisting Tracing's Siren Song"; and Vargas, "The Trouble with Tracing." For one, they seem to get the phenomenology of our blaming wrong. The target of our blame simply is not failing to ensure that at some point in the future they will have the capacities required to avoid cutting down the rainforest (though this might be an additional source of blameworthiness). Relatedly, for many kinds of actions we want to trace culpability back to, it simply is not the case that that specific wrongdoing was reasonably foreseeable. And even if it was foreseeable, tracing explanations misrepresent the degree of blame we experience. Tracing seems to entail, e.g., that someone who takes heroin one time, knowing it has a risk of addiction, is thereby *fully* blameworthy for *all* wrongdoing that occurs as a result of their addiction since there was one decision said wrongdoings can be traced back to.

15 Wallace, *Responsibility and the Moral Sentiments*.

16 Knobe, "Intentional Action in Folk Psychology."

3. AN ATTRIBUTIONIST ACCOUNT OF GROUP RESPONSIBILITY

Let us reconsider for a moment the group agent that has local structural deficits. Perhaps the group agent is locally blind—it cannot even consider certain kinds of reasons—or perhaps it is locally constrained—it can recognize some considerations in some cases but cannot perform certain kinds of actions. The puzzle is that typically, lack of control or capacity undermines culpability: if I cannot swim, or I am too far away to help, or I am ignorant that anyone needs my help, it seems to be these facts that explain why I am not culpable for failing to save a drowning child. Likewise, if a company fails to give aid to others because it is prohibited by law or because it lacks the means to help them, it will be excused. But when ACME Co. poisons the water supply because its decision-making structure requires it to maximize profits, a plea to be considered excused seems much less convincing than that which would be offered if ACME Co. had no idea a chemical was poisonous or was forced to dump poison because the government mandated it. "Sorry I poisoned your water supply, but I was unable to consider the harm it would cause you as a reason to abstain from harming you" seems to have much less effect tempering our outrage. Indeed, it is tempting to think this blindness is precisely what we want to blame this agent for.

Thompson is not the only person to think group agents are analogous to psychopaths; this similarity has also been noted by Hindriks and Bakan.[17] Though it is true that some philosophers take psychopaths to be excused precisely because they lack the capacity to understand moral reasons, another line of argument is that psychopaths differ from other agents we take to be excused due to a lack of capacity to avoid wrongdoing. Although they cannot understand moral reasons, psychopaths are still capable of making assessments of what kinds of things are or are not reason giving. They are still agents capable of guiding themselves by what reasons they take to be present. And importantly, they are capable of understanding the effects that their actions have on others, e.g., that stabbing someone will cause a lot of pain, frustrate the victim's desire to go on living, and result in death. These features suggest a difference between psychopaths and other agents who lack the capacity to do the right thing. Though they cannot understand the concept of moral status, and so cannot form the attitude "your moral status is a reason to not harm you," they can form the attitude "the fact that this would cause you pain is *not* a reason to abstain from harming you."

According to attributionist theories of moral responsibility, while agents are *typically* excused for wrongdoing when they lack a certain kind of control

17 Hindriks, "How Autonomous Are Collective Agencies?"; Bakan, *The Corporation*.

over their actions or certain kinds of capacities, they are not excused *in virtue of* this lack of control or capacity.[18] Rather, what makes them excused is that many kinds of lack of control prevent an action from being attributable to the agent in the right kind of way. In particular, lack of control often prevents an action from expressing the agent's evaluative judgment about other agents.[19] When someone fails to give aid because they are tied up, cannot swim, or are unaware that someone needs aid, the failure to give aid does not reveal anything objectionable about the agent's take on the person in need.

Importantly, not all capacity deficits are like this. Talbert argues that when we reflect on the agency of psychopaths, it is evident they are different from other agents we typically take to be excused.[20] Given that the psychopath can understand all of the nonmoral facts and express judgments about those facts, this is sufficient to make blaming them appropriate. It is their *denial* of our importance that we find objectionable and which we want to blame them for. And this reasoning is just as appropriate in the case of group agents. When ACME Co. poisons my water supply to increase profits, it is not that they show me ill will in particular. I barely even enter into their deliberation. But it is precisely this *lack* of concern that I and others care about.

This makes them very different from Wonka Co., which poisons the river because it does not understand what effects its effluent has (and has no reason to think it ought to check). The latter's actions do not express any evaluative take on the merits of doing things that poison me, and this is why they are not blameworthy. But ACME Co. is aware that the poison will harm us, and their knowledge of this fact, combined with the decision to dump poison anyway, means that they do have the attitude that the harm to us is *not* a reason to avoid dumping poison. Such attitudes are precisely what make ACME Co. blameworthy.

Whereas most philosophers who think group agents can be blameworthy locate this blameworthiness in the group agent being able to perform actions, the attributionist approach locates it in the group agent being able to have evaluative judgments. This seems notable because, as highlighted earlier, a key

18 This is not simply referring to the attributionist "face" of responsibility, as an earlier reviewer thought, which is often adopted by control theorists and developed in more detail by Shoemaker (*Responsibility from the Margins*). Though Smith, "Responsibility as Answerability," emphasizes responsibility as answerability, she and Talbert, "Blame and Responsiveness," take an agent's evaluative judgments to ground blameworthiness.

19 Hieronymi, "The Force and Fairness of Blame"; Smith, "Moral Blame and Moral Protest." See also Scanlon, *What We Owe to Each Other*. Note, though, that his views have since changed (see Scanlon, *Moral Dimensions*).

20 Talbert, "Blame and Responsiveness to Moral Reasons."

argument in favor of thinking that group agents are genuine agents and not reducible to the aggregate of its group members is that group agents can have attitudes none of its group members themselves endorse. This is a strength of the account: even if Volkswagen's group members individually believe that violating emissions laws is immoral, and strongly preferred that this not take place, the group agent Volkswagen remains blameworthy, and this seems to be because of how that group agent evaluated the merits of its options. In particular, the contribution to climate change that its actions would produce (and which it knew, or should have known, they would produce) is the kind of thing that reveals an objectionable attitude toward those people who will be negatively affected by climate change.

This is not to say that the actions of group members are irrelevant for our assessments, however; an attributionist model of moral responsibility generates some new insights for thinking about the ways that group members' actions influence group agent blameworthiness. Consider the familiar idea that blameworthiness comes in degrees. One factor relevant here is the degree of wrongdoing, which control-based accounts of responsibility can accommodate. But another relevant dimension concerns accounting for the intuition that actions can be more or less blameworthy in virtue of how strongly they are endorsed or how attributable they are to the agent's evaluative orientation. We typically think that someone who experiences significant internal conflict and then commits wrongdoing is less blameworthy than someone who knowingly commits wrongdoing with enthusiasm (though conflict alone surely does not get one off the hook). Control accounts might try to explain this by invoking difficulty as a factor that is relevant to blameworthiness, and which is often present when agents' experience does not fully endorse their actions.[21] But an attributionist approach seems to do a better job of directly accommodating degree of endorsement by taking group agents to be more or less blameworthy in virtue of the extent to which the action was endorsed by its members. For example, it seems that in many cases, all else being equal, if a government's immoral decision is the result of 100 percent of voters voting in favor, this is more blameworthy than an otherwise equivalent decision produced by only 51 percent of voters. The latter action is less attributable to the group agent, even if it remains sufficiently attributable to make the agent blameworthy.

However, the relationship between degree of support among group members and degree of blameworthiness is not always simple. Readers can no doubt recall various instances in which governments failed to act in ways that voters supported, but which seemed to make said governments *more* blameworthy,

21 Nelkin, "Difficulty and Degrees of Moral Praiseworthiness and Blameworthiness."

rather than less. This initially seems to be a problem for my above observation. But group agent theorists emphasize that the attitudes and decisions of group agents are determined not only by group members' votes, but also by how the votes are aggregated.[22] Group agents, which can easily act in ways that most group members strongly reject, are agents in which members have less control over the relevant attitudes and thus are more poorly designed. Poor design does not exculpate group agents, but it does allow us to see how acting in ways with lower support among group members can reduce blameworthiness in some circumstances and increase it in others. In cases where members' votes seem to have insufficient influence on group agent attitudes and actions, owing to a suboptimal aggregation or decision-making procedure, the failure to influence attitudes suggests the group agent has, in their agency, something like a *bias* toward forming certain judgments, and this is what explains our blaming.[23] That the group agent regularly forms certain attitudes or commits wrongdoing *despite* the group members' votes can show that the decision-making procedure, which is a stable part of the group agent's makeup, is having an outsized criticizable effect, and this is what explains our tendency to increase blame.

There is one last feature which may be affecting our intuitions that is worth identifying, and this concerns the level of stability in decision-making that the decision-making procedure allows. Group members might continually fail to have sufficient influence on the group agent's attitudes and decisions, but those attitudes and decisions might not manifest something like a bias because the resulting attitudes and decisions are too inconsistent, unstable, or haphazard. In short, we might have discordance that suggests the agent is less responsive to reasons altogether.[24] While it is common to talk of "being an agent" as if it were a threshold notion, agency, in fact, comes in degrees, evidenced by there being no clear point between birth and adulthood in which one becomes a

22 First-past-the-post voting, for instance, tends to produce different results than those of mixed-member proportional representation, even if each group member's vote remains the same under both systems.

23 Here I do not mean something analogous to implicit bias, which has received a lot of attention and is commonly taken to be defined by the way that it *conflicts* with the agent's explicit attitudes and has little impact on agential decision-making. I mean bias in the traditional sense in which someone, e.g., continually favors their own group despite there being no adequate justification for this, or selectively interprets evidence and misrepresents challenges to their view due to motivated reasoning.

24 A subtle point: there is a sense in which the incorrigible racist is not responsive to reasons in that he will not change when we argue with him. But what attributionists take to matter is that the agent's attitudes are responsive to what the agent takes to be the case, rather than whether their attitudes accurately reflect the reasons they, in fact, have (Smith, "Attitudes, Tracing, and Control," 125–26).

morally responsible agent. Likewise, a group agent whose decision-making procedure leads to decisions that are too haphazard seems to be less of an agent and could be more analogous to a young child or someone with certain mental disabilities. To be sure, agents need not be perfectly consistent; Smith emphasizes that someone can fear a spider while also sincerely claiming to believe the spider is perfectly safe, and both of these attitudes (along with the charge of irrationally holding inconsistent attitudes) will be attributable to the agent without impugning her status as an agent.[25] But too much inconsistency can put one out of the agency and responsibility game altogether.

4. Group Agents and Justifying Blame

Now that we have considered how an attributionist account of responsibility avoids the local structural deficits objection, we can also see how it avoids the inability to care objection. The fact that group agents cannot form attitudes that have people as their object is no barrier to blaming them. What concerns us is their take on what things are or are not reason giving, and this is something they are able to do.[26]

Admittedly, this result relies on adopting a different account of blame and blameworthiness. This may be unsatisfying for some readers who interpret Thompson's argument as a conditional (*if* Strawsonian accounts of moral responsibility are correct, group agents cannot be blameworthy) and take attributionists to be simply rejecting the premise. We would have more reason to support the attributionist story if there were independent objections to Thompson's argument, or reasons to not grant the premise. These can be provided by examining some background considerations regarding Strawsonian accounts of moral responsibility that Thompson is somewhat unclear on, which any future treatments on the responsibility of group agents should be sensitive to.

On Strawsonian accounts, our reactive attitudes are responses to others' quality of will or level of regard. But most philosophers take the reactive attitudes to respond to, and thus occur *downstream* of, the blameworthy-making features of agents (though the reactive attitudes can be good evidence of blameworthiness). On this *response-independent* interpretation of blameworthiness, showing that group agents cannot experience resentment, guilt, or indignation does not yet show that blame is inapt, as these are things that occur after someone has displayed poor quality of will. That we cannot create guilt in them now

25 Smith, "Attributability, Answerability, and Accountability," 579.

26 Note this account also avoids worries that without consciousness, group agents are analogous to zombies (O'Madagain, "Group Agents").

(meaning blame might be ineffective or pointless) will not affect whether they are blame*worthy*. Instead, Thompson would need to show that moral emotions are essential to displaying blameworthy quality of will (or essential to some other factor that grounds blameworthiness) and that group agents cannot experience these emotions.

Taking reactive attitudes to be downstream of blameworthy-making features of agents is distinct from the position most commonly attributed to Strawson, known as a *response-dependent* conception of responsibility.[27] On this account, the reactive attitudes are crucial to understanding responsibility because there is no external justification for holding people responsible beyond the fact that ill will or lack of regard is what properly trained human reactive attitudes respond to.[28] This seems to be closer to what Thompson has in mind. On this account, the reactive attitudes are not just an inseparable part of holding agents responsible; they are also constitutive of those agents *being* responsible. It is our proneness to experiencing the reactive attitudes to an agent's quality of will that makes the agent responsible.

Holding responsible is typically thought to be linked to being responsible via the demands our blame expresses.[29] Thompson, however, links them via blame's purported function. He thinks that blame's function is to produce moral alignment of the wrongdoer and victim's understanding of the wrong. In particular, it aims to produce guilt, which is linked to resentment (blame) because guilt and resentment view the same wrong from different perspectives. Importantly, he takes guilt to involve caring, and caring to require affective attitudes, evidenced by the fact that our blame does not cease if wrongdoers behave as if they feel guilt.

With this background clarified, we are now in a position to note some costs to the overall argument. One is that a response-dependent account leaves Thompson in something of a minority position; many philosophers and folk alike take our blame to be responding to facts about wrongdoers that are independent of our actual (properly trained) blaming practices which ground blameworthiness. For example, when asking what *makes* him blameworthy, *why* that feature makes him blameworthy, and whether he is *really* blameworthy,

27 For a review of the trickiness of articulating his exact position, see Todd, "Strawson, Moral Responsibility, and the 'Order of Explanation.'"

28 Shoemaker, "Response-Dependent Responsibility."

29 As Strawson puts it: "the making of the demand *is* the proneness to such attitudes" ("Responsibility as Answerability," 207). For more on this point, see also Darwall, *The Second-Person Standpoint*. Cf. McKenna, who argues that ability to make demands is intertwined with the ability to respect demands because our blaming interactions are similar to conversations that require understanding meaning (*Conversation and Responsibility*).

answering with an appeal to the mere fact that we have tendencies to blame agents like him seems unsatisfying.[30]

More substantive costs concern the proposal that blame's function is to produce moral alignment by making our target care, which is identified by reflecting on our practices of blaming and what its point is.[31] As Thompson points out, "We do not merely accept a form of behaviorism. We do not demand that others merely act *as if* they experience guilt for their actions."[32] But there are three problems here. The first is that if we are appealing to our actual practices to determine the proper objects of our blame, it seems that many people *do* blame group agents. Some readers who are skeptics about the responsibility of group agents may not personally have the intuition that such agents are responsible, but it is undeniable that many members of our moral community routinely blame group agents in a manner identical to the way they blame individual group agents. Claims about what *our* moral practices are require some explanation for why, though blame toward individual agents has the aim of making them care, seemingly identical blame toward group agents either does not have this function or can only be interpreted as a mistake.

Relatedly, the observation that we do not accept mere behaviorism seems difficult to square with the observation that many people's blame toward group agents does subside in response to how the group agent responds to its earlier wrongdoing. Group agents are capable of offering apologies, attempting to repair relationships, and signaling that their failures will not happen again. And, indeed, failures to do these things typically generate even more blame from us.

Finally, observing we want something for our blame to cease does not show that thing is what makes our blame appropriately begin. Such reasoning seems

30 That this is not how most philosophers writing on the responsibility of group agents would think about moral responsibility is evidenced by how the literature as a whole has progressed. As noted earlier, philosophers have been concerned with investigating whether group agents possess the same kind of relevant features that justify considering them *bona fide* agents, who perform actions, and whom we can *justify* blaming and holding responsible, or whether there are relevant differences between group agents and ordinary agents (e.g., ability to care). But for the response-dependent approach, this framing may be wrongheaded as there is no independent justification for our reactive attitudes. If we have granted that group agents are agents, the facts about which agents are responsible is determined by whatever (properly developed and informed) human sensibilities deem to be fitting targets of blame. And this, it seems, is to be determined by looking inward to our actual moral practices and the fittingness conditions of our emotions, rather than considering principled arguments regarding functionalism, agency and intentional objects.

31 It should also be noted that appeals to the benefits of aligning understanding (Thompson, "The Moral Agency of Group Agents," 524) risk appealing to an external source of justification.

32 Thompson, "The Moral Agency of Group Agents," 526.

to imply that if someone committed wrongdoing and felt guilt immediately after doing it (perhaps they even felt guilty in the lead-up to the wrong, realizing and caring about how the wrong would affect us), then our blame would be pointless. But if our blame is pointless, then it seems like our blame ought not even begin, meaning the agent *would not be blameworthy*. But that is clearly not the case; agents who commit wrong while experiencing considerable guilt, and a proper understanding of their actions, can nevertheless be appropriate targets of blame.

5. Moral Alignment and Protest

The previous section's objections concerned the way in which we use our moral practices to determine who counts as blameworthy. But even if we set aside our concerns with justification and blameworthiness, taking blame's function to be moral alignment seems to also generate a few notable discrepancies with our blaming practices. In particular, we do not think our blame is inapt or unjustified even if we know it will be ineffective, and there are a variety of ways that blame aimed at producing alignment can be rendered ineffective.[33] Perhaps our target is simply incorrigible. It is also difficult to see how any alignment could be achieved when we blame historical figures or the dead, or when no one is around to see our blame. Another difficulty is that blame seems to achieve little when our target already feels guilt and so seems to have already acquired the same understanding as us. And sometimes we blame expressly without the aim of producing any alignment. We often feel outrage and are not interested in what our target has to say in response or how they think about what they did. Sometimes it is apt to storm off and not talk to the wrongdoer. Sometimes it is apt to continue blaming after they feel guilty. And even if group agents did somehow achieve consciousness and could thereby take persons as intentional objects or feel guilt, it does not seem like the character of our blame would dramatically change; alignment still would not be the goal. Just like when we blame politicians on the television, most of the time we do not expect our blame toward corporations or governments to have much effect on the target. We do not expect any response from them to us *qua* individuals at all, and we do not always make an effort to make sure said agents notice our blame. If moral alignment is the goal, blame is often an inefficient way to do it.

One can try to account for these discrepancies by arguing that we are taking a paradigm-based approach to our theorizing, or argue that blame is a speech act and speech acts can be unexpressed, or try to locate the alignment in

33 Tierney, "Guilty Confessions."

someone other than the target of our blame, as Fricker does.[34] But at a certain point, such discrepancies look more like counterexamples and seem against the spirit of the Strawsonian emphasis on our actual moral practices. These observations at least motivate looking for an alternative.

An attributionist account of blame is much better placed to make sense of these aspects of our blaming tendencies. Rather than taking blame's function to be producing moral alignment between wrongdoer and blamer, attributionists instead take blame to be a form of *protest*.[35] Our blame's focus is on the wrongdoing and the threat that the attitude expressed in the wrongdoing poses to our moral standing. In blaming, we are standing up for ourselves or another victim. Although protest is communicative, it does not only communicate to the wrongdoer. It also expresses outrage to other members of the moral community (which may include group members of the group agent) who have a role in maintaining moral standards. Unlike the moral alignment account, which seems to make blame pointless once alignment is achieved, communication more clearly does not merely function to transmit information. Many forms of communication have an expressive point, such as telling a spouse you love them even when this was never in doubt. Protest also makes better sense of our goal in blaming. The moral alignment account had trouble accounting for the fact that many instances of blame seem to not explicitly aim at alignment, such as storming off, ceasing interaction altogether, or blaming when alignment has been achieved. But the protest account does much better because protest denounces or repudiates the attitude that was expressed in the group agent's wrong. When I storm off in response to your wrongdoing, I am protesting against the attitude that was expressed, whether or not you, in particular, get the message and whether or not you already perfectly understand how the wrong affected me. This, in turn, allows us to make sense of how blame that is directed at a group agent but which is not performed in a way likely to evoke a response (e.g., because they are a foreign government, and you are just in a university classroom) has not thereby misfired. Protest against something that is not the case, however, does seem pointless, and this matches the thought that blaming someone who has not culpably done anything wrong is inappropriate.

Even if one accepts that attributionist accounts of moral responsibility and blame can overcome the objections raised against the possibility of group agent moral responsibility, one might object to attributionist accounts on independent grounds. One might argue that attributionist accounts are implausible,

34 Fricker, "What's the Point of Blame?"

35 Hieronymi, "Articulating an Uncompromising Forgiveness"; Smith, "Moral Blame and Moral Protest"; Talbert, "Moral Competence."

citing various factors often used to support control-based accounts of moral responsibility.

While I cannot resolve the debate between control theorists and attributionists here, it is worth noting that traditional objections to attributionist accounts of moral responsibility are much less persuasive when it comes to considering the responsibility of group agents. One of the main fault lines in the debate over moral responsibility concerns what ultimately justifies blaming culpable wrongdoers.[36] Control theorists typically argue that the ability to avoid wrongdoing is needed because blame is a negative treatment or sanction which is against one's interests. For blame to be apt, it must be *deserved*, and people do not deserve things which they could not avoid.[37] This is partly what grounds the thought that psychopaths are not blameworthy: since they lacked the capacity to understand moral reasons, they could not choose to act on those reasons and so cannot be blameworthy for failing to do so.

However, a concern for desert is much less pressing when considering the responsibility of group agents. There is strong support for the idea that group agents do not merit as much consideration as ordinary agents, and List and Pettit believe we should not extend the same rights to group agents that we give to individuals, such as the right to vote.[38] Additionally, some control theorists argue that part of what makes blame unpleasant and deserved is that it induces guilt, understood here as a *pained* recognition of what wrongs one has done.[39] Such arguments also cannot be used to justify blaming group agents because their lack of phenomenal consciousness means they are unable to experience pain. And given that pain seems to be bad in virtue of its phenomenal quality, a functional analogue of pain is unlikely to be a sufficient alternative for our theory.

If we are already much less concerned with the interests of group agents, and our ordinary reason to be careful with our blame—that it induces pained recognition of wrongdoing—simply does not apply, then arguments that we should favor a control requirement on blameworthiness for group agents over a protest-based account of blame (and an attributionist account of blameworthiness)

36 Rudy-Hiller, "It's (Almost) All about Desert."

37 Levy, *Hard Luck*; Nelkin, "Desert, Fairness, and Resentment." Group agents can "deserve" things like prizes in virtue of meeting the conditions stipulated in contests, but this is not the same as *basic* desert, which is what most control theorists take to be at issue (Pereboom, *Free Will, Agency, and Meaning in Life*). For some reflections on how attributionists justify blame, see Hieronymi, "I'll Bet You Think This Blame Is about You."

38 Hindriks, "How Autonomous Are Collective Agents?"; List and Pettit, *Group Agency*, 180–82.

39 Carlsson, "Blameworthiness as Deserved Guilt."

are much harder to get off the ground. It is perhaps even possible that one could be a control theorist about individual moral responsibility while being an attributionist about group agent moral responsibility, but I will set aside the possibility of defending such a position.[40]

6. CONCLUSION

This paper has argued that an attributionist account of moral responsibility is well-suited to make sense of our practices of blaming group agents and holding them morally responsible. Even though group agents cannot experience guilt, cannot feel pain, and can sometimes lack the ability to avoid wrongdoing, these factors are not barriers to them being blameworthy. This is because group agents are the kinds of things which, in virtue of their reasons-responsive decision-making structure, are able to make assessments about what kinds of actions are worthwhile and, importantly, what kinds of considerations are *not* reason giving. When their actions reflect attitudes or assessments that are objectionable, group agents are blameworthy, and our blame toward them is warranted.

University of Notre Dame, Australia
adam.piovarchy@nd.edu.au

REFERENCES

Agule, Craig. "Resisting Tracing's Siren Song." *Journal of Ethics and Social Philosophy* 10, no. 1 (January 2016): 1–24.

Albertzart, Maike. "Monsters and Their Makers: Group Agency without Moral Agency." In *Reflections on Ethics and Responsibility: Essays in Honor of Peter A. French,* edited by Zachary J. Goldberg, 21–35. Cham, Switzerland: Springer, 2017.

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

Bakan, Joel. *The Corporation: The Pathological Pursuit of Profit and Power*. New York: Free Press, 2004.

Björnsson, Gunnar, and Kendy Hess. "Corporate Crocodile Tears? On the

40 Elsewhere I have endorsed control accounts of responsibility; these arguments should be considered independent.

Reactive Attitudes of Corporate Agents." *Philosophy and Phenomenological Research* 94, no. 2 (March 2017): 273–98.

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, 284–314. Oxford: Oxford University Press, 2013.

Carlsson, Andreas Brekke. "Blameworthiness as Deserved Guilt." *Journal of Ethics* 21, no. 1 (March 2017): 89–115.

Darwall, Stephen L. *The Second-Person Standpoint: Morality, Respect, and Accountability*. Cambridge, MA: Harvard University Press, 2006.

de Haan, Niels. "Collective Moral Agency and Self-Induced Moral Incapacity." *Philosophical Explorations* 26, no. 1 (2023): 1–22.

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

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

Fricker, Miranda. "What's the Point of Blame? A Paradigm Based Explanation." *Noûs* 50 no. 1 (March 2016): 165–83.

Gilbert, Margaret. "Collective Guilt and Collective Guilt Feelings." *Journal of Ethics* 6, no. 2 (June 2002): 115–43.

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

Gilbert, Margaret and Daniel Pilchman. "Belief, Acceptance, and What Happens in Groups: Some Methodological Considerations." In *Essays in Collective Epistemology*, edited by Jennifer Lackey, 189–212. Oxford: Oxford University Press, 2014.

Hieronymi, Pamela. "Articulating an Uncompromising Forgiveness." *Philosophy and Phenomenological Research* 62, no. 3 (May 2001): 529–55.

———. "The Force and Fairness of Blame." *Philosophical Perspectives* 18, no. 1 (December 2004): 115–48.

———. "I'll Bet You Think This Blame Is about You." In *Oxford Studies in Agency and Responsibility*, vol. 5, edited by D. Justin Coates and Neal A. Tognazzini, 60–87. Oxford: Oxford University Press, 2019.

Hindriks, Frank. "Collective Agency: Moral and Amoral." *Dialectica* 72, no. 1 (March 2018): 3–23.

———. "How Autonomous Are Collective Agents? Corporate Rights and Normative Individualism." *Erkenntnis* 79, no. S9 (October 2014): 1565–85.

Knobe, Joshua. "Intentional Action in Folk Psychology: An Experimental Investigation." *Philosophical Psychology* 16, no. 2 (2003): 309–24.

Levy, Neil. *Hard Luck: How Luck Undermines Free Will and Moral Responsibility*. New York: Oxford University Press, 2011.

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

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

Nelkin, Dana K. "Desert, Fairness, and Resentment." *Philosophical Explorations* 16, no. 2 (2013): 117–32.

———. "Difficulty and Degrees of Moral Praiseworthiness and Blameworthiness." *Nous* 50, no. 2 (June 2016): 356–78.

———. "Psychopaths, Incorrigible Racists, and the Faces of Responsibility." *Ethics* 125, no. 2 (January 2015): 357–90.

O'Madagain, Cathal. "Group Agents: Persons, Mobs, or Zombies?" *International Journal of Philosophical Studies* 20, no. 2 (2012): 271–87.

Pereboom, Derek. *Free Will, Agency, and Meaning in Life*. New York: Oxford University Press, 2014.

Rudy-Hiller, Fernando. "It's (Almost) All about Desert: On the Source of Disagreements in Responsibility Studies." *Southern Journal of Philosophy* 59, no. 3 (September 2021): 386–404.

Scanlon, T. M. *Moral Dimensions: Permissibility, Meaning, Blame*. Cambridge, MA: Belknap Press of Harvard University Press, 2008.

———. *What We Owe to Each Other*. Cambridge, MA: Belknap Press of Harvard University Press, 2000.

Shabo, Seth. "More Trouble with Tracing." *Erkenntnis* 80, no. 5 (October 2015): 987–1011.

Shoemaker, David. "Response-Dependent Responsibility; or, A Funny Thing Happened on the Way to Blame." *Philosophical Review* 126, no. 4 (October 2017): 481–527.

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

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

———. "Attributability, Answerability, and Accountability: In Defense of a Unified Account." *Ethics* 122, no. 3 (April 2012): 575–89.

———. "Moral Blame and Moral Protest." In *Blame: Its Nature and Norm*, edited by D. Justin Coates and Neal A. Tognazzini, 27–48. New York: Oxford University Press, 2013.

———. "Responsibility as Answerability." *Inquiry* 58, no. 2 (2015): 99–126.

Strawson, Peter. "Freedom and Resentment." *Proceedings of the British Academy* 48 (1962): 187–211.

Talbert, Matthew. "Blame and Responsiveness to Moral Reasons: Are Psychopaths Blameworthy?" *Pacific Philosophical Quarterly* 89, no. 4 (December 2008): 516–35.

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

Tanguay-Renaud, Francois. "To Fill or Not to Fill Individual Responsibility Gaps?" In *The Legacy of Ronald Dworkin*, edited by Wil Waluchow and Stefan Sciaraffa, 71–98. Oxford: Oxford University Press, 2016.

Thompson, Christopher. "The Moral Agency of Group Agents." *Erkenntnis* 83, no. 3 (June 2018): 517–38.

Tierney, Hannah. "Guilty Confessions." In *Oxford Studies in Agency and Responsibility*, vol. 7, edited by David Shoemaker, 182–204. Oxford: Oxford University Press, 2021.

Todd, Patrick. "Strawson, Moral Responsibility, and the 'Order of Explanation': An Intervention." *Ethics* 127, no. 1 (October 2016): 208–40.

Tollefsen, Deborah Perron. "Participant Reactive Attitudes and Collective Responsibility." *Philosophical Explorations* 6, no. 3 (2003): 218–34.

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

Tuomela, Raimo, Raul Hakli, and Pekka Mäkelä, eds. *Social Ontology in the Making*. Boston: De Gruyter, 2020.

Vargas, Manuel "The Trouble with Tracing." *Midwest Studies in Philosophy* 29, no. 1 (September 2005): 269–91.

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

Watson, Gary. "The Trouble with Psychopaths." In *Reasons and Recognition: Essays on the Philosophy of T.M. Scanlon,* edited by R. Jay Wallace, Rahul Kumar, and Samuel Freeman, 307–31. New York: Oxford University Press.

Ziv, Anita Konzelmann. "Collective Guilt Feeling Revisited." *Dialectica* 61, no. 3 (2007): 467–93.

Journal of Ethics and Social Philosophy
VOL. 25, NO. 3 · SEPTEMBER 2023
https://doi.org/10.26556/jesp.v25i3.2359

INCLUSIVE BLAMEWORTHINESS AND THE WRONGFULNESS OF CAUSING HARM

Evan Tiffany

Villanelle wants Eve to die. She aims a gun at Eve and pulls the trigger with the intention of killing her.[1] The bullet strikes Eve who dies as a result of the gunshot. Intuitively, it seems clear that Villanelle is blameworthy for killing Eve. If correct, this would seem to imply that Eve's death plays a role in determining Villanelle's blameworthiness such that she would be less blameworthy had she not killed Eve. However, this claim is in tension with another powerful intuition regarding the significance of luck. Suppose that Oxana also attempted to kill Eve in circumstances that were exactly like those of Villanelle (including any morally relevant facts about their thoughts and motives), except that a chandelier fell in the path of the bullet, thereby preventing it from reaching Eve's body. To many, it seems counterintuitive that Oxana deserves less blame than Villanelle, given that they both performed the same volitional act with the same malicious intent, the only difference being that Oxana's plan was foiled by an unforeseeable event disrupting the causal chain between act and intended result. However, if blameworthiness must be immune from luck, then it is difficult to see how Villanelle can be blameworthy for Eve's death, given that it was a matter of luck that the bullet hit Eve rather than a falling chandelier. The tension between these two common intuitions provides an illustration of what is known in the philosophical literature as the problem of moral luck, specifically resultant moral luck.

There are three broad strategies for responding to this problem.[2] One response is to simply accept resultant moral luck, to accept that how much blame a person deserves can be partly determined by factors outside of one's control. A second is to deny the first intuition that Villanelle is blameworthy for *killing* Eve. This view I refer to as *robust internalism*, as it holds that one can only be blameworthy for internal manifestations of agency, such as one's intentions,

1 The names for the running example are taken from the television program *Killing Eve*.

2 More precisely, three *nonskeptical* responses. While I certainly feel the force of hard determinist worries about the legitimacy of desert-based blameworthiness, I am setting aside general responsibility skepticism for the purposes of this paper.

attitudes, or values, and not for external actions, like killings. A third strategy is to attempt to reconcile both intuitions. This approach is adopted by Michael Zimmerman, who draws on the distinction between the *degree* and the *scope* of blameworthiness to argue that it is possible for something (such as a person's death) to increase the number of things for which one is to blame without increasing the degree or severity of blame one is deserving of.[3] The intuition that Villanelle is to blame for Eve's death is captured by the claim that Eve's death is within the *scope* of Villanelle's blame, that it is among the things for which Villanelle is to blame. The intuition that Villanelle is not more blameworthy than Oxana is captured by the claim that Eve's death does not increase the degree or magnitude of Villanelle's blameworthiness.

The aim of this paper is to defend resultant moral luck. My strategy for defending this view begins by outlining an independently plausible theoretical framework that I refer to as the *inclusive conception* of blameworthiness, according to which the degree of blameworthiness is a function of two independent variables: wrongfulness and responsibility. I take one of the primary dialectic contributions of the paper to consist of reframing the debate over resultant moral luck in terms of the contribution of harm to the comparative *wrongfulness* of an action. When framed in these terms, I take the inclusive conception of blameworthiness, together with resultant moral luck, to create a more plausible package of theoretical commitments than competing views. In brief, this package consists of the following claims:

1. One can be blameworthy only if there is something one is blameworthy *for*.
2. Agents are paradigmatically blameworthy for what they *do*.
3. For a given wrongful action α, how much blame a person deserves for doing α is a function of both their degree of responsibility for and the moral wrongfulness of α.
4. For two agents A and B and two action-descriptions α and β, it is possible that A's doing α is a morally worse offense than B's doing β, even if the fact that B did β rather than α is due only to the fact that factors outside of B's control prevented B from having the opportunity to do α.
5. For two agents A and B and two action-descriptions α and β, it is possible that A's doing α is a morally worse offense than B's doing β, even if the fact that B did β rather than α is due only to the fact that factors outside of B's control prevented B from bringing about the intended harm.

3 Zimmerman, "Moral Responsibility and Ignorance" and "Taking Luck Seriously."

Zimmerman is committed to denying 1, while the robust internalist is committed to denying 2. Claim 3 expresses the inclusive conception of blame. The view that I call "qualitative scoping" could accept all three claims but faces dialectic pressure to accept 4, which in turn puts dialectic pressure on accepting the kind of resultant moral luck expressed in 5.

1. RESPONSIBILITY AND DESERT-BASED BLAME

The central question this paper takes up is whether the harm caused by one's actions contributes to the degree of blame one is deserving of. Before answering this, it is important to clarify some of the key concepts involved in this question, beginning with the notion of blameworthiness.

In this paper, I shall understand blameworthiness in terms of Gary Watson's notion of "accountability," which he distinguished from "attributability."[4] Whereas attributability concerns the "aretaic face" of responsibility whereby we appraise a person as "an adopter of ends" or "an agent in a strong sense" in virtue of what their actions disclose about their "deep self," accountability concerns responsibility's "deontic face" whereby we appraise whether a person is an apt target of "adverse or unwelcome treatment" in virtue of failing to satisfy certain "expectations or demands or requirements."[5] While such an accountability conception of blame may be at odds with those who prefer to focus on the attributionist face of responsibility, I take it to be compatible with a wide variety of views regarding the nature of the sanctions associated with accountability.[6] For example, the currency of criminal punishment is often held to be suffering, but it could also take the form of a monetary penalty, the imposition of community service, or the (temporary) deprivation of certain rights and privileges

4 Watson, "Two Faces of Responsibility."

5 Watson, "Two Faces of Responsibility," 229, 237, and 235. While Watson does not explicitly use the term "deontic" to describe the accountability face of responsibility, Darwall, in "Taking Account of Character," does refer to attributability and accountability in terms of the aretaic and deontic dimensions of responsibility. However, in contrast to Watson, Darwall takes accountability to be disanalogous to attributability in that the former only has the negative dimension of blame.

6 Cf. Scanlon: "Questions of 'moral responsibility' are most often questions about whether some action can be attributed to an agent in the way that is required in order for it to be a basis for moral appraisal" (*What We Owe to Each Other*, 248). An attributionist could either hold that attributability is all there is to responsibility or hold that attributability is sufficient for accountability; the former would amount to a type of eliminativism and the latter a type of reductionism with respect to accountability.

of citizenship.[7] In the context of interpersonal morality, the adverse treatment could, *à la* Strawson, take the form of being the target of resentment or some other negative emotion.[8] Alternatively, it could take the form of a modification or withdrawal of one's friendship.[9] While proponents of the "moral ledger" view of blame may not have originally conceived of their view as putting forth a claim about the nature of moral sanction, I see no reason why a "negative mark" or "blemish" on one's "moral ledger" could not be understood as a type of unwelcome treatment in the same way that a negative score on one's credit report could be seen as adverse treatment in response to a failure to pay one's bills on time.[10]

When it comes to the *justification* of blame, I assume a broadly retributivist view in the sense that I take the justification for any particular sanction to be based primarily on backward-looking considerations of *desert* as opposed to forward-looking considerations regarding the expected benefits of that sanction. That is, I hold that desert provides a necessary condition on the appropriateness of a given sanction and that the degree of the sanction should be proportional to desert, especially with respect to the upper limit. When I say that blame is *primarily* a matter of desert, I leave open the possibility for non-desert-based reasons—including, *inter alia*, evidence of remorse and restitution, the prospects of rehabilitation, and the value of mercy and forgiveness—to factor in as well, especially with respect to justifying a sanction that is less than what is strictly deserved.[11] For the purposes of this paper, however, I shall focus exclusively on the question regarding the degree of blame that one *deserves*, for what skeptics of resultant moral luck typically deny is that one person "*deserves* . . . a harsher reaction than" another based on resulting harm,

7 For examples of the view that the currency of punishment is suffering, see Ross, *The Right and the Good*, 135–38; Hart, *Punishment and Responsibility*, 234–35; Tadros, *The Ends of Harm*, 63. For the point that the currency takes multiple different forms, see Brink, "The Nature and Significance of Culpability," 351.

8 Strawson, "Freedom and Resentment."

9 Scanlon explicitly states that "blame . . . is not a kind of sanction" (*Moral Dimensions*, 122); rather "to blame a person for an action . . . is to take that action to indicate something about the person that impairs one's relationship with him or her" (122–23). Thus, Scanlon seems to view the impairment of a relationship as the object of blame—or the external manifestation of the object of blame—rather than as a constituent of the blame itself. However, there seems to be conceptual space for thinking of accountability within a Scanlonian framework of interpersonal relationships such that the modification or termination of a relationship could be a kind of "treatment" that one is deserving of in response to conduct that has impaired (perhaps unforgivably) the relationship.

10 For examples of the ledger view, see Haji, *Moral Appraisability*; and Zimmerman, *An Essay on Moral Responsibility*.

11 This is the view that David Brink labels "predominant retributivism" ("Retributivism and Legal Moralism" and "The Nature and Significance of Culpability").

leaving "open the possibility that it would be morally justified to react more harshly toward [one who succeeded in causing harm] than toward [one whose attempt was thwarted] on grounds *other* than those having to do with desert."[12]

This last way of putting the point—that a person deserves a harsher reaction—points to an important clarification regarding what I mean in saying that one person is more blameworthy than another. Following Robert Hartman and Justin Coates, we can distinguish between the claim that a person is more (or less) deserving of blame from the claim that they are deserving of more (or less) blame.[13] The former reading indicates how strongly one stands in the desert relationship to blame; as Coates puts it: "for *A* to be more deserving of blame for a token of an *x*-type transgression than *B* is for a token of an *x*-type transgression, there are weightier reasons for blaming *A* than for blaming *B*."[14] In contrast, the latter reading is not about how weighty the reasons are for blaming a given person but how harsh or stringent a sanction they are deserving of. For example, to say that an adult offender is more blameworthy than a juvenile who has committed a type-identical offense in this latter sense is to say that the adult offender is deserving of a harsher punishment (e.g., a longer prison sentence). It is this second sense that I have in mind in this paper—when I say that Eve's death can make Villanelle more blameworthy than Oxana, I mean that it can make Villanelle deserving of greater sanction.

In explicating the sense of "more blameworthy" that I have in mind, I contrasted an adult and a juvenile offender committing a type-identical offense. While there are different views about why juvenile offenders are deserving of less punishment, one common view is that they are less responsible because they have less control over their actions, and they have less control because their agential or reasons-responsive capacities are not fully developed.[15] We can also contrast two agents who have type-identical control while committing different offenses. If *A* commits murder and *B* commits petty theft, then it is also the case that *A* is more blameworthy—is deserving of more punishment—than *B*, even if they have type-identical control over their respective wrongdoing. According to what I will call the *inclusive conception* of blameworthiness, these two examples illustrate each of two independent components of the desert base for blameworthiness: responsibility and wrongfulness.[16]

12 Zimmerman, "Taking Luck Seriously," 562, emphasis added.

13 Hartman, *In Defense of Moral Luck*, 34; Coates, "Being More (or Less) Blameworthy."

14 Coates, "Being More (or Less) Blameworthy," 235.

15 See, e.g., Brink, "Immaturity, Normative Competence, and Juvenile Transfer"; Scott, "Criminal Responsibility in Adolescence."

16 I use the label "inclusive conception" following David Brink's label "inclusive culpability" for the type of culpability that includes (is inclusive of) both wrongdoing and

On the inclusive conception, blame is a fitting response to wrongdoing for which one is culpable or responsible, where the culpability or responsibility and the wrongdoing are independent variables.[17] This view has the advantage of mapping onto the two main categories of culpability-denying defenses in law and morality: justification and excuse. In law, justification defenses, such as necessity or self-defense, deny wrongdoing; they deny that the defendant's action was criminal in nature and thus that a criminal offense has taken place.[18] In contrast, excuse defenses, such as insanity or duress, accept that a criminal offense was committed but deny that the defendant should be held criminally liable because they lacked the capacity or (fair) opportunity to avoid committing the offense.[19] Likewise, when a person is morally called to account for some *pro tanto* wrongdoing, they might justify their conduct by citing reasons that make the action all-things-considered morally permissible, thereby denying overall wrongdoing. Alternatively, they may accept that they ought not to have done what they did but offer an explanation that denies or mitigates their responsibility for the offense, e.g., by explaining how the circumstances led them to (mistakenly) believe they had good reasons for acting as they did.[20]

---

responsibility. In the context of criminal culpability, Brink understands "inclusive culpability" as encompassing both "narrow culpability" (elemental *mens rea*), which is an ingredient of the wrongdoing, and "broad culpability," which refers to the defendant's moral responsibility for the wrongdoing. See Brink, "The Nature and Significance of Culpability" and *Fair Opportunity and Moral Responsibility*. By "desert-base," I mean the base or "ground" of desert—i.e., that in virtue of which one is deserving of accountability blame.

17 See Brink, "Retributivism and Legal Moralism" and "The Nature and Significance of Culpability"; Moore, *Placing Blame*; Nozick, *Philosophical Explanations*.

18 See Dressler: "Justified conduct is conduct that under ordinary circumstances is criminal, but which under the special circumstances encompassed by the justification defense is not wrongful and is even, perhaps, affirmatively desirable" (*Understanding Criminal Law*, 208).

19 Some conceptualize duress as a justification, rather than excuse, because they take (*pro tanto*) wrongdoing committed in response to an unlawful threat to be all-things-considered justified in the circumstances; see, e.g., Westen, "Does Duress Justify or Excuse?" Craig Agule argues for a middle position, such that duress shares aspects of both justification and excuse ("Distinctive Duress"). Some prefer to conceptualize insanity as an "exemption" rather than an "excuse."

20 It is also standard in the responsibility literature to recognize an epistemic condition on responsibility. On my view, how this condition factors into the inclusive conception is a complex matter, as the ignorance or mistaken belief can function either as a moral analogue of a negating defense (by affecting which action descriptions can be aptly imputed to a person) or a moral analogue of an affirmative defense (by defeating or mitigating moral responsibility); see Tiffany, "Imputability, Answerability, and the Epistemic Condition on Moral and Legal Culpability."

Following Robert Nozick and David Brink, we can represent this view of blameworthiness in terms of the following formula:[21]

$$B \propto D\ (= W \times R) \qquad (1)$$

The $B \propto D$ part of the formula expresses the claim that the degree of blame that is appropriate or fitting is proportional to the target's desert, and the $D = W \times R$ component expresses the view that the desert-base (D) for punishment is wrongdoing (W) for which the target is responsible (R), where these are independent and scalar variables. Numerically, one can think of the "R" component as ranging from 0 to 1, with "1" representing full responsibility and "0" representing no responsibility (full excuse). It can be thought of as a "multiplier" for the degree and type of sanction associated with a given wrong, such that a person whose responsibility is diminished—e.g., because their cognitive or volitional capacities are diminished, or they are acting under coercive pressure—deserves less sanction than a person who commits the same offense in full possession of their rational capacities and free from external pressure. The "W" component should likewise be thought of in terms of a numerical representation of "seriousness of moral wrong" along some interval between "least wrong" and "most wrong." However, all of this should also be understood with the following caveats.

Despite the quasi-mathematical nature of the formula, we should not interpret it as indicating that deserved blame can be calculated with precision. For one thing, desert may only determine an appropriate interval, rather than a precise quantum, due either to genuine metaphysical indeterminacy regarding the desert base or to epistemic indeterminacy regarding our ability to accurately detect small differences in wrongdoing or responsibility.[22] The quasi-mathematical *representation* of the desert-base for blame should not be taken to indicate that there is some metaphysical fact of the matter as to the precise percentage of responsibility one bears or that the moral quality of any given offense can be precisely quantified and measured, much less that we have the epistemic capacity to detect and measure such things with precision. Whether it even makes sense to think in terms of a *quantum* of blame may depend on how one understands the nature or currency of blame. Criminal punishment, for example, is often expressed in quantifiable terms, such as days in prison or hours of mandatory community service. Similarly, the "moral ledger" view of

21 This version of the formula comes from Brink, "Retributivism and Legal Moralism," 498, and "The Nature and Significance of Culpability," 350, who adapts it from Nozick, *Philosophical Explanations*, 363. Whereas both Brink and Nozick use "P" for "punishment," I use "B" for "blameworthiness."

22 For example, the United States Federal Sentencing Guidelines reflect interval sentencing.

moral blame may admit of quantification in terms of the "number of demerits" one receives on one's moral "scorecard." In contrast, if one thinks of blame as an expression of the reactive emotions, it is more difficult to quantify the intensity of resentment that one is deserving of.[23] In some cases, it might be more appropriate to understand degree of blame in categorical, rather than continuously scalar, terms—for example, in terms of the distinction between moral disdain, ordinary resentment, and mere annoyance. The point is that, even when expressed in a more qualitative type of currency, the type or intensity of reactive attitude one deserves is a function of both the moral turpitude of the wrong and the degree of responsibility or control one had over that wrong. For example, if I learn that you did not intend to step on my hand, I may either withdraw or mitigate the intensity of my reaction, depending on whether I think your behavior was (nonculpably) inadvertent and so completely blameless or whether it still manifested some (lesser) moral failure, such as recklessness or negligence.

Caveats aside, the fundamental idea behind the retributivist formula is that the degree of blame one deserves is a product of both the magnitude of the wrongfulness of one's conduct and the degree of responsibility one had for that wrongdoing. This matters to the debate over resultant harm because the view that the actual harm caused by one's actions contributes to one's degree of blameworthiness is most plausibly interpreted as a claim about the contribution of that harm to the moral *wrongfulness* of that for which one is being blamed. While there is a sense in which some of the same concerns about the relevance of facts external to a person's agency, such as resultant harm, to a person's degree of responsibility will reappear as concerns about the relevance of those facts to the degree of wrongfulness, I do not think this merely replays the exact same debate in different language. At the very least, reframing the debate in terms of the inclusive conception changes the contours of the dialectic, as it is not clear that the same considerations about control apply equally to judgments about wrongfulness as they do to questions about responsibility.

This consequence is perhaps most clearly demonstrated in the context of the scoping strategy. When framed as a debate about the relevance of resultant harm to the wrongfulness of one's conduct, the scoping strategy occupies an unstable middle position, as the considerations that would support excluding results from an evaluation of the moral wrongfulness of one's actions would also support excluding them from the scope.

23 Coates develops this point in more detail in "Being More (or Less) Blameworthy," 239–41.

2. AGAINST ZIMMERMAN'S SCOPING STRATEGY

One of the theoretical advantages of the scoping strategy, according to which factors such as outcomes can affect the scope but not the degree of blameworthiness, is its ability to accommodate both the intuitions that Villanelle is responsible for Eve's death and that luck should not determine how much blame one is deserving of. The problem is that once the debate is reframed in terms of the contribution of harm to the degree of wrongfulness, the rationale that Zimmerman offers for the scoping strategy no longer makes dialectic contact with the relevant opponent. Consider how Michael Zimmerman initially articulates and defends the view:

> Although [Eve's] death may have added to the number of things for which [Villanelle] is to *blame*, it did not increase the degree to which she is to *blame*. Given [her] death, she may be to blame for more, but she is no more to blame than she would be had [Eve's] death not occurred. The reason for this is that [Villanelle] was only indirectly in control of [Eve's] death. That is, she was in direct control of something of which [Eve's] death was a consequence. . . . Her control did not extend beyond this something with respect to which she was directly free; there was no fresh injection of freedom beyond that point. Given that responsibility tracks freedom, there was therefore no fresh injection of responsibility beyond that point; her *responsibility* was not extended, its degree was not increased, by [Eve's] death.[24]

While he begins by stating the thesis in terms of blame, the rationale he offers speaks only to the kind of control that grounds responsibility.

In order to represent the scoping strategy in the formal terms introduced above, we can represent the scope of blameworthiness as objects of the variables, such that "$B(a)$" can be read as "blameworthy for a." Thus (1) can be rewritten as:

$$B(a) = W(a) \times R(a) \qquad (1')$$

According to the scoping strategy, how things turn out can increase the number of things for which one is to blame without increasing how much blame one deserves. This can be represented as follows:

$$B(a) = B(a + \beta) \qquad (2)$$

24 Zimmerman, "Moral Responsibility and Ignorance," 419, emphasis added.

The "β" variable indicates an additional element in the scope of blameworthiness, for example, a resulting harm, such as a person's death. According to the inclusive conception, this is equivalent to the following:

$$W(\alpha) \times R(\alpha) = W(\alpha + \beta) \times R(\alpha + \beta) \quad (3)$$

This way of representing the scoping strategy highlights the gap in Zimmerman's argument, quoted above, as the rationale he provides speaks only to the responsibility component. That is, Zimmerman may have given us a good reason for accepting:

$$R(\alpha) = R(\alpha + \beta) \quad (4)$$

But this would produce an equivalent degree of blameworthiness only if it is also the case that:

$$W(\alpha) = W(\alpha + \beta) \quad (5)$$

The scoper owes an argument for claim (5), as it is not sufficient to point out that there is no difference in control. Imagine a third assailant, Irina, who intends only to (nonfatally) wound Eve. Zimmerman would accept that Villanelle and Oxana are both more blameworthy than Irina even if they all have type-identical control. According to the inclusive model, this difference is plausibly explained by a difference in the wrongfulness of intending to kill versus intending to (nonfatally) wound. If correct, the moral luck skeptic owes an account of the nature of this wrongfulness such that it distinguishes between Villanelle and Irina, but not Villanelle and Oxana. When it comes to offering such an explanation, Zimmerman is in a particularly vulnerable dialectic position.

First, he accepts that states of affairs can be within the scope of blameworthiness. In the above quotation, he puts the point in terms of the victim's death; in a different paper, he writes: "I do not wish to deny that [the assassin] is responsible for killing [the victim] (or for [the victim's] death—the distinction between actions and their 'results' seems to me irrelevant here)."[25] In other words, the "β" variable in the above formulas can refer to the state of affairs in which the victim is dead. However, states of affairs are morally assessed in terms of their axiological value.[26] If some state of affairs is included in the scope

25 Zimmerman, "Taking Luck Seriously," 560.

26 Here I have in mind the distinction that Darwall draws between "ought-to-be" and "ought-to-do." Darwall reads G. E. Moore as holding that "what most fundamentally possesses intrinsic value for Moore is a *state of affairs* … the normative proposition entailed by a thing's having intrinsic value is that the state of its existing ought to be" ("How Should Ethics Relate to (the Rest of) Philosophy?" 26, emphasis original). This, according to Darwall, is Moore's fundamental metaethical mistake: "Moore's failure to understand

of *W* and the moral (dis)value of that state of affairs is a matter of the intrinsic (dis)value contained in that state of affairs, then it is not clear why that intrinsic (dis)value is irrelevant to the magnitude of *W*. Put differently: if state of affairs β is relevant to *W*, then the value of β would seem to be a blameworthy-relevant value; hence, if β's value is axiological, then axiological value would be a blameworthy-relevant value. In contrast, if axiological value is irrelevant to the magnitude of *W* and states of affairs are fundamentally bearers of axiological value, then, contra the scoping strategy, that would seem to be a reason to exclude states of affairs from the scope of *W*.

Zimmerman could resist the above argument by denying that the fact that some state of affairs β is within the scope of *W* entails that β's value is relevant to the value of *W*. That is, Zimmerman could include states of affairs within the scope of *W* (or *B*) while excluding the relevance of axiological value to the magnitude of *W* (or *B*) by denying that the moral status of *any* of the objects within the scope of *W* (or *B*) is relevant to the magnitude of *W* (or *B*).[27] This reading is supported by the fact that, in order to rule out the possibility of circumstantial moral luck, Zimmerman is willing to accept that a person can be blameworthy even if "the scope of [their] responsibility has dwindled to nothing," even when they are "not responsible *for* anything."[28] To illustrate, imagine that Villanelle, Oxana, and Dasha are all supposed to meet at a bar to discuss their mortal enemy, Eve; however, on her way to the meeting, Dasha's car breaks down, and she is unable to attend. At the meeting, Villanelle and Oxana end up placing a bet on who will be the first one to kill Eve, and then they each proceed as before with their murderous plans. Because Dasha was not there, she knows nothing about the bet and never conceives of killing Eve. However, we can suppose that had her car not broken down, she would have

reasons for action and ought-to-do's, which he reduced to the ought-to-be's he identified with intrinsic value" (26). While one could conceptualize actions and attitudes in terms of states of affairs—e.g., the state of affairs in which one values *x* or does *y*—I find it plausible to follow Darwall in emphasizing the normative difference between these ontological categories. If, for example, one accepts that "oughts gain their sense from norms; only what can be regulated by norms can be subject to normative judgment," then it seems plausible that "we must understand ought-to-be's as elliptical and underspecified, requiring completion by reference to something that can be normatively regulated," such as actions and attitudes ("How Should Ethics Relate to (the Rest of) Philosophy?" 27).

27 As Khoury puts it: "On his approach, to say that *S* is blameworthy for ϕ is to make a claim that, by Zimmerman's own lights, is irrelevant to anything that matters for blameworthiness" ("The Objects of Moral Responsibility," 1363; Khoury develops his argument against Zimmerman at 1361–63). See also Hartman *In Defense of Moral Luck*, ch. 4, for a thorough argument against Zimmerman's view (and counterfactual views more broadly).

28 Zimmerman, "Taking Luck Seriously," 364.

entered the bet and "*would* have freely killed [Eve], *if* [she] had the cooperation of certain features of the case."[29] For Zimmerman, this is sufficient for Dasha to be equally blameworthy as Villanelle, even though Villanelle actually killed Eve while Dasha never so much as contemplated killing Eve. Contra Zimmerman, my view accepts the following:

> *Blameworthy-for*: To be blameworthy, one must be blameworthy *for* something.

I find this so intuitive as to constitute a platitude. I am not sure how to even conceptualize a person's being blameworthy without being blameworthy *for* anything at all.[30] I find it especially problematic to think about degree of blameworthiness independent of what a person is blameworthy for.

Setting aside Zimmerman's particular counterfactual view, there is a way to accept Blameworthy-for and still embrace the spirit of the scoping strategy. One could hold both that Dasha is not at all blameworthy because she did not actually do anything wrong and that Oxana and Villanelle are equally blameworthy because what they did was equally wrong. This does mean that we have to slightly revise how to understand the scoping strategy, as it seems more accurate to say that Villanelle is blameworthy for a *different* action—murder, as opposed to attempted murder—than to say that she is blameworthy for *more* actions. Call this the "qualitative scoping strategy":

> *Qualitative Scoping*: When a person *P* performs an act *A* that causes some set of consequences *C*, *C* can affect the quality of *A* in the sense that it can affect what descriptions aptly describe the act for which *P* is being blamed, but *C* cannot increase the degree to which *P* is to blame for that act.

On this view, one can accept that Villanelle "did something wrong that [Oxana] did not," but this is merely to accept that there is an act description that aptly describes what Villanelle did but not what Oxana did, and that the action is wrong under that description.[31] What the scoper is committed to denying is that the particular wrongful action that Villanelle performed (murdering Eve) is morally worse than the particular wrongful action that Oxana performed (attempting to murder Eve). In the next section, I consider an internalist strategy for defending this claim.

29 Zimmerman, "Taking Luck Seriously," 567.

30 Zimmerman explicitly rules out that a person in Dasha's situation is blameworthy for having the kind of will such that she would have killed Eve had circumstances been different ("Taking Luck Seriously," 564).

31 Zimmerman, "Taking Luck Seriously," 561.

3. QUALITY OF WILL AND THE WRONGFULNESS OF CAUSING HARM

Within the literature on moral responsibility, it is common to distinguish between reasons-responsive or agential-control and quality of will or agential-revelation views of responsibility.[32] It is not implausible to think that the same considerations that support a quality of will view of moral *responsibility* would also support a quality of will view of moral *wrongfulness*. According to this view, the scalar dimension of wrongfulness that is relevant to inclusive blameworthiness is exclusively determined by what something reveals about the moral quality of one's will:

> *Quality of Will*: For any two objects within the scope of one's blame, α and β, α is more wrongful than β only if the quality of will manifested by α is morally worse than the quality of will manifested by β.

To return to the original example, the quality of will manifested by Villanelle is no worse than that manifested by Oxana; the fact that an unforeseeable event interrupted the causal chain from Oxana's action to the intended result does not seem to diminish the quality of will manifested by her attempt. Thus, if Quality of Will is correct, it follows that Oxana's attempted murder is not more wrongful than Villanelle's murder. Since we have already conceded that they have the same degree of responsibility, it would follow that they share the same degree of blame. Eve's death may determine whether we can aptly describe Villanelle's act as murder or (merely) attempted murder, but it does not make her any more blameworthy than if something had intervened to prevent Eve's death.

I accept that Quality of Will enjoys intuitive support. How far that support extends, I think, depends on what objects fall within the scope of blameworthiness. According to what I will call *robust internalism*, only direct internal manifestations of agency—intending, valuing, willing, and so on—can be objects of blameworthiness. If correct, then Quality of Will seems eminently plausible, perhaps even trivial. This is because the objects within the scope of

32 Cf. Yaffe: "It is common for theorists of responsibility to contrast quality of will views of responsibility with reasons-responsive views" (*The Age of Culpability*, 77); and Guerrero: "It will be useful to have two broad pictures concerning moral responsibility … the agential control view [and] … an agential revelation view" ("Intellectual Difficulty and Moral Responsibility," 208). Proponents of the control view include Fischer and Ravizza, *Responsibility and Control*; Wolf, *Freedom Within Reason*; Nelkin, *Making Sense of Freedom and Responsibility*; Brink, *Fair Opportunity and Responsibility*. Proponents of the quality of will view of moral responsibility include Arpaly, *Unprincipled Virtue*; Harman, "Does Moral Ignorance Exculpate?"; Hieronymi, "Reflection and Responsibility"; Smith, "Attributability, Answerability, and Accountability"; Talbert, "Moral Competence, Moral Blame, and Protest."

blameworthiness are all different ways of manifesting the kind of "will" that is relevant to a quality of will view of responsibility—what one intends, values, wills, and so on are all ways of manifesting a certain "quality of will." Thus, anything that could affect the moral status of these is, by definition, something that affects the quality of one's will.

This is the view defended by Peter Graham, who argues that "what people are most fundamentally blameworthy for are their attitudes to and mental bearing toward those things of intrinsic value around them."[33] Commenting on a resultant luck scenario in which Bloggs succeeds at shooting Gomez in the leg while Jiggles' bullet is knocked off course by a baseball, Graham claims "there is no level of resentment it is appropriate, for *any* sense of 'appropriate,' for Gomez to feel toward Bloggs that it would not be appropriate for her to feel toward Jiggles."[34] Since Graham follows the Strawsonian tradition in understanding blameworthiness in terms of resentment, he takes this to imply that Jiggles is no less blameworthy than Bloggs. And, since they have the same intention but performed two distinct actions—shooting Gomez versus shooting a baseball—Graham takes the fact that they are equally blameworthy to imply that "what they're blameworthy for fundamentally is not their actions, but rather their intentions."[35]

If we adapt Graham's view to the inclusive model, the same considerations apply to each of the components of blameworthiness: wrongfulness and responsibility. This view fits quite naturally with Quality of Will, for if we limit the *ground* of moral appraisal to those features that are internal to one's moral agency—e.g., one's intentions, motives, and values—it is natural to likewise limit the *scope* of moral appraisal. And if we limit the objects of moral appraisal to internal features reflecting quality of will, then it is not clear how the scope of things we are responsible *for* can extend beyond those same features.

I do not have an argument against this type of robust internalism. However, in excluding even actions from the scope of things one can be blameworthy for,

33 Graham, "The Epistemic Condition on Moral Blameworthiness," 163. See also Khoury, who argues that "the rejection of resultant moral luck entails we can only be morally responsible for elements of our mental life" ("The Objects of Moral Responsibility," 1358). While Khoury argues from the rejection of resultant moral luck to internalism, he does provide reasons for skepticism with respect to resultant moral luck, some of which I discuss in section 4. As I note in my commentary on Graham, I do not take this kind of robust internalism to be my primary dialectic opponent in this paper. One could read this paper as adopting the inverse of Khoury's strategy, as arguing from the rejection of robust internalism to resultant moral luck, while providing both intuitive and theoretical support for the latter.

34 Graham, "The Epistemic Condition on Moral Blameworthiness," 170.

35 Graham, "The Epistemic Condition on Moral Blameworthiness," 170.

the view is highly revisionary. It is also the case that this view is not available to the scoper, as what made the scoping strategy initially appealing was the fact that it was able to accommodate the very strong intuition that Villanelle is blameworthy for killing Eve. According to what could plausibly be termed the "standard" view of responsibility, we are paradigmatically responsible for what we *do*, which brings me to the second key claim that identifies the package of claims that characterize my view:

> *Blameworthy-Actions*: Agents are paradigmatically blameworthy for what they *do*.

I say "paradigmatically" blameworthy, as I do not deny that we can be blameworthy for attitudes or mental "acts" (e.g., how we direct or fail to direct our attention). Rather, I take it to be a feature of our blaming practices that, typically, it is wrong*doing* that merits a blaming response. Once we extend the scope of blameworthiness to include external acts, it is not clear why the resulting states of affairs that partly constitute those actions are not also relevant to our moral appraisal of those actions.

On a plausible view of action, whether some action can be properly ascribed to one depends on the description under which it is being ascribed. For example, in both law and ordinary language, the act of causing someone's death can only be described as "murder" if it involves either the direct intention to kill or some mental attitude that is morally equivalent to a direct intention, such as "depraved indifference." If one's mental attitude at the time of the offense lacks this quality, then what one did cannot be aptly imputed under the description "murder." It could, however, be imputed under some other description, such as "reckless homicide," and under that description the action could still be wrongful, albeit less wrongful than directly intending some harm.

In criminal law, this view regarding the relevance of one's mental attitude is captured by the *mens rea* condition. The American Legal Institute's Model Penal Code recognizes four categories of *mens rea*—intention, foreknowledge, recklessness, and negligence—which are plausibly viewed as "reflect[ing] four grades of culpability from greater to lesser culpability and sometimes define distinct offenses."[36] As Kenneth Simons puts it, we can "differentiate different mental states according to the relevant blameworthiness they display, holding constant a particular object of those mental states: intending to cause a death is more blameworthy than being reckless or negligent as to causing death."[37] On the inclusive conception, this is plausibly explained by a difference in the

36 Brink, "The Nature and Significance of Culpability," 360.

37 Simons, "Culpability and Retributive Theory," 367.

wrongfulness of intending to cause some harm versus foreseeing and causing without intending the harm versus recklessly causing the harm versus negligently causing the harm. However, in addition to the *mens rea* element, criminal offenses are also constituted by an *actus reus* component, which can be understood as a voluntary act that causes some social harm.[38] There are three important features of this model of criminal culpability that I think plausibly hold for moral blameworthiness as well.

First, it is noteworthy that the *actus reus* condition contains both the external act and the resulting harm as constituent components of the "guilty action." This is not just a feature of legal definitions or the nature of criminal law; ordinary language also includes resulting states of affairs as constituent components of action descriptions. As the external manifestation of a person's will, actions extend into the world, and how they do so partly constitutes the nature of the action. Causing the baseball to fly over the outfield wall in fair territory partly constitutes the action of "hitting a home run"; causing the washing machine to function properly partly constitutes "fixing the washing machine." If an outfielder's mitt intrudes into the ball's path, thereby preventing it from flying over the outfield wall, then one cannot aptly impute the batter's action under the description "hit a home run," even if the batter were the unlucky victim of a highly unlikely feat of athletic excellence on the part of the outfielder. Regardless of how much goodwill I put into my repair efforts, if I leave the washing machine no more functional than when I found it, I cannot accurately say that I "fixed the washing machine." Whether Eve dies or something intervenes to save her life determines whether we can aptly describe Villanelle's act as that of killing Eve or merely attempting to kill Eve. Both the nature and content of one's mental attitude (*mens rea*) and the nature and consequences of one's conduct (*actus reus*) combine to determine the description under which an action can be imputed and, hence, the description under which we are to assess a person's responsibility for and the wrongfulness of that action.

Second, just as the different grades of *mens rea* correspond to different grades of wrongfulness, a difference in the magnitude of harm that partly constitutes the *actus reus* corresponds to a difference in the wrongfulness of the offense for which one stands accused. This can be captured by the following principle:

> *Harm Matters*: For any two action descriptions α and β, it is possible for α and β to differ in their moral wrongfulness due only to the resulting harm(s), or lack thereof, that partly constitute α and β.

38 See Dressler, *Understanding Criminal Law*, 85.

It is important not to equate Harm Matters with consequentialism. Even if consequentialism is false and the moral quality of an act is not determined exclusively by its consequences, it can still be the case that consequences at least *partly* determine its moral quality. I provide some intuitive support for this claim below. First, though, it is important to bring out one final feature of the criminal model that is central to establishing moral blameworthiness.

While the criminal law does recognize the relevance of actual harm caused to the degree of punishment one deserves, it is important to recognize the way in which *mens rea* interacts with *actus reus* to establish culpability for a given harm. It is standardly held that the *mens rea* condition is what connects the agent to the *actus reus* in a way that grounds a defendant's blameworthiness for an offense.[39] One way to conceptualize this connection is in terms of control—*mens rea* is what accounts for control over the conduct and resulting harm captured by the *actus reus*. It is odd, for example, to say that Villanelle had *no* control over whether Eve died, given that Eve's death is precisely what Villanelle intended and what she went to great lengths to make happen. As Michael Moore presses the point:

> In the situation where some defendant *D* intends to kill victim *V*, and where *D* carefully loads his gun, checking all bullets to be sure none are duds; tests the firing mechanism of the pistol; isolates *V* from all possible help or medical attention; screens off all birds or other objects that could interfere; puts the gun at *V*'s head, pulls the trigger, and kills him—I would say that *D controlled V*'s death.... If *D* has control over his choices, despite not having control over all the possible preventers or disrupters of those choices, does he also have control over *what* he chooses, *viz*., that *V* die? His choice ... is the product of his practical reasoning process. But so are the bodily movements by which he executes that choice, and so is the intended effect of those bodily movements, viz., *V*'s death. It is *D*'s reasoning processes that cause all three of them: *D*'s choice, *D*'s act of moving his finger, and *V*'s death.[40]

39 This function of *mens rea* is often made explicit in discussions of negligence. Cf. Herstein, commenting on the puzzle posed by negligence: "A person's responsibility for conduct turns on a type of connection between one's conduct and one's practical agency.... The paradigm for this conception of responsibility is the *intentional action* (or omission), wherein the responsibility-establishing connection between conduct and agency is obvious" ("Nobody's Perfect," 110, emphasis added); and Stark arguing that a "bare conduct-based account of negligence ... would be unacceptable for the criminal law, for it would fail to draw a clear, personalized link between the defendant and her wrongdoing" (*Culpable Carelessness*, 181).

40 Moore, *Causation and Responsibility*, 28.

It is true that Villanelle has no control over whether her bullet is blocked by a falling chandelier, but it does not follow from that that Villanelle had *no* control over Eve's death. After all, she could have chosen not to shoot Villanelle. In the same way, imagine a person working on a construction site who recklessly tosses heavy debris onto the sidewalk below in full knowledge that people are using the walkway and, hence, that there is a strong likelihood of injury. If a pedestrian is struck by a thrown brick, it would seem an exceptionally poor attempt at excuse for the worker to claim, "I had no control over whether anyone was injured; they were just unlucky to be at the spot where and when the brick landed." The natural response to such a plea would be to point out that the worker did have control insofar as he could have easily avoided causing the injury by refraining from recklessly tossing bricks off the roof in the first place. Of course, the moral-luck skeptic holds that the degree of blame should be determined exclusively by the relevant attitude or *mens rea*, e.g., intention or recklessness. The point I am making is simply that one can hold that actual harm is *relevant* to degree of blame while also requiring that one have some sufficient degree of control over that harm via the moral analogue of a *mens rea* condition with respect to that harm, e.g., by intending the harm or being willfully reckless with respect to that harm.

It may be that the respective proponents of Quality of Will versus Harm Matters simply have different intuitions about crucial cases. I find the intuitive case for Harm Matters especially persuasive when looking at things from the perspective of the victim or the victim's family (in the case of murder). Consider, for example, a scene from the film *Dead Man Walking* in which the Sister Helen character (played by Susan Sarandon) is trying to get condemned murderer Matthew Poncelet (played by Sean Penn) to take responsibility for his crimes. Poncelet angrily remarks how the father of one of his victims said he wants to administer the lethal injection to Poncelet himself, and Sister Helen responds as follows:

> Well, think of how angry he must be.
> He's never gonna see his daughter again.
> He's never gonna hold her, love her, laugh with her.
> You have robbed these parents of so much.
> They have nothing in their lives but sorrow, no joy.
> That is what you gave them.[41]

The last line here is key: "That is what you gave them." These consequences are all part of what Poncelet *did*—he robbed the parents of their child, of all the

41 Robbins, *Dead Man Walking*.

joy, the laughter, the experiences they would have shared with their daughter. Those consequences strike me as relevant to a deontic evaluation of the action's wrongfulness. Imagine a parallel scenario in which Poncelet fails to kill the teenagers because of resultant luck. Now the above is no longer an accurate description of what he did—it is no longer the case that he has robbed the parents of their daughter, of all the joy, laughter, and love that they would share. Given the context, Sister Helen's words would lack the same force when reflecting the lack of actual harm: "He *almost* never saw his daughter again . . . that is what you *attempted* to give them."

To take a less extreme case, consider a variation on Graham's example of shooting a person in the leg. Eve and Evee are both professional dancers; Villanelle successfully shoots Eve in the leg, while Oxana's attempt to shoot Evee in the leg is foiled by a falling chandelier. Two years later, Evee is a principal dancer for the Joffrey Ballet while Eve is reduced to serving drinks in the lobby, her career cut short due to the permanent muscle damage suffered as a result of Villanelle's action. My intuitions differ from Graham's in that it strikes me as entirely appropriate for Eve to seethe with resentment of Villanelle in a way that it does not seem appropriate for Evee to feel toward Oxana. After all, Villanelle robbed Eve of her dreams, her career, the one thing that had given meaning to her life—that is what Villanelle *did* to her. While Oxana may have tried to do the same to Evee, she failed. Evee still has her career; she is still able to pursue what brings meaning to her life. Because of this, it strikes me that what Villanelle did to Eve is morally worse than what Oxana did to Evee.

Not everyone will share my intuitions about these cases. Others might acknowledge that these examples have some intuitive force but argue that this is outweighed by the counter-intuitiveness of allowing for luck to determine the moral status of one's action. It is this argument from moral luck to which I turn in the final section.

4. Moral Appraisal and Moral Luck

While the concept of moral luck is typically discussed in the context of moral *responsibility* or blame, Nagel originally introduces the topic by speaking about the "moral judgment of a person and his actions."[42] One reason to prefer Quality of Will over Harm Matters thus comes from a desire to insulate the moral appraisal of actions from luck. In this section, I draw on recent work in the philosophy of criminal attempts to argue that the consequences of completely

42 Nagel, "Moral Luck," 24. He later offers the following definition: "Where a significant aspect of what a person does depends on factors beyond his control, yet we continue to treat him in that respect as an object of moral judgment, it can be called moral luck" (26).

eliminating moral luck are at least as counterintuitive as that of accepting resultant moral luck. In brief, I argue that we face the following trilemma: (1) accept that a person who has merely taken a few initial steps in pursuit of a murder plot is equally blameworthy as a successful murderer, (2) accept resultant moral luck, or (3) allow for circumstantial but not resultant moral luck. Most of the section is devoted to articulating why ruling out moral luck commits one to 1. To my mind, this is far more counterintuitive than simply accepting resultant moral luck, which leaves options 2 and 3. I end with some reasons for doubting that there is a principled way of distinguishing between circumstantial and resultant moral luck.[43]

In the previous section, I suggested that there is both intuitive and theoretical support for thinking of criminal culpability and moral blameworthiness as structurally analogous in that the object of blameworthiness can be understood in terms of action under a description: paradigmatically, we are blameworthy for wrongful actions, the content of which is determined by the particular description under which we are being blamed. Both law and morality distinguish between murder and attempted murder, where those guilty of the former are held to be more blameworthy than those who are (merely) guilty of the latter. Since the only difference between a murderer and an attempted murderer is often a matter of resultant luck—as in the Villanelle and Oxana example—moral luck skeptics reject this view. In order to probe this debate, I find it instructive to look more closely at the way the criminal law treats criminal attempts.

The philosophical literature on moral responsibility often discusses attempts, such as attempted murder, in terms of fully *completed* attempts, such that the only difference between a murderer and an attempted murderer is resultant luck.[44] However, in criminal law there are good reasons to recognize criminal attempts prior to completion. For example, if police were to interrupt Oxana just seconds before she is able to pull the trigger, it is intuitively plausible to think that she should still be charged with attempted murder, that the

43 Robert Hartman also argues from circumstantial moral luck to resultant moral luck (*In Defense of Moral Luck*, 105–11). The argument presented here was developed independently of Hartman's and can be read as compatible with and supplemental to Hartman's argument. The principal difference between my version and Hartman's is that: first, the argument developed here draws on the nature of attempts as they are conceptualized in the context of criminal law, in particular by observing how criminal attempts need not be fully complete; second, my argument focuses on the difference in the wrongfulness of one's activity at various stages along the path to completion.

44 In commenting on a scenario in which a would-be assassin is unable to get his shot off because the target turned into a doorway, Zimmerman states: "In this sort of case there isn't even an attempt on [the victim's] life" ("Taking Luck Seriously," 563).

difference in mere seconds should not mean the difference between a charge of attempted murder and some lesser offense, such as breaking and entering. Within the philosophy of law, this leads to the question of how to determine when an attempt "vests," that is, when a person's conduct over some period of time is sufficient to establish a criminal attempt.

According to the view defended by Gideon Yaffe, we can approach this question counterfactually.[45] On Yaffe's view, for some agent *D* who is motivated by an intention to commit some criminal offense *C* across some time interval from t_1 to t_2, *D* counts as having committed a criminal attempt if they satisfy the following counterfactual:

> *Completion Counterfactual*: If (1) from t_1 to t_2 *D* has the ability and the opportunity to *C* and does not fall prey to "execution failure," and (2) *D* does not (at least until after t_2) change his mind, then *D* would *C*.[46]

This view is much less revisionary than Zimmerman's counterfactual view in that it requires a defendant to have a criminal intention and that they are actually motivated by that intention. It also fits comfortably with the quality of will view, as whether a person satisfies the counterfactual is arguably an indication of the quality of will from which they are acting.

It is plausible to suppose that the Completion Counterfactual has a determinate truth value at any point along the path to completion.[47] Thus, for any point along the path to completion, either the Completion Counterfactual is true or false of a given individual; if true, then that individual has committed a criminal attempt.[48] Yaffe takes this to be a feature of his view, as it preserves the (American) criminal law's presumption of bivalence, and it is able to account for the intuition that being interrupted by the police just before one is able to complete one final muscular movement of the finger should not determine whether one is guilty of attempted murder or some minor offense such as breaking and entering.[49]

45 Yaffe, *Attempts*.

46 Yaffe, *Attempts*, 94.

47 As Brink notes, there may be some indeterminacy regarding the precise degree of capacity and opportunity required, but "it might still be true that for any given precisification of the counterfactual it will always be determinately true or false for any given individual and any given point in time whether she would commit the crime if she had ability and opportunity to do so" ("The Path to Completion," 189n6).

48 In the legal context there is the added *epistemic* complication of determining what counts as sufficient evidence for the truth of this claim, but we can set this aside as my concern in this paper is with the metaphysics of blameworthiness.

49 The presumption of bivalence is the presumption that a defendant can only be guilty or not guilty; other jurisdictions allow for a finding of partial guilt.

Whether Yaffe's account offers a plausible analysis of criminal attempts within the context of US criminal law, I find the presumption of bivalence problematic from the standpoint of moral blameworthiness. Following David Brink, I find it more plausible to think of blameworthiness for criminal attempts as scalar or multivalent, rather than binary or bivalent.[50] As Brink observes, "attempts are often temporally extended, unfolding over a period of time," and "typically involve a series or sequence of actions, such as conceiving of the offense, preparation and planning for the crime, and a sequence of steps in executing the plan."[51] To fix some terminology, we can say that "prior to the last act, attempts are *partially complete,* and their degree of completion is roughly a matter of proximity to the last act."[52] The problem with relying exclusively on the Completion Counterfactual is that it "will often be true fairly early in the execution of a plan."[53] For example, imagine a variation on the scenario discussed in section 2 involving Villanelle, Oxana, and Dasha placing a bet on who will be the first to kill Eve. Suppose that Dasha does attend the meeting, joins the bet, and begins to pursue her goal of killing Eve with the kind of determination that would ground the truth of the Completion Counterfactual. However, let us further suppose that she is only able to take a few preliminary steps, such as tracking Eve's daily routine, before she is arrested on unrelated charges and deprived of the opportunity to even make significant progress on the attempt. I find it deeply counterintuitive to hold that Dasha is equally blameworthy as Oxana, much less Villanelle. Following Brink, I find it much more plausible to hold that "partially complete attempts deserve censure sanction *proportionate* to their degree of completion."[54]

I find the inclusive conception of blameworthiness to provide a plausible analysis of the scalar approach to attempts, as I think that the difference in degree of blameworthiness is plausibly accounted for by a difference in their degree of wrongfulness, where the wrongfulness of an attempt is proportionate

50 Brink, "The Path to Completion."

51 Brink, "The Path to Completion," 189.

52 Brink, "The Path to Completion," 190.

53 Brink, "The Path to Completion," 189.

54 Brink, "The Path to Completion," 192. As Brink also remarks, one can conceptualize multivalence either as continuously scalar or "lumpy." On the lumpy model, we can recognize different categories defined by different thresholds—e.g., anything less than 25 percent complete qualifies as *de minimus* (hence not prosecutable), anything over 75 percent counts as fully complete, and anything in between counts as partially complete. Given various practical and epistemic limitations on making very fine-grained distinctions in degree of blameworthiness, there may be good reasons for adopting the lumpy model with respect to criminal culpability, even if, metaphysically speaking, moral blameworthiness is continuously scalar.

to its degree of completion. This, in turn, I think can be supported by reflection on the doctrine of abandonment. Abandonment occurs when an agent abandons a criminal attempt at some point prior to completion, and it is sometimes held to exculpate one of blameworthiness even if the attempt has already vested.[55] In a criminal context, this doctrine functions in part to provide would-be criminals with an incentive to abandon their plans prior to completion, and the state has an interest in creating such incentives. On the view I am defending here, it is also relevant to a person's desert-based (inclusive) blameworthiness in virtue of being relevant to the wrongfulness of that for which one is being blamed. Specifically, I think the following is a plausible generalization:

> *Abandonment*: It is less wrong to begin and abandon a plan to bring about some harm than to fully complete an attempt to bring about that harm.

Crucially, even if the Completion Counterfactual is true of a given agent from t_1 to t_2, it is still possible for them to abandon the plan prior to completion. That is, the following could both be true: (1) at some time t, within the interval from t_1 to t_2, if Dasha were presented with the opportunity to kill Eve at t, she would take it, and (2) at some time t_3, after t_2 but prior to completion, Dasha decides to abandon her attempt to kill Eve. Because the Completion Counterfactual is true of Dasha at t, I take it that Dasha (at t) manifests the same quality of will as Oxana. However, because Dasha still has the opportunity to abandon her plan, I take it that what she has actually done (as of t) is less wrongful than what Oxana did; hence, she is less blameworthy (at t).

If the Brink-inspired scalar account of attempts is plausible, it follows that the wrongfulness of what one has done can be partly a matter of (circumstantial) luck. While resultant luck is what distinguishes a successful murder from an unsuccessful but completed attempt at murder, circumstantial luck is (typically) what distinguishes a partially completed attempt from a fully completed attempt. One could be prevented from completing an attempt for a variety of factors outside of one's control, such as getting caught by the police, the death of one's intended victim due to natural causes, a sudden and drastic weather event, or the onset of depression. Even in cases where one abandons one's plan out of a genuine change of heart, it could be that the change of heart was occasioned by a fortuitous occurrence such as the chance appearance of a child who closely resembles the would-be assassin's own child. In any number of ways, what a person ends up actually freely doing is susceptible to circumstantial luck.

55 The Model Penal Code §5.02(4) and some jurisdictions treat abandonment as an affirmative defense in cases where a plan has been abandoned due to a genuine change of heart (rather than, e.g., fear of apprehension), meaning that the abandonment provides grounds for acquittal of the charged attempt.

Thus, we are left either accepting Zimmerman's highly revisionary counterfactual view whereby a successful murderer is equally blameworthy as a person who never even contemplated murder but would have committed murder under suitable counterfactual circumstances, or accepting the following claim:

> *Opportunity Matters*: For two agents *A* and *B* and two action-descriptions α and β, it is possible that *A*'s doing α is a morally worse offense than *B*'s doing β, even if the fact that *B* did β rather than α is due only to the fact that factors outside of *B*'s control precluded *B* from having the opportunity to do α.

Once one accepts a principle such as Opportunity Matters, the existence of resultant moral luck comes down to whether there is a principled difference between resultant moral luck on the one hand and circumstantial and constitutive moral luck on the other. In favor of the view that there is such a difference, Andrew Khoury reasons as follows:

> The compatibilist has a principled reason for drawing a line between resultant moral luck and other forms of moral luck such as circumstantial and constitutive moral luck (luck in one's circumstances and luck in who one is). The compatibilist can hold that it is the quality of an agent's will that determines responsibility. It is particular qualities of an agent's willing that are the *bearers of responsibility relevant value*. To the extent that luck can affect those qualities of will, then luck can affect responsibility, . . . hence, the compatibilist can accept circumstantial and constitutive moral luck.[56]

The rationale provided in this passage seems to depend on the following principle:

> *Blameworthy-Relevant Luck*: For some *X*, if *X* is a determinant of blameworthy-relevant value, then luck can affect blameworthiness by affecting *X*.

This strikes me as a plausible principle. It also strikes me as correct that quality of will is a determinant of blameworthy-relevant value and that this explains why we should accept constitutive and (some) circumstantial moral luck. However, I do not think it can do all the work a resultant-moral-luck skeptic wants it to do for two reasons.

56 Khoury, "The Objects of Moral Responsibility," 1374, original emphasis. See also: "What Nagel . . . identifies as constitutive and circumstantial moral luck are significantly less problematic than resultant moral luck precisely because they, as it were, flow through a person's agency whereas resultant luck bypasses one's agency altogether" (Khoury, "Responsibility, Tracing, and Consequences," 203).

First, it cannot explain the difference in blameworthiness between Dasha (who has merely taken some preliminary steps in her murder plan) and Oxana (who has fully completed her attempted murder) because, as I argued above, there is no difference in the quality of will manifested by Dasha and Oxana. Thus, the resultant-moral-luck skeptic must either reject Abandonment, explain how Abandonment is consistent with holding Oxana and Dasha to be equally blameworthy, or find an alternative explanation for the difference in blameworthiness.

Second, and more fundamentally, Blameworthy-Relevant Luck cannot help to explain the difference between resultant moral luck on the one hand and circumstantial and constitutive moral luck on the other without begging the question against resultant moral luck. On the view being defended here, resulting harm is relevant to the moral wrongfulness of one's action. If correct, then harm *is* a determinant of blameworthy-relevant *value*, which means that luck can affect blameworthiness by affecting harm. If Harm Matters, it entails the following counterpart to Opportunity Matters:

> *Harm Matters-Corollary*: For two agents *A* and *B* and two action descriptions α and β, it is possible that *A*'s doing α is a morally worse offense than *B*'s doing β, even if the fact that *B* did β rather than α is due only to the fact that factors outside of *B*'s control prevented *B* from bringing about the intended harm.

I have intentionally formulated the two luck principles in contrastive terms, as I think that the case against resultant moral luck gains some intuitive plausibility when stated in contrastive terms—e.g., when we ask whether one had control over doing α *rather than* β. In the previous section, I argued that if a person acts volitionally with the intention to bring about some harm, such as a person's death, then it does seem that person has a morally relevant sense of control over the victim's death. But this can be true even if that person did not have control over whether they did α *rather than* β due to a lack of control over whether something intervenes to prevent the intended harm from coming about. But precisely the same thing is true of opportunity. One can have control over doing α, even if one does not have control over doing α *rather than* β due to a lack of control over whether something intervenes to deprive one of the opportunity to do β.

It is, of course, possible that there is a way of articulating a principled difference between resultant moral luck and circumstantial moral luck such that the former is morally problematic in a way the latter is not. However, I find it more parsimonious to simply accept that once we move "outside the head" and extend the scope of moral appraisal to include the ways in which we interact with the world, we invite various forms of moral luck.

5. CONCLUSION

My aim in this paper has been to contribute to the ongoing dialogue on resultant moral luck by reframing the central issue in terms of the inclusive conception of blameworthiness and the contribution of harm to the wrongfulness of that for which one is being blamed. While some of the same considerations regarding the relation between luck and responsibility will resurface regarding the relation between luck and wrongfulness, I have tried to show that reframing the debate in this way at least changes some of the contours of the dialectic, most importantly with respect to the role of control. Whereas a difference in control seems to necessarily correspond to a difference in responsibility, the same is not true of wrongfulness, which places the proponent of resultant moral luck in a stronger dialectic position.

Over the course of the paper, I have tried to defend a set of claims, including resultant moral luck, that combine to provide an intuitively plausible and theoretically sound package. I have argued that the scoping strategy—whether in Zimmerman's counterfactual version or the modified qualitative version—is in a particularly vulnerable dialectic position. Zimmerman's view involves the radically revisionary claim that one can be blameworthy without being blameworthy for anything. The qualitative scoper, on the other hand, must either accept that a person who has merely taken a few initial steps toward a murder plot is equally blameworthy as a successful murderer or make an *ad hoc* distinction between resultant and circumstantial moral luck. The robust internalist may be in the strongest dialectic position vis-à-vis the proponent of resultant moral luck insofar as they have an internally coherent position and can offer a principled explanation for why resultant moral luck does not affect blameworthiness. It may be that the debate between these positions comes down to the dull thud of clashing intuitions, but I take it to be a theoretical virtue of the view defended here that it is less revisionary in preserving the thought that we are typically blameworthy for what we do, where what we do extends beyond the mind.

Simon Fraser University
etiffany@sfu.ca

REFERENCES

Agule, Craig K. "Distinctive Duress." *Philosophical Studies* 177, no. 4 (April 2020): 1007–26.

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

Brink, David O. *Fair Opportunity and Responsibility*. Oxford: Clarendon Press, 2021.

———. "Immaturity, Normative Competence, and Juvenile Transfer: How (Not) to Punish Minors for Major Crimes." *Texas Law Review* 82 (2004): 1555–85.

———. "The Nature and Significance of Culpability." *Criminal Law and Philosophy* 13, no. 2 (June 2019): 347–73.

———. "The Path to Completion." In *Oxford Studies in Agency and Responsibility*, vol. 4, edited by David Shoemaker, 183–205. New York: Oxford University Press, 2017.

———. "Retributivism and Legal Moralism." *Ratio Juris* 25, no. 4 (December 2012): 496–512.

Coates, D. Justin. "Being More (or Less) Blameworthy." *American Philosophical Quarterly* 56, no. 3 (July 2019): 233–46.

Darwall, Stephen. "How Should Ethics Relate to (the Rest of) Philosophy? Moore's Legacy." In *Metaethics After Moore*, edited by Terry Horgan and Mark Timmons, 17–37. Oxford: Clarendon Press, 2006.

———. "Taking Account of Character and Being an Accountable Person." In *Oxford Studies in Normative Ethics*, vol. 6, edited by Mark Timmons, 12–36. New York: Oxford University Press, 2016.

Dressler, Joshua. *Understanding Criminal Law.* 6th ed. New York: LexisNexis, 2012.

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

Graham, Peter. "The Epistemic Condition on Moral Blameworthiness: A Theoretical Epiphenomenon." In *Responsibility: The Epistemic Condition*, edited by Philip Robichaud and Jan Willem Wieland, 163–79. New York: Oxford University Press, 2017.

Guerrero, Alex. "Intellectual Difficulty and Moral Responsibility." In *Responsibility: The Epistemic Condition*, edited by Philip Robichaud and Jan Willem Wieland, 199–218. New York: Oxford University Press, 2017.

Haji, Ishtiyaque. *Moral Appraisability: Puzzles, Proposals, and Perplexities.* New York: Oxford University Press, 1998.

Harman, Elizabeth. "Does Moral Ignorance Exculpate?" *Ratio* 24, no. 4 (December 2011): 443–68.

Hart, H. L. A. *Punishment and Responsibility: Essays in the Philosophy of Law.* Oxford: Clarendon Press, 1968.

Hartman, Robert. *In Defense of Moral Luck: Why Luck Often Affects*

Praiseworthiness and Blameworthiness. New York: Routledge, 2017.
Herstein, Ori J. "Nobody's Perfect." *Canadian Journal of Law and Jurisprudence* 32, no. 1 (February 2019): 109–25.
Hieronymi, Pamela. "Reflection and Responsibility." *Philosophy and Public Affairs* 42, no. 1 (Winter 2014): 3–41.
Khoury, Andrew C. "The Objects of Moral Responsibility." *Philosophical Studies* 175, no. 6 (June 2018): 1357–81.
———. "Responsibility, Tracing, and Consequences." *Canadian Journal of Philosophy* 42, nos. 3/4 (September/December 2012): 187–208.
Moore, Michael S. *Causation and Responsibility: An Essay in Law, Morals, and Metaphysics*. New York: Oxford University Press, 2009.
———. *Placing Blame: A General Theory of the Criminal Law*. New York: Oxford University Press, 1997.
Nagel, Thomas. "Moral Luck." In *Mortal Questions*, 24–38. New York: Cambridge University Press, 1979.
Nelkin, Dana Kay. *Making Sense of Freedom and Responsibility*. New York: Oxford University Press, 2011.
Nozick, Robert. *Philosophical Explanations*. Cambridge, MA: Harvard University Press, 1981.
Robbins, Tim, dir. *Dead Man Walking*. Gramercy Pictures, 1995.
Ross, David. *The Right and the Good*. Edited by Philip Stratton-Lake. Oxford: Clarendon Press, 2002.
Scanlon, T. M. *Moral Dimensions: Permissibility, Meaning, Blame*. Cambridge, MA: Harvard University Press, 2008.
———. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press, 1998.
Scott, Elizabeth S. "Criminal Responsibility in Adolescence: Lessons from Developmental Psychology." In *Youth on Trial: A Developmental Perspective on Juvenile Justice*, edited by Thomas Grisso and Robert G. Schwartz, 291–324. Chicago: University of Chicago Press, 2000.
Simons, Kenneth. "Culpability and Retributive Theory: The Problem of Criminal Negligence." *Maryland Journal of Contemporary Legal Issues* 5 (January 1994): 365–98.
Smith, Angela M. "Attributability, Answerability, and Accountability: In Defense of a Unified Account," *Ethics* 122, no. 3 (April 2012): 575–89.
Stark, Findlay. *Culpable Carelessness: Recklessness and Negligence in the Criminal Law*. New York: Cambridge University Press, 2016.
Strawson, Peter. "Freedom and Resentment." In *Free Will*, 2nd ed., edited by Gary Watson, 72–93. New York: Oxford University Press, 2003.
Tadros, Victor. *The Ends of Harm: The Moral Foundations of Criminal Law*. New

York: Oxford University Press, 2011.
Talbert, Matthew. "Moral Competence, Moral Blame, and Protest." *Journal of Ethics* 16, no. 1 (March 2012): 89–109.
Tiffany, Evan. "Imputability, Answerability, and the Epistemic Condition on Moral and Legal Culpability," *European Journal of Philosophy* 30, no. 4 (December 2022): 1440–57.
Watson, Gary. "Two Faces of Responsibility." *Philosophical Topics* 24, no. 2 (Fall 1996): 227–48.
Westen, Peter K. "Does Duress Justify or Excuse? The Significance of Larry Alexander's Ambivalence." In *Moral Puzzles and Legal Perplexities: Essays on the Influence of Larry Alexander*, edited by Heidi M. Hurd, 76–97. New York: Cambridge University Press, 2019.
Wolf, Susan R. *Freedom Within Reason*. New York: Oxford University Press, 1990.
Yaffe, Gideon. *The Age of Culpability: Children and the Nature of Criminal Responsibility*. New York: Oxford University Press, 2018.
———. *Attempts: In the Philosophy of Action and the Criminal Law.* New York: Oxford University Press, 2010.
Zimmerman, Michael J. *An Essay on Moral Responsibility*. Totowa, NJ: Roman and Littlefield, 1988.
———. "Moral Responsibility and Ignorance." *Ethics* 107, no. 3 (April 1997): 410–26.
———. "Taking Luck Seriously." *Journal of Philosophy* 99, no. 11 (November 2002): 553–76.

Journal of Ethics and Social Philosophy
VOL. 25, NO. 3 · SEPTEMBER 2023
https://doi.org/10.26556/jesp.v25i3.2346

HOW TO BE MORALLY RESPONSIBLE FOR ANOTHER'S FREE INTENTIONAL ACTION

Olle Blomberg

THE THESIS that an agent can be morally responsible and fully blameworthy for another agent's free and intentional action is likely to strike many as either wildly implausible or trivial. On the one hand, it seems right that, as Joel Feinberg emphatically stresses, *"there can be no such thing as vicarious guilt."*[1] One agent's blameworthiness for an action cannot be directly grounded in another's morally objectionable attitudes as opposed to her own.[2] On the other hand, many would acknowledge that an agent can be morally responsible and blameworthy for another agent's free and intentional action if she brings it about that the other performs it. For example, while Marya Schechtman claims that "a person can only be held responsible for her own actions," she immediately footnotes this statement with the qualification that "a person may be held responsible for the action of someone else if she somehow brought it about."[3] In the same vein, John Gardner writes:

> I am responsible for my actions, and you are responsible for yours. My actions are mine to justify or excuse, and your actions are yours to justify or excuse. And yet my actions include my actions of contributing to your actions. So there is a sense in which my responsibility for my actions can extend out to your actions.[4]

I agree. However, I will argue that Gardner's responsibility for his actions can extend to my actions in the same sense that his responsibility for his own basic actions—such as his decisions or bodily movements—can extend to his own

1 Feinberg, "Collective Responsibility," 676. What Feinberg means here by "guilt" is a kind of fault for wrongdoing, not the moral emotion.

2 This does not imply that *our collective* blameworthiness for a joint action or an outcome cannot be grounded in a combination of my attitudes and your attitudes. For my view of collective moral obligations, the violation of which would imply such collective blameworthiness, see Blomberg and Petersson, "Team Reasoning and Collective Moral Obligation."

3 Schechtman, *The Constitution of Selves*, 14, 14n15.

4 Gardner, "Complicity and Causality," 136.

nonbasic actions of bringing about bodily or worldly results. While my actions are indeed mine to justify or excuse, they may also be his to justify or excuse. Hence, I will argue that an agent can be morally responsible and fully (but not necessarily solely) blameworthy for another's free and intentional action in the relevantly same way that she is morally responsible and blameworthy for her own nonbasic actions.

To illustrate what my thesis entails, consider the following case, which I will make use of throughout the paper:

> *Testimony:* Stringer desires and intends Mouzone to be killed. He happens to know that Mouzone murdered Omar's beloved. However, Omar, a notorious stickup man, mistakenly thinks that his beloved's death was the result of an accidental fall from a balcony. Knowing what sort of person Omar is, Stringer knows that if he reveals to Omar the true cause of his beloved's death, then it is very probable (with probability 0.8) that Mouzone will die as a result of Omar deciding to kill him and then carrying out this decision. With intent to bring about Mouzone's death, Stringer reveals to Omar that his beloved was actually murdered by Mouzone. Upon receiving this information, Omar acquires a desire to avenge his beloved's death, but this desire is not irresistible. He freely decides to kill Mouzone, just as Stringer predicted. Omar then tracks down Mouzone, aims a handgun at him, and pulls the trigger. The bullet hits Mouzone, who dies immediately.[5]

In all legal systems of which I am aware, Stringer would be legally off the hook in this case. According to the so-called autonomy doctrine in Anglo-American criminal law, an intervening agent's free and intentional action, such as Omar's killing of Mouzone, breaks "the moral connection" between the first agent's action and its bad or forbidden consequence.[6] But according to my thesis, the moral connection is retained. Stringer can be morally responsible and fully blameworthy for the killing of Mouzone in *Testimony*, just as he is morally responsible and fully blameworthy for the killing of Mouzone in the following case:

> *Lone Killer*: Stringer desires and intends Mouzone to be killed. He tracks down Mouzone and aims a handgun at him. He knows that if he pulls the trigger, then it is very probable (with probability 0.8) that Mouzone

5 The case is loosely inspired by characters and events from season 2 of David Simon's TV series *The Wire*.

6 Williams, "*Finis* for *Novus Actus*?," 392.

will die as a result. With intent to kill, he pulls the trigger. The bullet hits Mouzone, who dies immediately.[7]

The fact that Stringer's agency with respect to Mouzone's death is mediated by an autonomous free agent in *Testimony*, but only by a short-barreled firearm in *Lone Killer*, is not, I claim, itself relevant for Stringer's moral responsibility and blameworthiness for the killing.[8] The difference in mediation is morally relevant in other ways, though. Omar is presumably also morally responsible and fully blameworthy for the killing of Mouzone, whereas the handgun is not morally responsible for anything. (By "fully" blameworthy, I mean unexcused and blameworthy to a degree proportional to the intended and foreseen moral badness of the wrongdoing.) In addition, perhaps Stringer is blameworthy for an additional wrong of corrupting another autonomous agent in *Testimony* by making Omar and not only himself blameworthy for the killing of Mouzone. However, my focus here is solely on the first agent's moral responsibility and blameworthiness for the second agent's intentional action.

My thesis need not imply that Stringer killed Mouzone in *Testimony*. If the meaning of "kill" rules out the involvement of an intermediary agent's intentional action, then Stringer did not kill Mouzone in *Testimony*.[9] But even if Stringer did not kill Mouzone, he can still stand in the moral responsibility relation to the killing (i.e., Omar's killing of Mouzone). Knowing who did it is one thing; knowing who is morally responsible for it is another.[10] "I didn't do it!" is often, but not always, a valid excuse.

Many philosophers of action and moral responsibility explicitly or implicitly deny my thesis.[11] Some would claim that while Stringer is morally respon-

7 While the moral connection is retained in *Testimony*, perhaps there are other reasons for accepting the autonomy doctrine as a legal policy (see section 6).

8 Cf. Bazargan-Forward, "Complicity," 330.

9 For this view of the semantics of "kill," see Davidson, "Agency," 22n18; Gardner, "Complicity and Causality," 134, 137; David Lewis, "Causation," 188; and Ludwig, *From Individual to Plural Agency*, 73. But did not Stalin kill Trotsky, even though it was Ramón Mercader who buried the ice axe in Trotsky's head? If the intermediary agent's action can appropriately be construed as having enabled the first agent to cause the victim's death, then it can arguably truly be said that the first agent killed the victim. (See Wolff, "Direct Causation in the Linguistic Coding and Individuation of Causal Events.")

10 Eric Wiland assumes that for an agent to be morally responsible for an action, the agent must either perform the action himself or genuinely perform it together with others ("(En)joining Others," 65–66). I reject this assumption.

11 They include, e.g., Hywel D. Lewis, "Collective Responsibility"; Sverdlik, "Collective Responsibility" and "Crime and Moral Luck"; Dretske, "The Metaphysics of Freedom"; Frankfurt, "What We Are Morally Responsible For"; Aguilar, "Interpersonal Interactions and the Bounds of Agency"; Ginet, "An Action Can Be Both Uncaused and Up to The

sible for revealing information and evidence to Omar in *Testimony*, only Omar could be morally responsible and blameworthy for killing Mouzone. Some allow that Stringer could be morally responsible for the result of Omar's action—that is, for Mouzone's death.[12] Others allow that he could also be morally responsible for the outcome that Omar killed Mouzone. Some might even allow that Stringer could be just as blameworthy for bringing about the outcome that Omar killed Mouzone in *Testimony* as he would be for shooting and killing Mouzone himself in *Lone Killer*.[13] As I explain in section 4, I do not substantively disagree with such a position. If this is your position, then my argument at least shows that there is no moral significance to the distinction between responsibility for an action and responsibility for the outcome of said action being performed.

Others acknowledge that an agent can be morally responsible and fully blameworthy for another agent's free and intentional action, but not simply by intentionally creating the conditions for the action in a way that causes it. According to David Atenasio, the first agent is only morally responsible for the other agent's action if she has authorized the other to act on her behalf.[14] Relatedly, Eric Wiland argues that an agent can be morally responsible for another's action if the two are engaged in a form of joint agency where the second agent takes direction from the first.[15] With a focus on similar cases, Daniel Story argues that an agent can be morally responsible for another agent's action if the other is acting *directly* on the first agent's intention—an intention that then continuously regulates the other's action.[16] What I will show is that such authorization, special mode of joint agency, or transmission of intention is not necessary for the social extension of moral responsibility for action.

Here I focus on the case where an agent *intends* another agent to perform an action, since I believe that such a case provides the strongest intuitive support

---

Agent"; Fischer and Tognazzini, "The Physiognomy of Responsibility"; Deery and Nahmias, "Defeating Manipulation Arguments"; and Khoury, "The Objects of Moral Responsibility." A denial of my thesis is also at least suggested by Davidson, in "Agency." (See note 56 in section 3 below.)

12 According to Fred Dretske, Stringer could cause and be responsible for Mouzone's death, but *not* for Omar's act of killing (see "The Metaphysics of Freedom"). For a decisive objection to this intriguing view, see McCann, "Dretske on the Metaphysics of Freedom," 622–23.

13 For an explicit defense of this kind of view, see Himmelreich, "Responsibility for Killer Robots."

14 Atenasio, "Co-responsibility for Individualists."

15 Wiland, "(En)joining Others."

16 Story, *Essays Concerning the Social Dimensions of Human Agency*, chs. 3–4. See also Roth, "Entitlement to Reasons for Action."

for my thesis.[17] But, plausibly, an agent can recklessly or negligently bring about another agent's free and intentional wrongdoing in a way that makes him responsible and (less than fully) blameworthy for that wrongdoing, just as an agent can act recklessly or negligently and thereby become responsible and (less than fully) blameworthy for his own future (unwitting) wrongdoing.[18] In addition, I focus on a case where an agent performs a positive action in order to bring about the outcome that another performs an action. I here leave aside cases where the first agent omits to act in order to let the outcome that the other performs an action come about. Furthermore, I focus on cases of socially extended blameworthiness for wrongdoing and leave aside cases of socially extended praiseworthiness for morally exemplary action.

Without this narrowed focus, my argument would be relevant for a wider range of real-world cases, but cases similar to *Testimony* do occur in the real world. For example, people sometimes reveal the identity of convicted criminals or political activists online with the intent that others harass or attack them. Some cases of legal or civil entrapment also resemble *Testimony*, although the first agent does not then simply intend the second agent to commit the wrongdoing but also that he be prosecuted or otherwise exposed for committing it—something that affects the first agent's degree of blameworthiness for the entrapped agent's wrongdoing.[19]

My argument and thesis also have theoretical implications. It helps make sense of how several agents can be jointly blameworthy for a joint intentional wrongdoing or conspiracy, as each of them could be fully morally responsible and blameworthy for the whole intended joint wrongdoing, including both their own intentional contribution and the others' intentional contributions.[20] In section 5, I show how the argument for my thesis undermines an attempt to respond to manipulation arguments that favor incompatibilism about moral responsibility and determinism. Furthermore, as I show in section 6, my argument may have consequences for how to best think about the difference between the legal responsibility and liability of principals and accomplices.

17 On intending that others act, see Bratman, "I Intend That We J"; Ludwig, *From Individual to Plural Agency*, 102–6, 207–10; and Núñez, "Intending Recalcitrant Social Ends."

18 See Smith, "Negligence." For an informative discussion of how an unwitting wrongdoing must be related to a "benighting act" to be traceable to it, see Robichaud and Wieland, "A Puzzle concerning Blame Transfer."

19 On such differences in the degrees to which the agents are blameworthy, see the final paragraph of section 2 below. On entrapment, see Hill, McLeod, and Tanyi, "The Concept of Entrapment."

20 See Blomberg and Hindriks, "Collective Responsibility and Acting Together"; and Ludwig, "From Individual Responsibility to Collective Responsibility."

In section 1, I provide sets of jointly sufficient conditions for moral responsibility and blameworthiness. I hope that most readers will find these jointly sufficient conditions acceptable. I also introduce a crucial distinction between basic and nonbasic moral responsibility. An agent is basically responsible only for that over which he has direct control—his basic actions—such as his decisions or bodily movements. Drawing on work by Carolina Sartorio, I provide principles (sufficient conditions) for how moral responsibility and blameworthiness can then be causally transmitted to outcomes and nonbasic actions of bringing those outcomes about.

In section 2, I present my positive argument: the symmetry argument. I argue that the jointly sufficient conditions for basic and nonbasic moral responsibility and blameworthiness yield the result that, other things being equal, Stringer can be morally responsible and fully blameworthy for the killing of Mouzone in both *Testimony* and *Lone Killer*. There is a perfect symmetry between the cases as far as Stringer's moral responsibility and blameworthiness for the killing are concerned. In both cases, Stringer is basically responsible for a decision or bodily movement (moving his vocal cords, tongue, and lips in *Testimony*; flexing his index finger in *Lone Killer*). He is blameworthy for this basic action in each case because he intended and foresaw that it would causally result in Mouzone's death. Since his basic action in each case did cause Mouzone's death in the way he intended and foresaw, he is in each case fully blameworthy for the killing of Mouzone as well as for the outcome that Mouzone died.[21]

In sections 3–6, I consider and respond to four different objections to this symmetry argument. The first three objections are grounded in ideas about free will, intentional agency, and the kind of control agents have of their own actions when they are morally responsible and fully blameworthy for them. The fourth objection is a normative policy-based objection, based on the autonomy doctrine in Anglo-American criminal law. According to this doctrine, Stringer could not be legally liable at all in *Testimony* for the murder of Mouzone. Whatever the legal justification for the doctrine might be, a moral version of the autonomy doctrine should be rejected. Even if Stringer's act of telling Omar the truth is not as such culpable, Stringer can nevertheless be responsible and fully blameworthy for Omar's killing of Mouzone.

21 Similar arguments have been offered by Moore ("Causing, Aiding, and the Superfluity of Accomplice Liability") and Bazargan-Forward ("Complicity") for the conclusion that a distinct kind of liability for accomplices is superfluous.

1. Moral Responsibility, Blameworthiness, and Action

An agent is morally responsible for an action or outcome if she stands in a relation to that action or outcome such that she would be an appropriate target of blame (praise) for it if it was morally bad (good). In this sense, an agent can be morally responsible not only for wrongdoing or otherwise morally significant actions but also for morally insignificant actions such as, say, drinking a glass of water or putting on a jacket in an ordinary context where such an action lacks moral significance.[22] With that said, for brevity's sake I will from now on use "responsible" and "responsibility" elliptically for "morally responsible" and "moral responsibility." So, the thesis I will be arguing for is that an agent can stand in the responsibility relation not only to her own intentional actions but also to the intentional actions of other agents. When the other agent's action is morally wrong or bad, both agents can be blameworthy for that action. The blame I take the agents to be worthy of here is, paradigmatically at least, moral anger from others and guilt on the part of the agents themselves. This does not mean that blame cannot take other forms, where these other forms are perhaps associated with distinct kinds of responsibility.[23] However, my focus is on the kind of responsibility for wrongdoing that makes an agent an appropriate target of moral anger or guilt in light of the wrongdoing.

Since I would like the argument for my thesis to be compatible with many plausible accounts of the kind of responsibility and blameworthiness that I focus on, I will start by suggesting a set of *jointly sufficient* conditions for such responsibility.

An agent S is responsible for ϕ-ing if

1. S has direct control over ϕ-ing (S freely ϕs);
2. S is aware of what S is doing in ϕ-ing;
3. S is aware of the moral significance (or lack thereof) of ϕ-ing;
4. S has the ability "to feel and understand moral sentiments and reactive attitudes" (such as moral indignation, guilt, gratitude); and
5. S's desires or values that motivated S to ϕ were not acquired by manipulation that bypassed S's reasoning capacities, but rather were acquired in a way that makes those desires or values her own.[24]

22 Here I follow Fischer and Ravizza, *Responsibility and Control*, 8n11; Talbert, *Moral Responsibility*, 1–2; and Sartorio, "Responsibility and Causation," 351–52. Some use "moral responsibility" more narrowly to refer to responsibility for actions, omissions, or outcomes that are morally significant: see McKenna, *Conversation and Responsibility*, 16–17; Vargas, *Building Better Beings*, 307–9; and Mele, *Manipulated Agents*, 4.

23 For an overview, see Jeppsson, "Accountability, Answerability, and Attributability."

24 Russell, "Responsibility and the Condition of Moral Sense," 293. Regarding historical conditions on an agent being the owner or source of her desires or values, see Kane, *The Significance of Free Will*; and Mele, *Manipulated Agents*.

Some of these conditions may not be necessary for *S* to be responsible for ϕ-ing, and perhaps some of them are not fundamental; for example, perhaps 2 is encompassed by 1, or 4 is encompassed by 3. Furthermore, the control or freedom involved in condition 1, as well as the ownership involved in condition 5, can be understood as requiring the ability to do otherwise (regulative control) or as only requiring the ability to guide behavior in a way that is responsive to reasons (guidance control).[25]

Conditions 1–5 are jointly sufficient for *basic responsibility*. We are basically responsible only for actions over which we exercise direct control, where this direct control can be understood as direct regulative control or direct guidance control.[26] Exercising control over a mental action such as making a decision, or over a bodily action such as flexing my right index finger, is normally not done indirectly by means of controlling some other more basic action (unless I flex my right finger indirectly by closing it with my left hand). Instead, we normally directly control these actions. Sartorio thus mentions "choices" as an example of an action that we might have direct control over, and that therefore could be an object of basic responsibility.[27] Randolph Clarke, in describing what he takes to be an attractive and widely held view, also includes bodily movements as possible objects of an agent's direct control and basic responsibility, although he excludes everything beyond the agent's body.[28]

On Donald Davidson's influential account of the nature of actions, which is often assumed within contemporary moral responsibility theory, the view described by Clarke would imply that we can be basically responsible only for our own actions.[29] According to Davidson, all actions are, strictly speaking, "primitive actions"—now more commonly known as "basic actions"—and these are actions that an agent can perform directly, not by means of performing some other action.[30] Davidson thinks that all such basic actions are

25 The terms "regulative control" and "guidance control" are from Fischer and Ravizza, *Responsibility and Control*.

26 See Sartorio, *Causation and Free Will*, 25; and Clarke, *Omissions*, 106–7, and "Responsibility for Acts and Omissions," 94–95. Sartorio calls basic responsibility "direct responsibility," but since "direct responsibility" is used by some philosophers in a way that allows for direct responsibility to overflow direct control (see section 4 below), I prefer "basic responsibility."

27 Sartorio, "Responsibility and Causation," 348; cf. Zimmerman, "Taking Luck Seriously," 564.

28 Clarke, "Responsibility for Acts and Omissions," 95.

29 Davidson's account seems to be assumed by, e.g., Sverdlik, "Collective Responsibility," 65–66, 72; Frankfurt, "What We Are Morally Responsible For," 290–92; and Fischer and Ravizza, *Responsibility and Control*, 82–83, 116.

30 See Davidson, "Agency," 10–11.

bodily movements: "We never do more than move our bodies: the rest is up to nature."[31] Moving one's body must be understood liberally though, to include mental actions such as making decisions.[32] Some philosophers of action restrict basic actions to *tryings*, so that even one's bodily movements turn out to be "up to nature" rather than up to oneself.[33] Either way, a basic action can then be picked out with descriptions that mention or imply its intended or unintended consequences—that is, that mention or imply events that are up to nature. To illustrate, Stringer's flexing of his index finger (or his trying to flex it) in *Lone Killer* could be picked out with the description "Stringer's killing of Mouzone," a description that implies the (intended) consequence that Mouzone dies.

These views are not supposed to capture what people—in a colloquial sense—*do*. When people do things, they typically make changes to the world beyond the movements of their bodies. As Davidson notes about his own view that basic actions are all the actions there are, it may come with a "shock of surprise."[34] But we can make room for what people do in a colloquial sense by allowing that there are nonbasic actions in addition to basic actions. When Stringer kills Mouzone by flexing his right index finger, besides the basic action of flexing his finger being performed, many nonbasic actions, such as the killing of Mouzone, are "generated" as well.[35] Alternatively, perhaps the basic action should be thought of as just one component of the larger nonbasic action.[36] We can also simply use a term other than "action," such as "conduct," to loosely refer to both actions (in the technical Davidsonian sense) and some of the outcomes of those actions in order to capture what people in a colloquial sense do.[37]

Since an agent is not in direct control over her nonbasic actions, she cannot be basically responsible for them. Once Stringer has aimed his handgun and flexed his index finger, he has no control over the immediate consequences of this basic action, which means that he does not have direct control over the nonbasic action of killing Mouzone. Here we have a small pocket of "local fatalism."[38] According to those who deny that there is resultant moral luck—

31 Davidson, "Agency," 23.

32 Davidson, "Agency," 11.

33 See Hornsby, *Actions*; and Khoury, "The Objects of Moral Responsibility."

34 Davidson, "Agency," 23.

35 See Goldman, *A Theory of Human Action*, 23.

36 Weil and Thalberg, "The Elements of Basic Action."

37 See McKenna, *Conversation and Responsibility*, 17.

38 Dennett, *Elbow Room*, 115–17. Of course, Stringer may have indirect control over the basic action's more distal consequences. If Mouzone does not die immediately, then whether he survives or dies from the gunshot wound may depend on whether Stringer calls an ambulance after flexing his index finger.

also known as "consequential" or "outcome" moral luck—an agent can only be responsible for that over which he has direct control.[39] This means that an agent cannot be responsible for his nonbasic actions. Whether the bullet from Stringer's gun actually hits Mouzone depends on many things beyond Stringer's control, such as whether a bird happens to fly by and stop the bullet before it reaches Mouzone. Hence, one might think that it would therefore be wrong to blame him for anything beyond what he directly controls, and wrong to adjust the degree to which he is deemed blameworthy in light of what is up to nature.[40]

Our practice of holding each other responsible for what we do—for our "conduct"—does make room for responsibility for nonbasic actions as well as outcomes. We do not hold each other responsible only for our tryings or bodily movements. In *Lone Killer*, we might not know what bodily movement Stringer made to bring about Mouzone's death, and even if we did know, the movement would not be our focus in holding him responsible and blaming him for killing Mouzone. Perhaps Stringer squeezed the trigger with his middle or ring finger rather than with his index finger, or perhaps he did not shoot Mouzone but instead stabbed or poisoned him (or indeed, perhaps he moved his lips and led someone else to kill him). While Stringer presumably moved his body in some way when he killed Mouzone, the bodily movement is not the primary object of blameworthiness. If it were, then the fact that we pick it out with a verb that implies the particular consequence that Mouzone died would just be a matter of convenience. We could pick it out by any of infinitely many alternative descriptions that do not imply that Mouzone died. But our focus and the object of blameworthiness is *the killing of Mouzone*, where this includes Mouzone's death. So, assuming a Davidsonian view of action, when we hold an agent responsible for what he "does," we typically primarily hold him responsible for an outcome that he brought about.

In other words, when we hold each other responsible for what we have done, we typically hold each other responsible for what is partly up to nature. I think that our practice of holding people responsible ought to be this way. However, it is beyond this paper's scope to argue for resultant moral luck. I will simply assume that such moral luck should be accepted.[41] Without it, an agent could

39 For an overview, see Nelkin, "Moral Luck."

40 Such an antiluck view is endorsed by, e.g., Sverdlik, "Crime and Moral Luck"; Frankfurt, "Three Concepts of Free Action," 123, and "What We Are Morally Responsible For," 290–93; and Khoury, "The Objects of Moral Responsibility." Sverdlik and Frankfurt take actions to be bodily movements, while Khoury identifies them with tryings.

41 My view is that resultant moral luck affects both what agents are blameworthy for—the "scope" of blameworthiness—and the degree to which they are blameworthy for it. For defenses of such a view, see Hartman, *In Defense of Moral Luck*; and Lang, *Strokes of Luck*.

not be responsible for *anything* beyond the boundary of his own body or will.[42] My thesis would thus be excluded from the get-go.[43] My goal is thus to convince those who accept resultant moral luck to accept my thesis. [44]

How is an agent's basic responsibility for decisions or other basic actions extended to nonbasic responsibility for outcomes and the nonbasic actions of bringing those outcomes about?[45] The following principle provides a plausible *sufficient* condition for nonbasic responsibility for outcomes and nonbasic actions:

For the view that resultant moral luck only affects the scope of blameworthiness, see Zimmerman, "Taking Luck Seriously."

42 Even without resultant moral luck, there need not be any significant asymmetry with respect to Stringer's responsibility and blameworthiness between *Lone Killer* and *Testimony*. Stringer's trying or bodily movement could in both cases be picked out with the description "Stringer's killing of Mouzone" (or "Stringer's killing* of Mouzone," where "kill*" is like "kill" except that it allows for the involvement of an intermediary agent's intentional action; see note 9 above). In *Lone Killer*, this description would pick out the flexing of his index finger; in *Testimony*, it would pick out the movements of his vocal cords, tongue, and lips. Since Omar directly controlled his basic action with the intention of killing/killing* Mouzone in both cases, and his basic action caused Mouzone's death as intended, one could argue that it would be appropriate to hold Stringer responsible for his action under the description "his killing/killing* of Mouzone" in both cases (see Khoury, "The Objects of Moral Responsibility," 1365–66). However, Stringer could just as well be held responsible for this action under the description "trying to kill/kill* Mouzone by moving his body," since whatever happens beyond Omar's direct control is supposed to be irrelevant. Because of this, antiluckists David Enoch and Andrei Marmor argue that "what we need is a reason to hold Brian [Stringer] morally responsible for his reckless drunken driving [trying to kill/kill* Mouzone] under the description of a killing [/killing*]. . . . But any such reason will just be a reason to acknowledge moral luck" ("The Case against Moral Luck," 411). Either way, *Lone Killer* and *Testimony* could be symmetrical with respect to Stringer's moral responsibility and blameworthiness.

43 That is, unless one agent's basic action can have another agent's basic action as a part (see Ford, "The Province of Human Agency," 715–16) or be identical to it (see Blomberg, "Socially Extended Intentions-in-Action").

44 Denial of resultant moral luck arguably motivated Hywel D. Lewis's widely quoted rejection of the possibility of responsibility for another agent's action: "If I were asked to put forward an ethical principle which I considered to be especially certain, it would be that no one can be responsible, in the properly ethical sense, for the conduct of another" ("Collective Responsibility," 3). Several passages, including the following, suggest that Lewis would not accept that an agent can be responsible for outcomes or even for behavior: "We want to be sure that our estimation of [a person's] moral worth is not prejudiced by considerations relating only to outward action" (4).

45 See Clarke, "Responsibility for Acts and Omissions," 94–96; and Sartorio, "Responsibility and Causation," 348–55. Clarke uses the terms "indirect responsibility" or "derivative responsibility," and Sartorio uses "derivative responsibility." But some philosophers tie these terms to so-called tracing cases (see section 4). Nonbasic responsibility covers not only responsibility in such tracing cases.

> *Intended Causal Transmission of Responsibility*: If *S* is responsible for ϕ-ing, and the ϕ-ing caused outcome *C*, and *S* intended and foresaw that the ϕ-ing would (or was likely to) cause *C* in roughly the way that the ϕ-ing did cause *C*, then *S* is responsible for *C* and for bringing *C* about.[46]

If responsibility for outcomes or nonbasic actions is accepted at all, I take it that this principle is relatively uncontroversial.[47] Since Stringer in *Lone Killer* is basically responsible for flexing his index finger (or for deciding to do so), and since this bodily movement (or decision) caused Mouzone to die in roughly the way that Stringer intended and foresaw, he is nonbasically responsible for the outcome that Mouzone dies.[48] Allowing for nonbasic actions, Stringer would also be nonbasically responsible for killing Mouzone, since Stringer is nonbasically responsible for bringing about Mouzone's death in a way that amounts to killing him.

Under what conditions is an agent *blameworthy* for what she is basically responsible for? To get plausible jointly sufficient conditions for *S* to be blameworthy for the ϕ-ing, we need to add the following two conditions to our previous conditions 1–5:

6. the ϕ-ing is morally wrong; and
7. the ϕ-ing manifests *S*'s "ill will or indifference or lack of concern" toward others or toward morality.[49]

46 The principle is adapted from Sartorio's principle "S" (Responsibility and Causation," 349–51). S only includes the condition that the agent foresees the outcome, not that the agent also intends the outcome to be a result of the action. My principle explicitly includes a clause meant to exclude cases of overly deviant causation, where the intended and foreseen outcome does occur, but not at all in the way that the agent intended or foresaw.

47 Davidson suggests that "we may indeed extend responsibility or liability for an action to responsibility or liability for its consequences … by pointing out that his original action had those results" ("Agency," 23). Likewise, Fischer and Ravizza, when discussing a case where "Sam is morally responsible for his action of shooting and killing the mayor," submit that "it seems very plausible to say that Sam can also fairly be held morally responsible for the consequence, *that the mayor is shot*" (*Responsibility and Control*, 93).

48 An agent is not basically responsible for his primitive action by causing it: "Doing something that causes my finger to move does not cause me to move my finger; it *is* moving my finger" (Davidson, "Agency," 11). See also Sartorio, *Causation and Free Will*, 25–26. This does not rule out the possibility that the agent's earlier action, such as his decision to later move his finger, caused him to later move his finger. The agent would then be responsible for the movement both basically and nonbasically.

49 Strawson, "Freedom and Resentment," 199. Michael McKenna argues that an agent being responsible for a moral wrongdoing is insufficient for her being blameworthy for it, since the wrongdoing must also manifest her bad "quality of will" (*Conversation and Responsibility*, 19–20).

In *Lone Killer*, other things being equal, Stringer's basic action of flexing his index finger is morally wrong. It is morally wrong because it causes Mouzone's death in a way that Stringer intended and foresaw. Stringer's flexing of his index finger also manifests his ill will toward Mouzone. For an agent's "ill will or indifference or lack of concern"—her "bad quality of will"—to be manifest in a decision or basic action, it simply needs to be rationalized and caused by intentions, desires, or beliefs about reasons for action that are morally objectionable. Conditions 6 and 7 are thus satisfied in *Lone Killer*: Stringer is blameworthy for flexing his finger.

Just as responsibility for a basic action can be transmitted by causation to intended and foreseen outcomes of the basic action, so can blameworthiness:

> *Intended Causal Transmission of Blameworthiness:* If *S* is blameworthy for ϕ-ing partly or wholly because *S* intended and foresaw that the ϕ-ing would likely causally result in *C*, and the ϕ-ing resulted in *C* in roughly the way *S* intended and foresaw, then *S* is blameworthy for *C* and for bringing *C* about.[50]

This principle closely mirrors the intended causal transmission of responsibility principle. But what does it mean for an agent to be blameworthy for an outcome? It means that the outcome, and not only the basic action that causes it, manifests the agent's bad quality of will. The term "manifest" may misleadingly suggest that in order for an agent to be blameworthy for an action, the agent's morally objectionable intentions, choices, or judgments about reasons must be publicly expressed and on full display in the action. But this would not be a plausible requirement since an agent can be blameworthy for a wrongdoing or a morally bad outcome while hiding his morally objectionable motivations and aims.[51] It is sufficient if the agent's morally objectionable intentions, choices, or judgments about reasons nondeviantly cause and rationalize the wrongdoing. Since Stringer is blameworthy for flexing his index finger because he intended and foresaw that it would cause Mouzone's death, and this basic action did cause Mouzone's death in roughly the way Stringer intended and foresaw, he is also blameworthy for the outcome that Mouzone died and for bringing this outcome about. The fact that Stringer nondeviantly brought about Mouzone's death means that this result, and not only Stringer's flexing of his index finger, manifests Stringer's bad quality of will. This upshot, I take it, will accord with most people's intuitions.

50 The principle is adapted from Sartorio's "Principle of Derivative Blameworthiness" (*Causation and Free Will*, 77).

51 Cf. McKenna, *Conversation and Responsibility*, 92–94.

2. THE SYMMETRY ARGUMENT

Stringer can be responsible and fully blameworthy for the killing of Mouzone in *Testimony*, just as he is in *Lone Killer*. Setting aside wholesale skepticism about basic responsibility or morality, conditions 1–7 can in both cases be satisfied with respect to Stringer's basic action. In both cases, there is intended causal transmission of both responsibility and blameworthiness such that Stringer is nonbasically responsible for the killing of Mouzone. Since I have already used *Lone Killer* to illustrate the conditions and principles in the previous section, I will here focus on Stringer's responsibility and blameworthiness for the killing in *Testimony*.

Stringer can be basically responsible for moving his vocal cords, tongue, and lips when he reveals the truth to Omar. If he decided to do this freely, is aware of what he is doing, is aware of its moral significance, and so on, then he is basically responsible for this basic action. Partly because he intends and foresees that this basic action will bring about Mouzone's death by causing Omar to kill Mouzone, he is also blameworthy for performing the basic action. If Stringer himself is a lousy shooter, then the probability that the movements of his vocal cords causally result in Mouzone's death in *Testimony* may be just as high or higher than the probability that the flexing of his index finger causally results in Mouzone's death in *Lone Killer*.

In light of the intended and foreseen causal connection between Stringer's morally objectionable reasons for action and intention and Omar's intention and action, Omar's killing of Mouzone manifests Stringer's (as well as Omar's) ill will toward Mouzone.[52] Recall that a wrongdoing manifests an agent's ill will if the wrongdoing was nondeviantly caused and rationalized by the agent's morally objectionable intentions, choices, or judgments about reasons. Given that Stringer's action nondeviantly caused Mouzone's death by causing Omar to kill Mouzone in the way Stringer intended and foresaw, Stringer is responsible and fully (but not solely) blameworthy for Omar's killing of Mouzone. What makes Stringer responsible and blameworthy here has parallels in the case of *Lone Killer*; it is just that there is in *Lone Killer* only a bullet fired from a gun and not also a free intentional action that mediates and transmits responsibility and blameworthiness from Stringer's executed intention to Mouzone's death. In both cases, Stringer is nonbasically responsible and blameworthy for the killing of Mouzone. There need be no differences between *Lone Killer* and *Testimony* relevant for Stringer's responsibility and blameworthiness for the killing.

52 As David Shoemaker puts it, an action can be "overdetermined" by the wills of several agents ("Responsibility without Identity," 123–24).

That concludes the positive argument for my thesis. In the rest of the paper, I will respond to various objections to the argument. But before turning to the first objection, let me clarify that I take each agent to be blameworthy on the basis of their own quality of will, insofar as that quality of will is manifest in the wrongdoing.[53] This means that the degree of blameworthiness for the wrongdoing can differ between the first and the second agent, along with the quality of their wills, in a case such as *Testimony*. If Omar kills Mouzone in an especially brutal way, but Stringer had no reason to believe that the killing he caused would be especially brutal, then the brutality does not manifest Stringer's quality of will. But if Stringer intended and foresaw that the killing would be brutal, then it does manifest his quality of will. If Stringer intends the killing to be brutal, but Omar instead kills Mouzone quickly and painlessly while Mouzone is asleep, then the killing does not manifest Stringer's full degree of bad will, and the remainder of the bad will left out of the action would not add to Stringer's degree of blameworthiness for the killing.[54] Note also that intentionally causing someone else to perform an action in a foreseeable way and actually performing an action will often require different skills and efforts as well as virtues or vices. These are factors that may be relevant for our assessment of the agent in light of the action. But as far as blameworthiness for the wrongdoing as such goes, the agent that intentionally causes another to do wrong would be responsible for it in the relevantly same way that he is responsible for his own nonbasic actions. No special grounds of responsibility and blameworthiness for the actions of others are needed.

3. First Objection: Free Intentional Actions Cannot Be Caused

Some might object that Stringer's action could not cause Omar to kill Mouzone, at least not if Omar's decision to kill Mouzone was up to him—that is, if his killing of Mouzone was a free intentional action.[55] If true, then this would of course also mean that Stringer could not correctly foresee that his action would *cause* Omar to freely kill Mouzone. This would also mean that Omar's action of killing Mouzone could not manifest Stringer's ill will toward Mouzone. Successful manifestation requires that the manifesting action causally depend on

53 An exception where this might not be true is a "fission" case where a postfission successor is blameworthy for the prefission predecessor's wrongdoing on the basis of the prefission predecessor's quality of will, rather than on the basis of their own quality of will. See Shoemaker, "Responsibility without Identity." For the contrary view, see Köhler, "Moral Responsibility without Personal Identity?"

54 Cf. Markovits, "Acting for the Right Reasons," 209–14.

55 See, e.g., Ginet, "An Action Can Be Both Uncaused and Up to the Agent."

the quality of will manifested. In other words, assuming that no noncausal principle of responsibility- and blameworthiness-transmission is applicable in *Testimony*, Stringer would not, on this view, be responsible and blameworthy for Omar's killing of Mouzone.

In response to this, the first thing to say is that this sort of noncausal libertarian view is *prima facie* implausible.[56] It becomes difficult to make sense of ordinary social interaction such as the asking for and giving of directions without such social interaction involving agents performing actions that deterministically or indeterministically cause intentional responses performed by interlocutors. There are, in general, good reasons for thinking that one agent can cause as well as causally control another's free intentional action.[57] There may be good reasons for thinking that one could not cause another's intentional action in such a way that the other agent could not avoid performing it, but that is a different matter.[58]

On any plausible libertarian view, the fact that it is up to an agent T (the second agent; Omar) whether to ψ (kill Mouzone) should not exclude the possibility that S's ϕ-ing (Stringer's act of assertion) could be a nondeterministic cause of T's ψ-ing, which is possible as long as it is up to T to allow S's ϕ-ing to become, or prevent it from becoming, a cause of T's ψ-ing.[59] If Stringer puts deadly poison in Mouzone's food, then it may be up to Omar, who has the antidote, to allow Stringer's act to become, or prevent it from becoming, a cause of Mouzone's death. Similarly, if Stringer tells Omar that Mouzone murdered his beloved, then it may be up to Omar, who has the gun, to allow Stringer's act to become, or prevent it from becoming, a cause of Mouzone's death. Hence, even if the control involved in condition 1, or the ownership involved in condition 5, requires S to have libertarian free will, Stringer could nevertheless be responsible and fully blameworthy for Omar's free and intentional killing of Mouzone. After all, Stringer could still have the ability to foresee Omar's decision to kill Mouzone. Given that Stringer knows what sort of person Omar is—what his

56 Dretske, who is a compatibilist, also argues that "when the actions are intentional, the causal buck—and, therefore, the responsibility—stops at the [intermediary] actor" ("The Metaphysics of Freedom," 8; see also note 12 above). Davidson ("Agency," 16n10) suggests that it "could be said" that the transitivity of causality breaks down in cases where an intermediary agent intentionally brings about the result of an agent's action, but this is best interpreted as a pragmatic point about our ordinary use of "cause." See also Hart and Honoré, *Causation in the Law*, 42–44.

57 See Feinberg, "Causing Voluntary Actions"; Dennett, *Elbow Room*, ch. 3; and Capes, "Freedom with Causation."

58 See Alvarez, "Actions, Thought-Experiments and the 'Principle of Alternative Possibilities,'" 67, 72.

59 See Capes, "Freedom with Causation."

fears, aims, and values are—Stringer could make an informed and reasonable judgment regarding how Omar will react to the information he is about to receive, even if Omar's reaction is genuinely up to him. It might be that in order for condition 5 to be satisfied, Omar must have faced a "torn" decision, where he made an undetermined "self-forming willing" such that it would have been impossible in advance to foresee or assign a higher than 50 percent probability to his choice, but that does not mean that Omar's killing of Mouzone needs to be the direct result of such a torn decision.[60] At most, it must be the result of desires or values that are in part the result of such torn decisions in the past.

4. Second Objection: "Second-Class" Responsibility

One might object that the symmetry argument merely shows that an agent can be responsible and blameworthy for the outcome that another agent performs an action. It does not show that the agent can be responsible and blameworthy for the other's action itself. On the Davidsonian view, according to which there are only basic actions, Stringer's responsibility for his own killing of Mouzone in *Lone Killer* (i.e., for his bodily movement that is describable as his killing of Mouzone) will be an instance of responsibility for an action, while his responsibility with respect to Omar's killing of Mouzone in *Testimony* will be an instance of responsibility for an outcome, namely, the outcome that Omar kills Mouzone.[61] Stringer is not responsible for Omar's bodily movement itself, which is describable as Omar's killing of Mouzone. According to this objection, an agent could thus not be directly responsible for another agent's action in the same way that the other agent herself is. This conclusion can also be reached from other accounts of what actions are. According to Alvin Goldman, Stringer's basic action of flexing his right index finger in *Lone Killer* "causally generates" his nonbasic action of killing Mouzone; it does not *cause* it.[62] (If Stringer took an electric scooter to find Mouzone before killing him, and pressed down the accelerator button with his right index finger, then the flexing of his index finger could be a cause of his later action of killing Mouzone, but this would be a different case.) Indeed, if one action causes another, then it cannot causally generate it, and *vice versa*.[63] Similarly, on a componential view of action, the basic action cannot cause the nonbasic action because the former is a part of

60 On torn decisions and self-forming willings, see Kane, *The Significance of Free Will*, ch. 8.

61 See Aguilar, "Interpersonal Interactions and the Bounds of Agency," 228–31.

62 Goldman, *A Theory of Human Action*, 23.

63 Goldman, *A Theory of Human Action*, 23.

the latter.[64] Again, this shows that an agent is not responsible for her own nonbasic action (by performing a basic action that causally generates or is part of this action) in the same way that she is responsible for the other's action or for her own later action (by causing it).

However, these action-theoretic distinctions do not show that there is a difference between *Lone Killer* and *Testimony* such that Stringer stands in different kinds of responsibility relations to the killing of Mouzone in the two cases, nor such that Stringer is more blameworthy for the killing in the former case than in the latter. There is something like a mundane distinction between responsibility for an agent's action and responsibility for outcomes that are not part of an agent's action, and this distinction is morally significant. This is because the mundane distinction is typically used to distinguish intentional wrongdoing from recklessness or negligence. When we explicitly hold an agent responsible for an outcome, the outcome is typically a result of the agent's recklessness or negligence. Suppose that Mouzone has an inept bodyguard, called Lamar, who fell asleep at his post, resulting in Mouzone being killed. We can then imagine someone saying to Lamar, "Mouzone is dead, and it's your fault!" When we instead hold an agent responsible for an action (in a nontechnical sense), the agent has typically intended to produce the bad outcome. While the mundane distinction is morally significant then, we should not be misled into thinking that the superficially similar action-theoretic distinction between responsibility for an action and responsibility for an outcome is similarly morally significant.

The objection to the symmetry argument can also be put in terms of "direct" and "indirect" responsibility for action. In discussions of complicity, it is said that an agent's responsibility for his own actions is "direct," whereas his responsibility for another agent's actions is, at most, "indirect."[65] In discussions about individual moral responsibility, there is a parallel intra-agential distinction tied to so-called tracing cases, where an agent is indirectly responsible and blameworthy for an action even though he does not satisfy the conditions for basic responsibility, but where his blameworthiness can be traced back to, or inherited from, some earlier reckless, negligent, or malicious action for which he is directly responsible.[66]

64 E.g., Weil and Thalberg, "The Elements of Basic Action."

65 E.g., Gardner, "Complicity and Causality," 136.

66 See, e.g., McKenna, *Conversation and Responsibility*, 15–16, 188, 191; Levy, *Consciousness and Moral Responsibility*, 3; Mele, *Manipulated Agents*, 10–11; and Vargas, *Building Better Beings*, 34–35. Vargas uses the terms "original responsibility" and "derivative responsibility." McKenna occasionally also uses this latter term.

In tracing cases, an agent is responsible for his own action partly or wholly by virtue of being responsible for an earlier action of his that causes the later action in a foreseeable way. At the time of the later action, he fails to satisfy either the control condition 1 or the epistemic condition 2 and cannot therefore be directly responsible for that action. But since responsibility can be traced back to his earlier blameworthy action, he is nevertheless blameworthy also for the later action. We can think of *Testimony* in an analogous way, where the earlier action is Stringer's action of revealing to Omar that his beloved was murdered by Mouzone, and the later action is Omar's killing of Mouzone.[67] Omar, of course, meets all conditions for being responsible and fully blameworthy for the action of killing Mouzone. When it comes to Stringer, he intentionally causes Omar's killing of Mouzone in a way that he can foresee, but he does not satisfy what is normally a plausible personal identity condition on responsibility for action—"I didn't do it!" is typically a valid excuse. However, his responsibility and blameworthiness for the killing can in this case be traced back to his responsibility and blameworthiness for the earlier action of revealing the truth to Omar. Moreover, since Stringer performs the earlier action with the intention that Omar kill Mouzone, he is (just like Omar) responsible and fully blameworthy for the killing.[68]

Is it significant that Stringer's responsibility for the killing is direct in *Lone Killer* but indirect in *Testimony*? The distinction between direct and indirect responsibility is different from my and Sartorio's and Clarke's distinction between basic and nonbasic responsibility. Unlike basic responsibility, direct responsibility overflows direct control: an agent can be directly responsible for nonbasic actions and perhaps also for negligence. To illustrate the distinction between direct and indirect responsibility, Mele considers a case where an agent called Don intentionally illuminates a room by flipping a switch, knowing that the room's becoming illuminated "is a signal for his accomplices to perform a dastardly deed."[69] Since signaling to his accomplices "is not a basic action, he

67 Holly M. Smith considers a case where a doctor negligently fails to update a colleague on a recent finding that the traditional treatment for premature infants has a harmful side effect ("Negligence," 3). The colleague uses the treatment on an infant who is harmed. Smith submits that the doctor is blameworthy for this harm, but it is equally true that the doctor is blameworthy for the colleague's action of using the treatment.

68 Fischer and Ravizza present tracing as a component of their account of responsibility for (basic) actions rather than of their account of responsibility for outcomes (*Responsibility and Control*, 49–51). If their drunk driver's responsibility for killing a pedestrian is an instance of responsibility for action in addition to responsibility for *the outcome that he performs that action*, then Stringer, arguably, could be responsible for Omar's killing of Mouzone and not only for the outcome that Omar kills Mouzone.

69 Mele, "Direct versus Indirect," 571.

does not exercise direct control regarding it," but he is nevertheless responsible for it, as Mele puts it, "in a first-class way."[70] Mele motivates the distinction between direct and indirect responsibility as follows:

> What motivates appeals to indirect moral responsibility are reasonable judgments that agents are morally responsible for some of their actions in a second-class way. Actions for which agents are indirectly morally responsible are said to inherit (some of) their moral responsibility from actions for which the agent is morally responsible in a first-class way. Recall the drunk driver, for example. He has first-class moral responsibility for some action or actions that preceded the crash and second-class moral responsibility for killing the pedestrians, and his moral responsibility for the killings is said to be inherited from his moral responsibility for the pertinent earlier actions.[71]

Mele does not elaborate on what is implied by responsibility being "first class" or "second class," but a natural reading is that Mele is suggesting that it is worse (in terms of degree of blameworthiness) for an agent to be responsible for a wrongdoing in a first-class way than it is for her to be responsible for it in a second-class way.[72] If this were correct, my thesis would be false. Stringer would be responsible in a first-class way (directly) for the nonbasic action of killing Mouzone in *Lone Killer*, but he would only be responsible in a second-class way (indirectly) for Omar's killing of Mouzone in *Testimony*. But I am arguing that *Testimony* illustrates that an agent can be responsible and blameworthy in a first-class way for another agent's free and intentional action, even though the first agent's responsibility for this action is wholly inherited from his direct responsibility for his own action of influencing the other agent.

However, the drunk driver's responsibility for the killing of the pedestrians is second class in this sense not because it is inherited from direct responsibility for another action but because it is the upshot of the driver's recklessness or negligence rather than a malicious intent. If the driver got drunk because he desired and intended to drive out of control through the streets and kill pedestrians, then his responsibility for later killing them would be first class. Similarly, Stringer can be responsible for Omar's killing in a first-class way, despite his responsibility for this action being inherited, since the killing is the result of Stringer's malicious intent rather than his recklessness or negligence.

70 Mele, "Direct versus Indirect," 570.

71 Mele, "Direct versus Indirect," 570–71.

72 Mele has clarified that this was not the reading he intended (personal communication).

On another natural reading of the quoted passage, the "reasonable judgments" that Mele refers to are action-theoretic judgments about what falls "inside" and "outside" the boundaries of an action, rather than judgments about degrees of blameworthiness. While it is reasonable to judge that Don's moving of his body generates, or is part of, his nonbasic action of signaling to his accomplices, it is less reasonable to judge that the drunk driver's moving his body (when drinking too much) causally generates, or is part of, a nonbasic action of crashing into and killing the pedestrians. In the former case, we have an extension of Don's direct responsibility on the basis of a part-whole relation within the same complex action, or a generation relation within one and the same "act-tree."[73] In the latter case, we instead have an inheritance relation between Don's responsibility for two separate actions or act-trees. Similarly, in *Testimony*, there is a relation of inheritance from Stringer's (indirect and second-class) responsibility for Omar's killing of Mouzone to his (direct and first-class) responsibility for his own testimony. But in *Lone Killer*, there is a relation of causal generation, or a part-whole relation, such that Stringer is responsible (directly and in a first-class way) for killing Mouzone.

These are indeed reasonable judgments regarding the extensions of different agents' nonbasic actions. However, they do not show that there is a difference between Stringer's responsibility relation to the killing in *Lone Killer* and his responsibility relation to the killing in *Testimony*, nor do they support the claim that he is more blameworthy for the killing in the former case than in the latter.

5. Third Objection: The Other's Action Is Caused Too Sensitively

Lone Killer and *Testimony* appear to differ in causal structure in a way that may seem relevant for the kind of control that Stringer has over the outcome that Mouzone is killed, and hence, one might think, for his responsibility for that outcome. According to Fred Dretske, as well as Marius Usher, this difference would explain why Stringer kills Mouzone in *Lone Killer* but not (allegedly) in *Testimony*.[74] Given an intimate connection between robust causal control and responsibility and the degree of blameworthiness, the difference in causal structure would also support the view that while Stringer is responsible and blameworthy for the killing of Mouzone in *Lone Killer*, he could not be responsible and blameworthy for (Omar's) killing (of) Mouzone in *Testimony*—at least not to anything like the degree to which Omar is responsible and blame-

73 On the notion of an act-tree, see Goldman, *A Theory of Human Action*, ch. 2.

74 See Dretske, "The Metaphysics of Freedom"; and Usher, "Agency, Teleological Control and Robust Causation."

worthy for the killing.[75] Similarly, Oisín Deery and Eddy Nahmias argue that an agent is responsible for a result only if he is the *causal source* of it.[76] An agent is the causal source of the result if and only if, roughly, his behavior is the prior event that, among all prior events, most robustly causes it. Since a second autonomous agent's behavior is caused by a holistic web of beliefs, desires, and other mental states that are continuously tracking and adapting to changes in the world, the first agent's intention will typically not be the most robust cause of the second agent's action. Rather, the most robust cause will typically be the second agent's own intention. Because of this, the first agent cannot be responsible and fully blameworthy for the second agent's action. In effect, the transmission of the first agent's responsibility and blameworthiness along the line of intended causation is blocked by the second agent's intention.

Dretske and Usher draw on David Lewis's idea that the dependence between a cause and an effect can be more or less insensitive/robust. According to Lewis, C_1 causes E_1 more insensitively/robustly than C_2 causes E_2 if the range of nearby possible worlds in which C_1 causes E_1 is wider than that in which C_2 causes E_2.[77] Dretske uses this notion of insensitive causation to argue that "the special kind of causal dependency required to make an action (e.g., killing) out of a causal relation (causing someone's death) is … an insensitive causal dependence."[78] Similarly, Usher takes the kind of control that responsible agents have over their actions to be such that their intentions insensitively cause their intended effects.[79] While I think that intermediary autonomous agents are compatible with insensitive causal dependence, it is true that they often introduce a significant measure of sensitivity.[80] In *Testimony*, since Omar is an autonomous agent rather than just a tool such as a handgun, there will probably be many nearby possible worlds where Stringer's action of revealing the truth to him would not result in Mouzone's death. Omar plausibly desires many things besides Mouzone's death, and the acquisition of some new information could easily change his behavior so that he would not kill Mouzone (say, if he spotted a police car outside Mouzone's house). Given that this sort of insensitive causal relation between an agent and an event would be required not only for the agency relation but also for the responsibility

75 See Usher, "Agency, Teleological Control and Robust Causation," 309–12.

76 Deery and Nahmias, "Defeating Manipulation Arguments."

77 See David Lewis, "Causation"; and Woodward, "Sensitive and Insensitive Causation."

78 Dretske, "The Metaphysics of Freedom," 11. Lewis himself discusses sensitivity of causation in relation to killing and causing death, but he is more cautious in his conclusions than Dretske.

79 Usher, "Agency, Teleological Control and Robust Causation," 308–9.

80 Cf. Usher, "Agency, Teleological Control and Robust Causation," 318.

relation, Dretske's argument could be interpreted as an argument against the idea that an agent can be responsible for the result of another agent's free and intentional action.

However, while intermediary autonomous agents often introduce this kind of sensitivity, they do not always do so. Omar's disposition to take revenge on anyone who hurts him and his loved ones may be so strong and stable that he will robustly cause the death of Mouzone. Agreements and hierarchical authority relations also normally make for robust causation (as well as foreseeability) through intermediary autonomous agents. Furthermore, note that the causal chain between Stringer's intention to have Mouzone killed and the death of Mouzone need not be sensitive even if the particular chain that runs through Omar's free intentional agency is sensitive. Suppose that Stringer is determined to get Mouzone killed come what may, so that if turning Omar's vengefulness against Mouzone were to fail, then Stringer would take his own gun and himself shoot Mouzone, effectively turning the case into *Lone Killer*. Admittedly, this might only make Stringer's intention a robust cause of Mouzone's death, without necessarily making it a robust cause of Omar's killing of Mouzone.[81]

More importantly, there can be sensitive causal relations, with or without intermediary agents, that do not preclude responsibility for killing. Consider the following writing prompt from the website Reddit: "You are a serial killer who uses Rube Goldberg Machines to kill his victims."[82] For an extreme example of responsibility for another agent's action through a sensitive causal chain involving an intermediary agent, consider also Mele's well-known zygote case, where the "supremely intelligent being"

> Diana creates a zygote *Z* in Mary. She combines *Z*'s atoms as she does because she wants a certain event *E* to occur thirty years later. From her knowledge of the state of the universe just prior to her creating *Z* and the laws of nature of her deterministic universe, she deduces that a zygote with precisely *Z*'s constitution located in Mary will develop into an ideally self-controlled agent [called Ernie] who, in thirty years, will

81 Multiple potential intermediary agents may also provide a robust causal relation between the first agent and the victim's death. Cf. Tierney and Glick, "Desperately Seeking Sourcehood," 960.

82 Rube Goldberg was an American cartoonist who drew complex contraptions that were designed to perform a simple task in an indirect and complicated way, through a very sensitive causal chain.

> judge, on the basis of rational deliberation, that it is best to *A* and will *A* on the basis of that judgment, thereby bringing about *E*.[83]

Mele is interested in what this case suggests about Ernie's responsibility, or lack thereof, for the *A*-ing. But my interest is rather in Diana's responsibility for the *A*-ing. Suppose that the *A*-ing here is "killing Mouzone." Given that Diana is sane and morally competent, Diana would arguably be responsible for Ernie's killing of Mouzone. Mele agrees.[84] She would be responsible for the killing despite the extremely sensitive causal chain that runs from her intention to Ernie's *A*-ing. (If the sort of "manipulation" involved in Mele's case undermines *Ernie's* freedom and responsibility for *A*-ing, then the case does not directly support my thesis, but my point here is just to show that responsibility for action is compatible with extreme sensitivity of causation.)

Plausibly, the background conditions had to be exactly right for Diana to successfully get Ernie to kill Mouzone. Being supremely intelligent, Diana has Laplace's demon-like knowledge and predictive powers that enable her to exploit this unique opportunity the universe provides her with. On this reading of the case, Diana would nevertheless be responsible and fully (but perhaps not solely) blameworthy for Mouzone's death. But on Deery and Nahmias's theory, she would only be responsible for bringing about Mouzone's death if her behavior was the event, among all events prior to his death, that most robustly caused it.[85] While Diana would be *a* cause of his death, only Ernie would be a causal source of it, and hence only Ernie would be responsible and fully blameworthy for killing Mouzone. According to Deery and Nahmias, Diana would be "merely getting lucky" in causing the wrongdoing in my reading of Mele's case.[86] But this is a mistake. While it is true that Diana would be circumstantially lucky to get the opportunity to modify a zygote to become an agent who performs her desired action thirty years later, she can nevertheless settle that this action is performed once she, thanks to her vast knowledge and awesome predictive powers, becomes aware of this fortunate opportunity. Since Diana knows which possible world is the actual world, she has no need for robust causation.

83 Mele, *Free Will and Luck*, 185, 188.

84 Mele, *Free Will and Luck*, 198n16. See also Sartorio, *Causation and Free Will*, 167–69.

85 Or, as Deery and Nahmias would put it, if Diana's behavior bore "the strongest causal invariance relation to [Mouzone's death] among all the prior causal variables" ("Defeating Manipulation Arguments," 1263).

86 Deery and Nahmias, "Defeating Manipulation Arguments," 1273.

In retelling Mele's zygote case, Deery and Nahmias subtly modify it in ways that suit their theory.[87] Diana is "a powerful Goddess," not merely a supremely intelligent being, who can manipulate many other background conditions in the universe besides the constitution of the zygote: "Diana is stipulated to be capable of controlling for a maximally wide range of possible changes to the background conditions."[88] Furthermore, she can design other agents besides Ernie to ensure that someone brings about the result she desires.[89] Unsurprisingly, Diana is then the most robust cause of the wrongdoing performed thirty years later. On their theory, this makes Diana, but not the agent grown from the zygote, responsible for the later wrongdoing. However, Diana is also responsible and fully blameworthy in Mele's original case, despite her lack of causal sourcehood with respect to the wrongdoing that occurs thirty years later.

Usher as well as Deery and Nahmias are right that there is a connection between robust causation and the control required for responsibility, but the connection is contingent and defeasible. As Lewis puts it: "*Ceteris paribus*, shortness and simplicity of the chain will make for insensitivity; insensitivity, in turn, will make for foreseeability."[90] Given that Stringer has the right kind of foresight, Stringer can be responsible and blameworthy for Omar's killing of Mouzone in *Testimony*, whether or not the causal dependence between his bodily movements and the death of Mouzone is more robust in *Testimony* than in *Lone Killer*. What is important for whether the agent is responsible and fully blameworthy for a killing is whether he can intend and foresee that a causal pathway from his own action will eventually result in the victim's death. This is in general what is essential for (nonbasic) responsibility for action, not insensitivity itself.[91]

87 Deery and Nahmias, "Defeating Manipulation Arguments," 1257. My interpretation of the zygote case agrees with Usher's ("Agency, Teleological Control and Robust Causation," 320). In contrast, Tierney and Glick's interpretation agrees with Deery and Nahmias's (Tierney and Glick, "Desperately Seeking Sourcehood," 958n6). Some of Mele's later retellings of the zygote case seem more in line with Deery and Nahmias's interpretation (e.g., *Free Will*, 15–16).

88 Deery and Nahmias, "Defeating Manipulation Arguments," 1272n15. Note that Mele does have another thought experiment involving a "libertarian goddess in an indeterministic universe" who is also called Diana (*Free Will and Luck*, 7).

89 Deery and Nahmias, "Defeating Manipulation Arguments," 1264.

90 David Lewis, "Causation," 187. Later in the same paragraph, Lewis writes: "If a chain is insensitive enough that you can predict it, then it is insensitive enough that you can kill by it. . . . What if you are much better than I am at predicting chains that are somewhat sensitive? I am inclined to say that if so, then indeed you can kill in ways that I cannot."

91 See Zimmerman, "Intervening Agents and Moral Responsibility," 356. Grinfeld et al. argue that people judge an agent to be more causally responsible for an event in cases where

6. FOURTH OBJECTION: THE AUTONOMY DOCTRINE

The last objection I will consider is based on a *normative policy* rather than some feature of agency or responsibility as such. When it comes to criminal responsibility for criminalized acts, a normative policy that sharply distinguishes between cases such as *Testimony* and *Lone Killer* is indeed widely accepted. But if my thesis is true and criminal responsibility ought to track moral responsibility, then the normative policy often referred to as the "autonomy doctrine" would be in jeopardy.[92]

Here is Glanville Williams's characterization of this policy:

> The first actor who starts on a dangerous or criminal plan will often be responsible for what happens if no one else intervenes; but a subsequent actor who has reached responsible years, is of sound mind, has full knowledge of what he is doing, and is not acting under intimidation or other pressure or stress resulting from the defendant's conduct, replaces him as the responsible actor. Such an intervening act is thought to break the moral connection that would otherwise have been perceived between the defendant's acts and the forbidden consequence.[93]

But *why* would it break "the moral connection"? Why would the second agent's intervening action make the first agent's action morally permissible, or make it morally wrong in a different and lesser way? Note that the moral connection need not be broken if the first agent uses threats, lies, or authority to induce the second agent to commit a crime. Consider first the following case:

> *Authority*: Stringer is a powerful and ruthless acting leader of a criminal organization. He desires and intends Mouzone to be killed. To that end, he commands Roland, a lower-ranking drug enforcer, to kill Mouzone. Roland does as he was ordered. He tracks down Mouzone, aims a handgun at him, and pulls the trigger. The bullet hits Mouzone, who dies immediately.

---

the agent's action more robustly causes the event ("Causal Responsibility and Robust Causation"); but their experiments do not disentangle the robustness of the causation and the agent's ability to foresee what will result from her action. People's judgments may therefore be sensitive to the latter rather than to the former.

92 For critical discussion of the autonomy doctrine, see Moore, "Causing, Aiding, and the Superfluity of Accomplice Liability"; Bazargan-Forward, "Complicity"; and du Bois-Pedain, "*Novus Actus* and Beyond."

93 Williams, "*Finis* for *Novus Actus*?," 392. See also Kadish, "Complicity, Cause and Blame," 327; and Hart and Honoré, *Causation in the Law*.

In *Authority*, I take it to be relatively unproblematic that Stringer would be morally responsible for the killing of Mouzone. Saba Bazargan-Forward and David Atenasio would each argue that what makes Stringer responsible and blameworthy for Roland's action in *Authority* is an agreement that authorizes Roland to act on Stringer's behalf.[94] Such an agreement would be implicit in the issuing and uptake of Stringer's command to Roland. On their views, it is this authorization agreement itself, rather than the foresight and indirect control that it engenders, that is normatively significant. On my view, on the other hand, Stringer's authority over Roland is relevant for what he is responsible for precisely because it enables him to foresee that his order will cause Mouzone's death. In *Testimony*, Stringer's knowledge of what sort of person Omar is likewise enables him to foresee that telling Omar the truth will result in Mouzone's death. Stringer is therefore morally responsible and fully blameworthy for the killing in *Testimony*, just as in *Authority*.

Turn now to the following case, where Stringer lies to Omar in order to make him kill Mouzone:

> *Deception*: Stringer desires and intends Mouzone to be killed. He knows that the beloved of the notorious stickup man Omar has died as a result of accidentally falling from a high balcony. But knowing what sort of person Omar is, Stringer knows that if he deceives Omar into thinking that his beloved was actually murdered by Mouzone, then it is very probable (with probability 0.8) that Mouzone will die as a result of Omar deciding to kill him and then carrying out this decision. With intent to bring about Mouzone's death, Stringer provides fabricated evidence to Omar that convincingly frames Mouzone as the murderer of Omar's beloved. Upon receiving the fabricated evidence, Omar acquires a desire to avenge his beloved's death, but this desire is not irresistible. He decides to kill Mouzone just as Stringer predicted. Omar then tracks down Mouzone, aims a handgun at him, and pulls the trigger. The bullet hits Mouzone, who dies immediately.

In the United States, Stringer could be convicted for instigating murder in *Deception*. If *Deception* (as well as *Authority*) was set in Berlin rather than Baltimore, then Stringer could also be convicted as a perpetrator of the murder in accordance with the doctrine of "the perpetrator behind the perpetrator" (*Der Täter hinter dem Täter*), in such a way that Stringer and Omar (or Roland)

94 Bazargan-Forward, "Complicity"; Atenasio, "Co-responsibility for Individualists."

could each be convicted as a perpetrator of one and the same murder.[95] Somewhat similarly, Swedish criminal law allows for the relabeling of the roles of those involved in a crime, such that an agent who "merely" instigates rather than performs the criminal act can nevertheless end up being convicted as a perpetrator.[96] In contrast, Stringer would be completely off the legal hook in *Testimony*, irrespective of whether the case was set in Baltimore, Berlin, or Borås.

Whatever the local legal doctrine is, I take it that many will judge Stringer to be *morally* responsible and blameworthy for Omar's killing of Mouzone in *Deception*—or at least, for the outcome that Omar killed Mouzone. Now, there is arguably no difference between *Deception* and *Testimony* such that Stringer could be morally responsible and blameworthy for the killing in the former case but not in the latter. To make the cases as nearly parallel as possible, suppose that Stringer in *Testimony* is the only person besides Mouzone himself who is privy to the information that Mouzone murdered Omar's beloved. Stringer's ability to foresee the result of the lie he tells Omar in *Deception* might for all practical purposes be identical to his ability to foresee the result of his truthful and sincere testimony to Omar in *Testimony*.

In some sense, Stringer constrains Omar's autonomy when he lies to Omar, but not when he tells Omar the truth. Perhaps this diminishes Omar's blameworthiness for killing Mouzone in *Deception*. It is tempting to think that this diminishment "makes room" for Stringer to be fully blameworthy for the killing in a way that is ruled out in *Testimony* by Omar's full blameworthiness for the killing. However, this thought requires a mistaken "pie model" of blameworthiness, where blameworthiness for a wrongdoing comes in a fixed amount that has to be distributed among those responsible for the wrongdoing. This pie model has been frequently and convincingly criticized.[97] The degree to which

95 See Ambos and Bock, "Germany," 327–30. Regarding this doctrine, Ambos and Bock write:

> Imagine, for example, that D knows that A wants to kill V_1 and falsely points out V_2 who is at the moment passing by, and tells *A* that this is V_1, although *D* is fully aware that this is not the case. As expected, *A* shoots and kills V_2 assuming that he is V_1. *A* has committed the crime of murder as a principal. His mistake concerning the identity of his victim (*error in persona*) does not affect his intent to kill the person in front of him and is thus irrelevant. Despite the fact that *A* is fully criminally liable, it is *D* who has "transferred" *A*'s intent from V_1 to V_2. Thus, *D* has killed V_2 through the "blind" *A*. (328)

If *D* can perpetrate the murder of V_2 by *transferring A*'s murderous intent in this way, then in *Deception* Stringer could perpetrate the murder of Mouzone by *creating* Omar's murderous intent.

96 See Asp and Ulväng, "Sweden," 442–45.

97 See Mellema, "Shared Responsibility and Ethical Dilutionism"; Zimmerman, "Intervening Agents and Moral Responsibility," 355; Sverdlik, "Collective Responsibility," 71–72;

Omar is blameworthy for killing Mouzone does not itself make any difference to Stringer's blameworthiness for the killing.

Perhaps something like the autonomy doctrine is, generally speaking, a good legal policy. When the law gives its verdict on a case such as *Testimony*, it does so from a third-person point of view and after the fact. Given limitations in epistemic access to what was actually going on, and to what kind of foresight Stringer was capable of, the law will typically be justified in assuming that there is a significant difference between Stringer's criminal responsibility for the murder of Mouzone in *Testimony* and in *Lone Killer*. Since the use of authority, deception, and coercion is typically evidence of ill will, foreseeability, and control, *Authority* and *Deception* may be more similar to *Lone Killer* than to *Testimony* with respect to Stringer's *legal* responsibility. However, my concern here is with moral responsibility and blameworthiness, which do not depend on the evidence available to third parties about the agent's quality of will, knowledge, and foresight.

7. CONCLUSION

Philosophers working on agency and responsibility sometimes take it for granted that one cannot be responsible for another agent's intentional action, at least when the other performs that action freely—without being coerced or otherwise manipulated, and without acting on behalf of the first agent (in the sense of acting under the first agent's authority). In this paper, I have argued that an agent can be responsible and fully blameworthy for another agent's intentional action when the second agent acts freely and in the absence of any authorization agreement or special kind of joint agency. Stringer can be responsible and fully blameworthy for Omar's killing of Mouzone, just by intentionally creating the conditions that causes Omar to freely and intentionally kill Mouzone. I have also argued that when we hold each other responsible for what we do in a colloquial sense, we typically hold each other responsible for events that we created the conditions for, rather than for our basic actions—that is, the actions by which we create those conditions. An agent can thus be responsible and blameworthy for another agent's intentional action in the relevantly same way that he is responsible and blameworthy for his own intentional action. If we hold Stringer responsible for killing Mouzone in *Lone Killer*, then we are holding him responsible and fully blameworthy for something that he "merely" created the conditions for, by flexing his index finger in a certain

Ludwig, "From Individual Responsibility to Collective Responsibility"; and Kaiserman, "Responsibility and the 'Pie Fallacy.'"

context. Similarly, we can hold Stringer responsible and fully blame him for Omar's killing of Mouzone in *Testimony*, where he created the conditions for this killing by revealing the truth to Omar.[98]

University of Gothenburg
olle.blomberg@gu.se

REFERENCES

Aguilar, Jesús H. "Interpersonal Interactions and the Bounds of Agency." *Dialectica* 61, no. 2 (June 2007): 219–34.

Alvarez, Maria. "Actions, Thought-Experiments and the 'Principle of Alternate Possibilities.'" *Australasian Journal of Philosophy* 87, no. 1 (March 2009): 61–81.

Ambos, Kai, and Stefanie Bock. "Germany." In *Participation in Crime: Domestic and Comparative Perspectives*, edited by Alan Reed and Michael Bohlander, 323–39. Abingdon: Routledge, 2013.

Asp, Petter, and Magnus Ulväng. "Sweden." In *Participation in Crime: Domestic and Comparative Perspectives*, edited by Alan Reed and Michael Bohlander, 433–50. Abingdon: Routledge, 2013.

Atenasio, David. "Co-responsibility for Individualists." *Res Publica* 25, no. 4 (November 2019): 511–30.

Bazargan-Forward, Saba. "Complicity." In *The Routledge Handbook of Collective Intentionality*, edited by Marija Jankovic and Kirk Ludwig, 327–37. Abingdon: Routledge, 2018.

Blomberg, Olle. "Socially Extended Intentions-in-Action." *Review of Philosophy and Psychology* 2, no. 2 (June 2011): 335–53.

98 I am grateful to this journal's as well as to *Philosophical Quarterly*'s anonymous reviewers, all of whom significantly helped me make this paper better. For helpful comments and questions, thanks also to Gunnar Björnsson, Stephanie Collins, Hadi Fazeli, Thor Grünbaum, Yuliya Kanygina, Olof Leffler, Yair Levy, Al Mele, Per-Erik Milam, Julius Schönherr, David Shoemaker, Daniel Story, Erik Svensson, and Matthew Talbert. In addition, I am grateful for audience feedback at Linköping University (January 2021), the University of Gothenburg (March 2021), the Moral Responsibility over Time and between Persons workshop (online, March 2021), the Empirical Approaches to Rationality and Action discussion group (online, September 2021), the Third CNY Moral Psychology Workshop (online, December 2021), Peking University (May 2022), the 2022 Swedish Congress of Philosophy (Lund, June 2022), and the Workshop on Agency and Responsibility (Copenhagen, June 2022). Finally, thanks to Sara Ash Georgi for her truly brilliant copyediting of this paper. My research was funded by the Lund Gothenburg Responsibility Project (PI: Paul Russell), which is in turn funded by the Swedish Research Council.

Blomberg, Olle, and Frank Hindriks. "Collective Responsibility and Acting Together." In *The Routledge Handbook of Collective Responsibility*, edited by Saba Bazargan-Forward and Deborah Tollefsen, 142–54. Abingdon: Routledge, 2020.

Blomberg, Olle, and Björn Petersson. "Team Reasoning and Collective Moral Obligation." *Social Theory and Practice* (forthcoming). Published ahead of print, January 27, 2023. https://doi.org/10.5840/soctheorpract2023120177.

Bratman, Michael E. "I Intend That We J." In *Contemporary Action Theory*, vol. 2, *Social Action*, edited by Ghita Holmström-Hintikka and Raimo Tuomela, 49–63. Dordrecht: Kluwer, 1997.

Capes, Justin A. "Freedom with Causation." *Erkenntnis* 82, no. 2 (April 2017): 327–38.

Clarke, Randolph. *Omissions: Agency, Metaphysics, and Responsibility*. New York: Oxford University Press, 2014.

———. "Responsibility for Acts and Omissions." In Nelkin and Pereboom, *The Oxford Handbook of Moral Responsibility*, 91–110.

Davidson, David. "Agency." In *Agent, Action, and Reason*, edited by Robert Binkley, Richard Bronaugh, and Ausonio Marras, 3–25. Toronto: University of Toronto Press, 1971.

Deery, Oisín, and Eddy Nahmias. "Defeating Manipulation Arguments: Interventionist Causation and Compatibilist Sourcehood." *Philosophical Studies* 174, no. 5 (May 2017): 1255–76.

Dennett, Daniel C. *Elbow Room: The Varieties of Free Will Worth Wanting*. Cambridge, MA: MIT Press, 1984.

Dretske, Fred. "The Metaphysics of Freedom." *Canadian Journal of Philosophy* 22, no. 1 (March 1992): 1–14.

du Bois-Pedain, Antje. "*Novus Actus* and Beyond: Attributing Causal Responsibility in the Criminal Courts." *Cambridge Law Journal* 80, supplement 1 (September 2021): S61–S90.

Enoch, David, and Andrei Marmor. "The Case against Moral Luck." *Law and Philosophy* 26, no. 4 (July 2007): 405–36.

Feinberg, Joel. "Causing Voluntary Actions." In *Metaphysics and Explanation: Proceedings of the 1964 Oberlin Colloquium in Philosophy*, edited by William H. Capitan and Daniel D. Merrill, 29–47. Pittsburgh: University of Pittsburgh Press, 1966.

———. "Collective Responsibility." *Journal of Philosophy* 65, no. 21 (November 1968): 674–88.

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

Fischer, John Martin, and Neal A. Tognazzini. "The Physiognomy of

Responsibility." *Philosophy and Phenomenological Research* 82, no. 2 (March 2011): 381–417.
Ford, Anton. "The Province of Human Agency." *Noûs* 52, no. 3 (September 2018): 697–720.
Frankfurt, Harry G. "Three Concepts of Free Action." *Aristotelian Society Supplementary Volume* 49, no. 1 (July 1975): 113–25.
———. "What We Are Morally Responsible For." In *Perspectives on Moral Responsibility*, edited by John Martin Fischer and Mark Ravizza, 286–95. Ithaca, NY: Cornell University Press, 1993.
Gardner, John. "Complicity and Causality." *Criminal Law and Philosophy* 1, no. 2 (May 2007): 127–41.
Ginet, Carl. "An Action Can Be Both Uncaused and Up to the Agent." In *Intentionality, Deliberation, and Autonomy: The Action-Theoretic Basis of Practical Philosophy*, edited by Christoph Lumer and Sandro Nannini, 243–55. Aldershot, UK: Ashgate, 2007.
Goldman, Alvin. *A Theory of Human Action*. Englewood Cliffs, NJ: Prentice-Hall, 1970.
Grinfeld, Guy, David Lagnado, Tobias Gerstenberg, James F. Woodward, and Marius Usher. "Causal Responsibility and Robust Causation." *Frontiers in Psychology* 11 (May 2020): 1069. https://www.frontiersin.org/articles/10.3389/fpsyg.2020.01069/.
Hart, H. L. A., and Anthony M. Honoré. *Causation in the Law*. 2nd ed. Oxford: Oxford University Press, 1985.
Hartman, Robert J. *In Defense of Moral Luck: Why Luck Often Affects Praiseworthiness and Blameworthiness*. New York: Routledge, 2017.
Hill, Daniel J., Stephen K. McLeod, and Attila Tanyi. "The Concept of Entrapment." *Criminal Law and Philosophy* 12, no. 4 (December 2018): 539–54.
Himmelreich, Johannes. "Responsibility for Killer Robots." *Ethical Theory and Moral Practice* 22, no. 3 (June 2019): 731–47.
Hornsby, Jennifer. *Actions*. Abingdon: Routledge, 1980.
Jeppsson, Sofia. "Accountability, Answerability, and Attributability: On Different Kinds of Moral Responsibility." In Nelkin and Pereboom, *The Oxford Handbook of Moral Responsibility*, 73–88.
Kadish, Sanford H. "Complicity, Cause and Blame: A Study in the Interpretation of Doctrine." *California Law Review* 73, no. 2 (March 1985): 323–410.
Kaiserman, Alex. "Responsibility and the 'Pie Fallacy.'" *Philosophical Studies* 178, no. 11 (November 2021): 3597–616.
Kane, Robert. *The Significance of Free Will*. Oxford: Oxford University Press, 1996.
Khoury, Andrew C. "The Objects of Moral Responsibility." *Philosophical Studies*

175, no. 6 (June 2018): 1357–81.
Köhler, Sebastian. "Moral Responsibility without Personal Identity?" *Erkenntnis* 86, no. 1 (February 2021): 39–58.
Lang, Gerald. *Strokes of Luck: A Study in Moral and Political Philosophy*. Oxford: Oxford University Press, 2021.
Levy, Neil. *Consciousness and Moral Responsibility*. Oxford: Oxford University Press, 2014.
Lewis, David. "Causation." In *Philosophical Papers*, vol. 2, 159–213. Oxford: Oxford University Press, 1987.
Lewis, Hywel D. "Collective Responsibility." *Philosophy* 23, no. 84 (January 1948): 3–18.
Ludwig, Kirk. *From Individual to Plural Agency*. Vol. 1 of *Collective Action*. Oxford: Oxford University Press, 2016.
———. "From Individual Responsibility to Collective Responsibility: There and Back Again." In *The Routledge Handbook of Collective Responsibility*, edited by Saba Bazargan-Forward and Deborah Tollefsen, 78–93. Abingdon: Routledge, 2020.
Markovits, Julia. "Acting for the Right Reasons." *Philosophical Review* 119, no. 2 (April 2010): 201–42.
McCann, Hugh J. "Dretske on the Metaphysics of Freedom." *Canadian Journal of Philosophy* 23, no. 4 (December 1993): 619–30.
McKenna, Michael. *Conversation and Responsibility*. Oxford: Oxford University Press, 2012.
Mele, Alfred R. "Direct versus Indirect: Control, Moral Responsibility, and Free Action." *Philosophy and Phenomenological Research* 102, no. 3 (May 2021): 559–73.
———. *Free Will and Luck*. Oxford: Oxford University Press, 2006.
———. *Free Will: An Opinionated Guide*. Oxford: Oxford University Press, 2022.
———. *Manipulated Agents: A Window to Moral Responsibility*. Oxford: Oxford University Press, 2019.
Mellema, Gregory. "Shared Responsibility and Ethical Dilutionism." *Australasian Journal of Philosophy* 63, no. 2 (June 1985): 177–87.
Moore, Michael S. "Causing, Aiding, and the Superfluity of Accomplice Liability." *University of Pennsylvania Law Review* 156, no. 2 (December 2007): 395–452.
Nelkin, Dana K. "Moral Luck." *Stanford Encyclopedia of Philosophy* (Summer 2021). https://plato.stanford.edu/archives/sum2021/entries/moral-luck/.
Nelkin, Dana K., and Derk Pereboom, eds. *The Oxford Handbook of Moral Responsibility*. Oxford: Oxford University Press, 2022.

Núñez, Carlos. "Intending Recalcitrant Social Ends." *Erkenntnis* 87, no. 2 (April 2022): 477–98.

Robichaud, Philip, and Jan Willem Wieland. "A Puzzle concerning Blame Transfer." *Philosophy and Phenomenological Research* 99, no. 1 (July 2019): 3–26.

Roth, Abraham Sesshu. "Entitlement to Reasons for Action." In *Oxford Studies in Agency and Responsibility*, vol. 4, edited by David Shoemaker, 75–92. Oxford: Oxford University Press, 2017.

Russell, Paul. "Responsibility and the Condition of Moral Sense." *Philosophical Topics* 32, nos. 1 and 2 (Spring and Fall 2004): 287–305.

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

———. "Responsibility and Causation." In Nelkin and Pereboom, *The Oxford Handbook of Moral Responsibility*, 348–62.

Schechtman, Marya. *The Constitution of Selves*. Ithaca, NY: Cornell University Press, 1996.

Shoemaker, David. "Responsibility without Identity." *Harvard Review of Philosophy* 18, no. 1 (Spring 2012): 109–32.

Smith, Holly M. "Negligence." *International Encyclopedia of Ethics*, edited by Hugh LaFollette. Hoboken, NJ: Wiley-Blackwell, 2013. https://doi.org/10.1002/9781444367072.wbiee199.

Story, Daniel. "Essays concerning the Social Dimensions of Human Agency." PhD diss., UC Santa Barbara, 2020. https://escholarship.org/uc/item/1xh8p10s.

Strawson, Peter F. "Freedom and Resentment." *Proceedings of the British Academy 1962* 48 (1963): 187–211.

Sverdlik, Steven. "Collective Responsibility." *Philosophical Studies* 51, no. 1 (January 1987): 61–76.

———. "Crime and Moral Luck." *American Philosophical Quarterly* 25, no. 1 (January 1988): 79–86.

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

Tierney, Hannah, and David Glick. "Desperately Seeking Sourcehood." *Philosophical Studies* 177, no. 4 (April 2020): 953–70.

Usher, Marius. "Agency, Teleological Control and Robust Causation." *Philosophy and Phenomenological Research* 100, no. 2 (March 2020): 302–24.

Vargas, Manuel. *Building Better Beings: A Theory of Moral Responsibility*. Oxford: Oxford University Press, 2013.

Weil, Vivian M., and Irving Thalberg, "The Elements of Basic Action." *Philosophia* 4, no. 1 (January 1974): 111–38.

Wiland, Eric. "(En)joining Others." In *Oxford Studies in Agency and Responsibility*, vol. 6, edited by David Shoemaker, 64–84. Oxford: Oxford University Press, 2019.

Williams, Glanville. "*Finis* for *Novus Actus*?" *Cambridge Law Journal* 48, no. 3 (November 1989): 391–416.

Wolff, Phillip. "Direct Causation in the Linguistic Coding and Individuation of Causal Events." *Cognition* 88, no. 1 (May 2003): 1–48.

Woodward, James. "Sensitive and Insensitive Causation." *Philosophical Review* 115, no. 1 (January 2006): 1–50.

Zimmerman, Michael J. "Intervening Agents and Moral Responsibility." *Philosophical Quarterly* 35, no. 141 (October 1985): 347–58.

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

Journal of Ethics and Social Philosophy
VOL. 25, NO. 3 · SEPTEMBER 2023
https://doi.org/10.26556/jesp.v25i3.1968

UNCERTAINTY AND INTENTION

Benjamin Lennertz

IT IS COMMON to express intentions using future-tensed indicative sentences that seem grammatically and semantically fit to express beliefs.[1] Let us consider the following example, which we can call *Sunday Eggs*. On Sunday morning, your roommate reports that you are out of eggs. Your day is wide open, so you say:

(1) I will go to the store today.

By uttering (1) in the scenario Sunday Eggs, you have expressed an intention to go to the store on Sunday, and it seems that you have also expressed a belief that you will do so. The expression of the intention and the expression of the belief in uttering one sentence makes the attitudes seem at least intimately related, if not identical; this is one of multiple reasons in favor of a popular thesis about the relationship between intention and belief:

Intention Implies Belief: If *S* intends to ϕ, then *S* believes *S* will ϕ.[2]

Furthermore, it seems like in uttering (1) in Sunday Eggs, you are expressing an intention that commits you to going to the store today. Here is a general statement inspired by this scenario:

1 Anscombe, *Intention*; Velleman, "What Good Is a Will?"; Marušić and Schwenkler, "Intending Is Believing."

2 For the suggestion that the fact that we simultaneously express intentions and beliefs is a reason in favor of Intention Implies Belief, see Marušić and Schwenkler, "Intending Is Believing," 319. So-called cognitivists about practical rationality or intention accept and often argue in favor of Intention Implies Belief. See Harman, "Practical Reasoning"; Setiya, "Cognitivism about Instrumental Reason"; Velleman, "What Good Is a Will?"; Marušić and Schwenkler, "Intending Is Believing." Grice proposes the view about acceptance rather than belief ("Intention and Uncertainty"). There are also many arguments against this view, including Davidson, "Intending"; Bratman, *Intention, Plans, and Practical Reason* and "Intention, Belief, Practical, Theoretical"; Holton, "Partial Belief, Partial Intention"; Kolodny, "The Myth of Practical Consistency," 372–73; Brunero, "Against Cognitivism about Practical Rationality."

Intention Implies Commitment: If *S* intends to ϕ, then *S* is committed to ϕ-ing.[3]

So a first pass way of thinking of ascriptions like (1) can make both Intention Implies Belief and Intention Implies Commitment seem attractive. In this paper, I will explore an intention ascription similar to (1) that challenges, rather than supports, the combination of Intention Implies Belief and Intention Implies Commitment.

Consider:

Friday Eggs: You bake a lot of cookies that week, and on Friday morning, your roommate notes that you are again out of eggs. Friday is busier than Sunday, so you say:

(2) I will probably go to the store today.

In uttering (2) in Friday Eggs, it seems, again, that you have expressed a sort of intention. But there is a difference between (1) and (2), signaled by the addition of "probably."[4]

3 Bratman, *Intention, Plans, and Practical Reason*, 107–10. I will discuss this further in section 4.2. Some authors mentioned in note 2 above construe these commitments as identical to the commitments of beliefs. See Velleman, "What Good Is a Will?"; Marušić and Schwenkler, "Intending Is Believing." However, it seems possible to distinguish these conceptually, as when Hieronymi says "an intention is a commitment to doing something, where a belief is a commitment to a claim as true" ("Controlling Attitudes," 56). A nearby point is that intending to ϕ *settles* what one will do. See Mele, "Intention, Belief, and Intentional Action," 26; Hieronymi, "Controlling Attitudes," 56.

4 A reviewer notes that neither of the following, which seem related to (2), are quite felicitous in Friday Eggs:

(2a) I intend to probably go to the store today.

(2b) I probably intend to go the store today.

There are two main hypotheses about the semantics and pragmatics of "probably." The traditional position claims that "Probably *p*" is context sensitive and can be paraphrased as: the relevant probability function in the context assigns a high value to *p*. See Dowell, "A Flexible Contextualist Account of Epistemic Modals"; Lennertz, *Reasoning with Uncertainty and Epistemic Modals*. But it would be strange in an ordinary case like Friday Eggs to have an intention about the value that the relevant probability function assigns to going to the store; this is one possible explanation of the infelicity of (2a). By contrast, the nontraditional, popular view says that sentences of the form "Probably *p*" do not encode propositions at all; they are used to express the speaker's high confidence or credence in *p*. Here are two quick arguments in support of this claim, though there are others, e.g., the animal/baby thought argument in Yalcin, "Epistemic Modals," 997, and "Nonfactualism about Epistemic Modality," 308; Price, "Conditional Credence," 19; Frankish, "Partial Belief and Flat-out Belief." First, no concept of probability makes sense of what is said using ordinary utterances, like (3):

(3) The Sparks probably won last night.

In this paper, I argue that a speaker who utters a sentence like (2) in a scenario like Friday Eggs expresses attitudes that are inconsistent with either Intention Implies Belief or Intention Implies Commitment. I do so by exploring what a speaker who utters (2) expresses in a range of situations. I find that there are two plausible accounts of what you express in an ordinary situation like Friday Eggs: an ordinary intention without a belief—so that Intention Implies Belief is false—or a partial intention that does not commit you to going to the store—so that Intention Implies Commitment is false.[5]

When a speaker utters (3), they are not talking about notions of probability like the relative frequency of possible Sparks wins last night to possible Sparks games last night; nor are they talking about the objective quantum chance of a Sparks win last night. The same consideration, purveyors of this argument claim, applies to any other notion of probability. See Maher, "The Irrelevance of Belief to Rational Action," 367; Christensen, *Putting Logic in Its Place*, 18–20; Ross, *Acceptance and Practical Reason*, 189; Eriksson and Hájek, "What Are Degrees of Belief?," 206–7; Staffel, "Can There Be Reasoning with Degrees of Belief?," 5357; Konek, "Probabilistic Knowledge and Cognitive Ability," 514; Moss, *Probabilistic Knowledge*, 2; though see Moon and Jackson, "Credence," for a reply. Second, "Probably *p*" does not embed well under attitude verbs that take propositions—other than acceptance ones like "believe" and "know":

(4) Jane fears probably being confined in small spaces.

(5) Sally hopes that her son probably gets a good grade.

It is hard to understand what is meant by (4) and (5) without removing "probably" altogether in interpretation, suggesting that there is no proposition expressed by "Probably *p*". If propositions are the objects of intentions, as well as hopes and fears, that would mean that (2a), like (4) and (5), is not well formed. What about (2b)? It is used to either express high confidence that you have an intention to go to the store (the popular view) or convey that the relevant probability function in the context assigns a high value to the proposition that you intend to go to the store (the traditional view). Either is a strange thing to express or convey, given that we often assume a thinker's intentions are transparent to them. The reviewer suggested that to talk naturally in this way, we must dispel this assumption of transparency, as in "Since starting psychoanalysis, I have come to think that I probably intend to go to the store." Without the suggestion of the failure of transparency—implied by the psychoanalysis clause—it makes sense that (2b) would be infelicitous. It is also worth noting that the strangeness of (2a) and (2b) in Friday Eggs contrasts with the naturalness of (2), suggesting that we should not try to analyze (2) as either (2a) or (2b).

5 The first diagnosis interprets our case as generally analogous in structure to proposed counterexamples to *Intention Implies Belief*. Bratman uses an example where a person is simultaneously playing two video games (*Intention, Plans, and Practical Reason*, 113–15); Mele gives an example involving an uncertain free-throw shooter ("Intention, Belief, and Intentional Action," 19–20). The second diagnosis interprets the case as related to those proposed to motivate the existence of partial intentions. See Chan, "A Not-so-Simple View of Intentional Action"; Holton, "Partial Belief, Partial Intention"; Shpall, "The Calendar Paradox"; Goldstein, "A Preface Paradox for Intention"; Beddor, "Fallibility for Expressivists"; Jian, "Rational Norms for Degreed Intention (and the Discrepancy between Theoretical and Practical Reason)." Our investigation is interesting in a way that goes beyond those

1. PRELIMINARIES: COMMUNICATING WITH SENTENCE (1)

It is easiest to start our explanation of what you express in uttering (2) in Friday Eggs by fleshing out our picture of what goes on when you utter (1) in Sunday Eggs:

(1) I will go to the store today.

As we said, it appears that you do at least two things: you express a belief that you will go to the store today, and you express an intention to go to the store today. The default linguistic explanation of how this would happen is that one of these speech acts is connected in a close way to the meaning of the sentence while the other is less explicit. We can call the first act the direct speech act. In Sunday Eggs, you express the belief that you will go to the store today as the direct speech act. Additionally, you use the direct speech act as a way of performing an indirect speech act—one that is less closely connected with the meaning of the sentence. In Sunday Eggs, you express the intention to go to the store today as an indirect speech act. Direct and indirect speech acts are analogous to Grice's notions of saying (or making as if to say) and implicating, respectively.[6]

The mechanism by which your roommate can infer your intention from your utterance of (1) is broadly Gricean, based on general reasoning about your state of mind. They could reason that in uttering (1), you expressed the belief that you will go to the store today. And they could wonder why you believe that. Since going to the store today is something you would want to do in order to get eggs, and since it seems like doing so is under your control, they could

---

earlier treatments for two reasons. First, it can be easier to have judgments about cases involving conversations, and this can highlight the problems for the accepted views in more natural examples than some discussed in the literature—like Bratman's video game case in *Intention, Plans, and Practical Reason*, 113–15, or the preface paradox for intention in Shpall, "The Calendar Paradox," and Goldstein, "A Preface Paradox for Intention." Second, it is independently interesting how intention-like attitudes are ascribed using sentences like (2), and understanding this helps us better appreciate or, as we will see in section 4.3, critique arguments in the literature that rely on "I will ϕ" constructions as canonical for expressing intention. Thanks to Jay Jian for discussion.

6 Grice, "Logic and Conversation." Expressing intentions calls for thinking of speech acts (or expressing mental states) in general, rather than just assertion or saying (or expressing beliefs). Grice realizes this general point: "I have stated my maxims as if this purpose [of communication] were a maximally effective exchange of information; this specification is, of course, too narrow, and the scheme needs to be generalized to allow for such general purposes as influencing or directing the actions of others" ("Logic and Conversation," 28).

conclude that you are not merely predicting that you will go to the store but that you intend to do so.[7]

There are a number of reasons to prefer this picture to one that says that expressing an intention to go to the store in uttering (1) is the direct speech act—closely tied to the meaning of (1). First, as has been widely noted, many sentences with (1)'s form are not used to express an intention at all.[8] For instance:

(6) I will be sick.

It would be strange if the meaning of (1) constrains its direct speech act to be an expression of an intention, while the meaning of (6)—which has the same form—does not. More importantly, whether an intention is expressed by a sentence of this form depends on the context in which the sentence is uttered.[9] In the nonstandard but possible context where one swallows something one should not, one can utter (6) to express an intention to be sick. And in nonstandard contexts, one can utter (1) without expressing an intention:

> *Sleepwalker*: You know you are a predictable sleepwalker. When you take an afternoon nap, you always sleepwalk to the store. You are going to nap on Sunday, and your roommate asks where you will sleepwalk to. You reply with (1).

In uttering (1) in Sleepwalker, you express a belief that you will go to the store but do not express an intention to go to the store.

Finally, it appears that we can cancel the expression of an intention, even in a context in which it appears natural to assume it has been performed:

(7) I will go to the store today. Sandy will drag me there as she always does. I will try as hard as I can to get out of it, but I am sure I will fail.[10]

7 This phenomenon may be like *generalized* conversational implicature, where "the use of a certain form of words in an utterance would normally (in the absence of special circumstances) carry such-and-such an implicature" (Grice, "Logic and Conversation," 37). Those who think that forms of sentences like (1) are particularly well suited to defeasibly express intentions might also use the notion of *standardization* as a model, from Bach and Harnish, *Linguistic Communication and Speech Acts*.

8 Anscombe, *Intention*.

9 Asarnow uses a similar example to make this point ("Noncognitivism in Metaethics and the Philosophy of Action," 6–7).

10 For the claim that we can cancel the expression of an intention like an implicature, see Asarnow, "Noncognitivism in Metaethics and the Philosophy of Action," 17–18.

By way of comparison, the reader can note that a speaker who uses (1) will express their belief in any context in which they are sincere and that this expression cannot be cancelled by adding more information (as in [7]) without retracting one's utterance of (1) itself. This is strong evidence to think that, in uttering a sentence like (1) in a scenario like Sunday Eggs, you directly express your belief and indirectly express your intention.[11]

2. Possible Indirect Speech Acts Performed in Uttering (2)

Now, consider again sentence (2):

(2) I will probably go to the store today.

In what respects are the speech acts a speaker performs in using (2) like those performed in using (1)? It seems natural to think that in uttering (2) you directly express high confidence (credence) that you will go to the store today.[12] This is analogous to the belief you express in uttering (1). The key question is what,

11 As a reviewer notes, the arguments in this paragraph assume two features of what is directly expressed by a speaker. First, it is not cancellable; we cannot take back what is directly expressed without revoking our commitment to what is conveyed in using those words. Second, it is context insensitive in the following weak sense: the same sentence cannot be used in different contexts to directly express different sorts of mental states. (This leaves open that language is context sensitive in the more standard way of having the content of the state of mind directly expressed depend on the context, as the belief that is directly expressed by a use of "I am hungry now" depends on who is speaking and when.) These two assumptions are standard in Gricean and similar paradigms, though they might be denied by more radical views. For instance, the essays of Travis, *Occasion-Sensitivity*, present a paradigm in which it is natural to deny the latter assumption. However, not only are these assumptions standard, they strike me as extremely natural. It is hard to even state the first assumption, since a direct speech act is just the one that the speaker is fundamentally committed to—and, so, it could not be cancelled without revoking commitment to what was directly done in uttering the sentence. Denying the second assumption is comprehensible but, I think, implausible. I do not have space to fully discuss the issue here, but I suspect that a picture of the connection between language and its use that allowed such a radical sort of context sensitivity lacks the systematicity that would be required for language being as vastly helpful as it is for communication.

12 Yalcin suggests that "in asserting something like [(2)] one may express an aspect of one's credal state, without describing that state. One expresses one's confidence, that is, without literally saying that one is confident" ("Bayesian Expressivism," 125). Rothschild makes a similar proposal, that one suggests that conversational participants adopt a given credal state in "Expressing Credences," 103. See also Price, "Does 'Probably' Modify Sense?"; Yalcin, "Epistemic Modals," "Nonfactualism about Epistemic Modality," and "Context Probabilism"; Moss, "On the Semantics and Pragmatics of Epistemic Vocabulary" and *Probabilistic Knowledge*; Swanson, "The Application of Constraint Semantics to the Language of Subjective Uncertainty." See note 4 above for related discussion.

if any, indirect speech act you perform in uttering (2). As with most cases of indirect speech acts, which one is performed by a speaker who utters (2), if any, will be different in different contexts. There are many indirect speech acts that could occur in less typical contexts that I will not discuss.[13] But I will survey some common ones in this section, eventually coming to what you plausibly express by using (2) in the ordinary context at issue, Friday Eggs.

2.1. None

In some contexts, a speaker uses a sentence like (2) without performing any indirect speech acts related to their intentions:

> *Sleepwalker**: You know you are a predictable sleepwalker, but not quite as predictable as in Sleepwalker. When you nap, you *usually* sleepwalk to the store. You are about to nap, and your roommate asks where you will sleepwalk to. You reply with (2).

You express your high credence that you will go to the store, but you do not communicate anything indirectly about your intentions.

However, not all utterances of (2) are like this. Many *do* involve you indirectly communicating something about your intentions. To deny this would be to treat all utterances of (2) quite differently than utterances of (1), which very often involve you indirectly communicating something about your intentions.

2.2. Conveying Confidence in a Future Intention

One case of this sort occurs when a speaker uses (2) to communicate that they are confident they will, in the future, form an intention to go to the store:

> *Combos*: You currently think going to the store is a bad idea, so you do not have an intention to do so. But you think it is likely that you are going to have some drinks later, and you know that if you do have those drinks, buying a party-size bag of Combos will seem like the thing to do. So you think that if you come to be under the influence, you will intend to go to the store. Thus, you utter (2) to your roommate.

At the time of your utterance, you do not have an intention about what you will do. But it does seem that in addition to directly expressing your high credence that you will go to the store, you also indirectly express high credence that you

13 Even in ordinary contexts, we can perform all sorts of indirect speech acts, e.g., conveying that grandma is not feeling well, that the mechanic fixed the car, or even that I will not go to the store today (using sarcasm).

will come to intend to go to the store. That is, you indirectly communicate your current state of mind about what you will intend in the future.[14]

I suspect that cases like Combos are less common than cases where you express a current intention-like state. Friday Eggs seems to be one where you express something intention-like. When you reply to your roommate with (2), it does not seem that you are expressing or conveying uncertainty about either your present *or your future* state of mind. Furthermore, in Friday Eggs you are subject to some characteristic norms on intending when you utter (2), not merely later after making a further decision. For instance, after uttering (2) in Friday Eggs (but not in Combos), lending your car to a friend for the entire day would violate a norm if driving your car is a known necessary means to going to the store. This suggests that the intention that generates this norm (intend the necessary means to your end) is something you have at the time of utterance.[15]

2.3. Expressing a Conditional Intention

One reason to utter (2) rather than (1) is that you do not know whether things will work out for you to go to the store.[16] You might not know whether you will finish your other errands, or whether the store stays open until nine o'clock, or whether you will still have the motivation to go after dinner. Let us look at a slight variant of Friday Eggs:

> *Friday Eggs**: You bake a lot of cookies that week, and on Friday morning, your roommate notes that you are again out of eggs. Friday is busier than Sunday. You tell your roommate that you cannot make it to the store until 8:15 PM and that you suspect, but are not sure, that the store is open until 9:00 PM on Fridays. You then utter (2).

14 Thanks to a referee for helpfully suggesting a case with the structure of Combos, rather than my earlier attempts, which made this point less effectively. David Braun (personal communication) notes that there are cases where you do not presently intend to ϕ but believe you will come to intend to ϕ, while being confident, but not believing, that you will, when the time comes, ϕ based on that intention. Here, your uncertainty is about carrying out, rather than forming that intention. As in Combos, there seems something amiss. Either your present self views your future intention as irrational (as in Combos) or you violate a plausible reflection principle, where believing you will rationally have an intention in the future rationally requires you to have that intention now.

15 Thanks to Justin Snedegar for discussion.

16 Thanks to Laura Tomlinson Makin for suggesting and showing me a way of reasoning into the position discussed in this section. And thanks to Luis Rosa for discussion and for showing me that some of my previous arguments against this view were unconvincing.

It seems that in uttering (2) in Friday Eggs*, you indirectly express a *conditional* intention to go to the store today if the store stays open until nine o'clock. If you knew that it did, then you would have uttered (1) rather than (2).

It is plausible that there are conditional intentions, which are not judgments about what you will intend in the future, if you learn that the condition obtains. Rather, they are intention-like right now. Ludwig, for instance, thinks that as intentions are states that guide us in planning, conditional intentions are states that guide us in contingency planning.[17] And conditional intentions give rise to intention-like norms.[18] It is plausible that in uttering (2) in Friday Eggs*, you indirectly express this sort of attitude.

But you do not do so in the original Friday Eggs.[19] The key difference between Friday Eggs and Friday Eggs* is that in the latter you give your roommate much more information about what your future actions are contingent on. A natural view of what goes on in an instance of communication is that a speaker utters some bits of language in their surroundings to express or convey something, and a hearer processes the utterance with the surroundings also in mind to come to grasp what the speaker expressed or conveyed. A hearer who understands the words in (2) in Friday Eggs will not, in general, be able to figure out the condition of the purportedly expressed conditional intention. In hearing your utterance of (2) in Friday Eggs, will your roommate take you to have expressed a conditional intention to go to the store if you finish your other errands, or if the store stays open until nine o'clock, or if you still have the motivation to go after dinner, or if some other condition obtains?[20] As an

17 Ludwig, "What Are Conditional Intentions?"; see also Bratman, *Intention, Plans, and Practical Reason*; Lennertz, "Quantificational Attitudes."

18 Ferrero, "Conditional Intentions"; Ludwig, "What Are Conditional Intentions?"; Lennertz, "Quantificational Attitudes."

19 The considerations discussed in what follows are similar to those that Shpall uses to make the case against a conditional intention construal of what he calls inclinations (or partial intentions) ("The Calendar Paradox," 818–19).

20 One might think that the hearer will be able to figure out some plausible condition. But there are two worries here. First, this does not explain how a hearer might report your utterance of (2) to a third party, without sharing the details of the context of utterance. Nonetheless the third party can know what you communicated, suggesting that the condition, and, so, the conditional intention, was not essential to it. Second, though the hearer might be able to, in many cases, take up your utterance in a way that seems somewhat plausible, there is no good reason to think that they will grasp a particular condition that you intended to convey; in many cases, it seems that there *is not* any particular condition that you intended to convey. I suspect the most plausible way forward for the presser of this objection is to accept the heterodox idea that a speaker and hearer can communicate without the latter grasping what the former has expressed or conveyed. See Buchanan, "A Puzzle about Meaning and Communication."

analogy, suppose that you utter to a passerby on the street, "I am ready." The passerby has no way of knowing what you have expressed; are you ready for breakfast, for your job interview, for the apocalypse, for something else? This is not successful communication. If we are only focused on which conditional intention you express, we should expect a similar breakdown of communication in uttering (2) in a situation where there is not a lot of shared background information, like Friday Eggs (though not in Friday Eggs*, given your more robust shared beliefs).[21] But we do not see such a breakdown.

It is helpful to think by analogy to how we express beliefs. Consider again:

(3) The Sparks probably won last night.

Why might you utter (3) rather than (8)?

(8) The Sparks won last night.

One reason is that you do not know whether things turned out in the way that would make the Sparks win. You might not know whether their star player, Nneka Ogwumike, fouled out or whether they held onto their third-quarter lead. So instead of uttering (8), you utter (3). We might conclude that in uttering (3), you are expressing a conditional belief—perhaps the conditional belief that the Sparks won last night if Ogwumike did not foul out or the conditional belief that the Sparks won last night if they held onto their third-quarter lead. This suggestion might diagnose what you convey on particular occasions, but it is not plausible for all cases. Though your judgment that the Sparks probably won last night might be grounded in contingency reasoning about what happens if Ogwumike fouls out or if they blow their lead, these thoughts are not the judgment itself. So they are not what you typically express in uttering (3). Likewise, things are similar for the analogous case of conditional intentions. In our scenario, whatever intention-like state you indirectly express by uttering (2)

21 Perhaps what is indirectly conveyed is that there is some condition or other, such that you conditionally intend to go to the store if that condition obtains. Knowledge of this, together with what is directly expressed—your high credence that you will go to the store—might tell your roommate a lot about your mental state. I agree that worries about communication/coordination do not apply here. But I do not think this diagnosis suffices to explain the strength of the attitude you convey that you have in uttering (2) in Friday Eggs. This is because if you have a conditional intention whose condition does not obtain, you can neither succeed nor fail at carrying it out. See Ferrero, "Conditional Intentions"; Ludwig, "What Are Conditional Intentions?"; Lennertz, "Quantificational Attitudes." Ferrero calls such conditional intentions *moot* ("Conditional Intentions," 705). But it seems that by uttering (2) in Friday Eggs, you express the sort of attitude that can be carried out or not when the time comes, regardless of any conditions obtaining or not. Thanks to Luis Rosa for discussion.

in Friday Eggs may be supported by conditional intentions to go to the store if various contingencies do or do not obtain and your confidence that they will obtain. But these conditional intentions are not what you express in uttering (2) in Friday Eggs.[22]

2.4. Expressing a Partial Intention

If we continue to think by analogy to belief, we will be struck by another possibility for what you indirectly express in uttering (2) in Friday Eggs. It is commonly said that (2) is used to directly express a partial version of the state that (1) is used to directly express (belief).[23] Perhaps in many scenarios, including Friday Eggs, (2) is used to indirectly express a partial version of the state that (1) is used to indirectly express (intention).[24]

22 As an analogical consideration, this is not a knockdown argument. One disanalogy is that the credence expressed by a speaker who utters (3) is expressed *directly*, while a proponent of the claim that a speaker who utters (2) in Friday Eggs expresses a conditional intention says they do so only *indirectly*. However, it is not clear why this would make the conditional intention picture for (2) more plausible than the conditional belief one is for (3).

23 As I will discuss further in note 48 below, Moon shows that conceiving of these attitudes as partial beliefs is incorrect in "Beliefs Do Not Come in Degrees." But that is no impediment to the analogy in the text, if conceived of more carefully, as one where partial intentions relate to ordinary intentions as attitudes of confidence, or credences, relate to belief. It is worth noting that some authors take the heterodox position that (1) and (2) are, at least often, both used to express the same sort of attitude (belief), though the latter is toward a probabilistic content. See Lance, "Subjective Probability and Acceptance"; Hawthorne and Stanley, "Knowledge and Action"; Moss, *Probabilistic Knowledge*; Dogramaci, "Rational Credence through Reasoning"; Moon and Jackson, "Credence"; Lennertz, "Noncognitivism and the Frege-Geach Problem in Formal Epistemology."

24 Authors who have motivated the existence of partial intention by analogy to partial belief or credence have not considered using these states to explain what speakers express in using sentences like (2). See Chan, "A Not-so-Simple View of Intentional Action"; Holton, "Partial Belief, Partial Intention"; Shpall, "The Calendar Paradox"; Goldstein, "A Preface Paradox for Intention"; Beddor, "Fallibility for Expressivists." Jian casts doubt on the analogy between partial intentions and partial beliefs or credences in "Rational Norms for Degreed Intention (and the Discrepancy between Theoretical and Practical Reason)." Marušić and Schwenkler take themselves to be doing something similar, though their analogy is not to credences ("Intending Is Believing," 322–28). Instead, they take partial intentions to be either conditional intentions or what they call *weak intentions*. We have already discussed how the former relate to our case. As for the latter, Marušić and Schwenkler say that one has a weak intention when "she anticipates that she may abandon this intention in the face of difficult or tempting circumstances" (326). I suspect that, in contrasting weak from other intentions, they are overplaying the settledness of ordinary intentions and beliefs. It is rare to intend something regardless of what temptations arise. For instance, almost everything I intend to do would be reconsidered if I were offered a billion dollars to do something else. In a similar way, almost everything I believe would be reconsidered if I were to obtain overwhelming evidence against it. Ludwig and Bratman

This relies on there being such things as partial intentions. But what are partial intentions? Holton says they

> are certainly like all-out intentions in many respects. They play the same roles of curtailing deliberation, resolving indeterminacy, and enabling coordination that intentions play: you fix on a small number of plans from the many that occurred to you and that you might have pursued, and as a result of this you can coordinate around your other plans … and with other people. … What distinguishes the states you are in from normal intentions is simply that they are partial: they stand to all-out intentions much as partial beliefs stand to all-out beliefs.[25]

It seems reasonable that in the scenario described above, you have a partial intention of the sort characterized here (though, as I will discuss below, our scenario is not consistent with Holton's full account of partial intentions). For instance, you do not leave the question about what you will do today totally open. Furthermore, you indirectly express a state that induces some other requirements on you. Remember what we said above in the scenario where you require the car to get to the store, but loan it to your neighbor all day after sincerely uttering (2) to your roommate. Your roommate is likely to say "Dude, what the hell?" You appear to violate a sort of means-end consistency norm stemming from the attitude you expressed.

There is more to say about different conceptions of partial intentions, and there are reasons to doubt that, on some of these conceptions, you indirectly express a partial intention in uttering (2) in Friday Eggs. We will return to this topic in section 3. I will now investigate another promising account.

2.5. Expressing an Ordinary Intention

The sorts of considerations I discussed in the previous subsection do not differentiate two hypotheses for what you indirectly express in uttering (2) in Friday

---

offer similar considerations against Ferrero's picture of almost all intentions as conditional. See Ludwig, "What Are Conditional Intentions?"; Bratman, "Simple Intention." See Ferrero, "Conditional Intentions." This may mean that the fact that intentions are often weak in Marušić and Schwenkler's sense in similar ways to belief can help them avoid the sort of objection they address when they discuss weak and conditional intentions. See Bratman, *Intention, Plans, and Practical Reason*, 37; Mele, "Intention, Belief, and Intentional Action," 19–20. But it does not seem to help in our project of understanding what you express in uttering (2) in Friday Eggs. Finally, Nathaniel Baron-Schmitt suggested to me that you might express an attitude that is distinct from though similar in some ways to partial intention: (strongly) considering. Muñoz discusses this sense of considering and distinguishes it from partial intention in "Thinking, Acting, Considering" (255–56).

25 Holton, "Partial Belief, Partial Intention," 41.

Eggs: a partial intention to go to the store and an ordinary (full) intention to go to the store.[26]

Consider, again, the continuation of our scenario where you lend the car to your neighbor, and your roommate responds with "Dude, what the hell?" Notice your roommate's reaction could be the same if you had uttered (1) rather than (2). It seems that you violate the same sort of norm in each case. A straightforward explanation of this fact is that you do violate the same norm in each case because you have an intention-like attitude in each case, which generates that norm. That attitude might, as suggested in the previous section, be a partial intention. But the parallel of the cases suggests it might be more reasonable to think that it is an ordinary intention.[27]

3. COMPARING THE PARTIAL AND ORDINARY INTENTION DIAGNOSES

So what do you indirectly express in uttering (2) in Friday Eggs: a partial or an ordinary intention? In this section, I will compare the *partial intention diagnosis* and the *ordinary intention diagnosis*. I will show their relative advantages, though I will leave it to the reader to decide which is more plausible. Either way, as I will show in section 4, there is an important challenge for some natural principle about intention.

3.1. In Favor of the Partial Intention Diagnosis

A first-pass reason for preferring the partial intention diagnosis to the ordinary intention diagnosis of your utterance in Friday Eggs is easy to find. It seems you express something weaker when you utter (2) in Friday Eggs than when you utter (1) in Sunday Eggs. So we should think that you express a partial, rather than an ordinary intention. However, this first pass reason is not decisive on its own. The advocate of the ordinary intention diagnosis could explain the intuitions about differing strength in terms of what is directly expressed—high confidence rather than belief.

26 This is what Chan calls an intention *par excellence* ("A Not-so-Simple View of Intentional Action," 5).

27 My descriptions of the scenarios Sleepwalker, Sleepwalker*, Combos, and Friday Eggs* include more intricate and somewhat obscure details than my description of the original scenario, Friday Eggs. We might think that we could differentiate a partial intention version of Friday Eggs from an ordinary intention version if we filled in the details. That may be true. But in many ordinary situations there often are not that many relevant details known to the conversational participants. My interest in Friday Eggs is an interest in what is going on in these ordinary situations. Thanks to Jay Jian for discussion.

There might be reason to think not just that you express something weaker with (2) in Friday Eggs than with (1) in Sunday Eggs, but that the *intention* expressed in uttering (2) is weaker than the *intention* expressed in uttering (1). Consider, for instance, if something comes up and you do not go to the store. Some have suggested that it is natural for your roommate to confront you about not going to the store in the case of Sunday Eggs, but it seems less natural for them to do that in the case of Friday Eggs. I think we should distinguish two questions: (i) Did you carry out your intention? (ii) Are you blameworthy? What it takes for you to carry out your intention in both Friday Eggs and Sunday Eggs is the same: you go to the store. So that would not explain your roommate's variable response. But it might be that you are blameworthy in Sunday Eggs but not in Friday Eggs. A natural thought is that in Sunday Eggs you express an ordinary intention which involves a full commitment toward going to the store, while in Friday Eggs you express a partial intention that does not involve a full commitment. Your roommate might think that you should be blamed since you gave them reason to rely on you completely in Sunday Eggs, but not in Friday Eggs.[28]

While this sort of example is a reason in favor of the partial intention diagnosis, it does not refute the ordinary intention diagnosis. An advocate of the ordinary intention diagnosis might give a slightly amended explanation where what licenses your roommate to rely on you is not the strength of your commitment, which might be full in both cases, but the strength of the belief you express, which is full in Sunday Eggs but not in Friday Eggs. This explanation is consistent with the ordinary intention diagnosis. I lack a firm judgment about whether this explanation is as good as the one from the previous paragraph, so it is not clear to me whether the partial intention diagnosis has an advantage here and, if so, how strong it is.[29]

In personal communication, Nathaniel Baron-Schmitt has suggested another reason to prefer the partial intention diagnosis, which comes from looking at a different utterance in the same scenario as Friday Eggs:

> *Friday Eggs May*: On Friday morning, your roommate notes that you are again out of eggs. Friday is busier than Sunday, so you say:

28 Luis Rosa suggested this sort of case to me. Note that what stops you from going to the store must not be catastrophic or unforeseeable. A catastrophic or unforeseeable event would exonerate you in both cases.

29 This is because I lack a firm judgment about whether blame in Sunday Eggs, if you did not go to the store, comes from a negative evaluation of your failing to carry out your plan/commitment or from a negative evaluation of your asserting something false that your roommate relies on. If it is the latter, the ordinary intention diagnosis's defense from this objection will likely be satisfying. Thanks to Jay Jian for discussion.

(9) I may go to the store today.

Your utterance is felicitous, and if you express an intention, it seems implausible that it is an ordinary intention. We can see so by thinking about what would happen if you uttered (10) rather than (9) in Friday Eggs May:

(10) I may go to the store today, but I may not.

If sentences like (9) are used in scenarios like Friday Eggs May to indirectly express ordinary intentions, then it would seem that (10) would be used in Friday Eggs May to indirectly express contradictory intentions—to go to the store and to not go the store. But you could utter (10) without expressing contradictory intentions. The partial intention diagnosis could easily explain what is going on by saying that what you express is a partial, not ordinary intention—and a partial intention to go to the store and a partial intention to not go to the store can be consistent. Since we should want a diagnosis of what goes on in Friday Eggs that can also explain the very similar goings on in Friday Eggs May, it appears that the partial intention diagnosis has an advantage here.

There are two plausible responses that the ordinary intention diagnosis could give. First, they could deny that either (9) or (10) is used to express any sort of intention (at least in Friday Eggs May). One reason to think this is that if one says (9) and then goes to the store, it is not clear that we should say they carried out their intention.[30] Second, an advocate of the ordinary intention diagnosis could say that even though sentences like (9) are sometimes used to express ordinary intentions (in scenarios like Friday May Eggs), sentences like (10) never are. This might appear *ad hoc* since (10) is just the conjunction of two sentences with (9)'s form. But it is important to remember that the expression of an intention is an indirect speech act; it is a pragmatic, rather than an encoded, compositional phenomenon. And one thing that could surely cancel an implicature generated by a use of (9), even in a context that would otherwise support it, is the claim that the opposite may happen. If either of these explanations is correct, then the ordinary intention view may still be right.

3.2. In Favor of the Ordinary Intention Diagnosis

One reason for preferring the ordinary intention diagnosis to the partial intention diagnosis is parsimony. The partial intention view posits a new sort of

30 One worry with this response is that I have been told that (2) and (9) are translated to the same sentence in Mandarin. I do not have space to adequately explore the fallout of this here, but it seems to be a *prima facie* reason to think that if you express an intention in uttering (2) in Friday Eggs, you can do so by uttering (9) as well. Thanks to Jay Jian and Mengqun Sun for discussion.

mental state; our picture of the mind needs to include not only ordinary intentions but partial ones. However, theorists have posited partial intentions for reasons other than explaining what you express in uttering a sentence like (2) in a scenario like Friday Eggs. For instance, Chan argues that partial intentions make sense of variation in the stringency of consistency constraints on intentions, especially in cases like Bratman's video game player.[31] Marušić uses partial intentions to understand Bratman's example where he intends to stop at the bookstore but does not believe that he will.[32] Goldstein and Shpall use partial intentions to solve a preface paradox-like problem involving intentions.[33] These examples show that there are independent motivations for partial intentions.

A more serious worry for the partial intention diagnosis is that the belief analogy used to motivate it is not so strong.[34] For instance, partial intentions do not obviously come with the fineness of grain that credence does. Suppose that in Friday Eggs you utter, instead of (2):

(11) It is 75 percent probable that I will go to the store today.

You directly express a credence of degree 75 percent that you will go to the store today. But it is not clear that you indirectly express any sort of partial intention (e.g., if you do later go to the store, it does not sound natural to say that you have carried out your intention). And if you do express a partial intention, it is not clear that it is a state whose degree can be measured quantitatively, as 0.75.

A related puzzle is why, if there are partial intentions with an analogous structure to credences, it is difficult to indirectly express low-degreed intentions. For instance, imagine that in Friday Eggs you utter:

(12) It is improbable that I will go to the store today.[35]

31 Chan, "A Not-so-Simple View of Intentional Action"; Bratman, *Intention, Plans, and Practical Reason*, 113–15.

32 Marušić, *Evidence and Agency*, 58–63; Bratman, *Intention, Plans, and Practical Reason*, 37.

33 Goldstein, "A Preface Paradox for Intention"; Shpall, "The Calendar Paradox."

34 For extensive discussion of similar points, see Jian, "Rational Norms for Degreed Intention (and the Discrepancy between Theoretical and Practical Reason)."

35 The fact that we do not use sentences like (12) to express intention-like attitudes might be abductive reason not only against the existence of partial intentions, but also for a substantive condition forbidding ordinary intentions toward actions that you have low confidence that you will perform. Mele says:

> Ordinary speakers of English are disinclined to attribute intentions to *A* to agents who estimate their chances of succeeding in A-ing as less than even. What accounts for this, I suspect, is not just that there is something very odd about such *assertions* as "I intend to *A* but I believe that I probably will not *A*," but also that the ordinary concept of intention incorporates a confidence condition—perhaps only a negative one. ("Intention, Belief, and Intentional Action," 28)

If I do go, we should not say that I carried out my intention. These points suggest that either there are no partial intentions, we do not indirectly express them in this sort of language, or partial intentions do not have the fineness of grain of credences.[36] In the first case, we should agree that it is an ordinary intention that you indirectly expressed by uttering (2) in Friday Eggs. In the second case, we are left with a puzzle about why we can easily express some sorts of our partial intentions (strong, qualitative ones) but not others. In the third, partial intentions will not have quantitative strengths. I will not try to dissolve the puzzle generated if the second possibility is true, though I am not ruling out this way forward. Instead, I will investigate a popular picture of partial intentions from Richard Holton which denies that they have quantitative strengths.[37]

Holton maintains the analogy between belief and intention by claiming that neither intention nor belief comes in quantitative degrees. What makes an intention partial for him is that there is an alternative intention to achieve the same end. For instance, since your end is to procure eggs, in order for your intention to go to the store (as a means to getting eggs) to be partial, you must have an alternative intention designed to get eggs. But Marušić notes that this should not be a general requirement for having a partial intention.[38] It is not required in Friday Eggs. Your attitude directed at going to the store may be accompanied by no other attitudes that set getting eggs as a goal. So either that state is not a partial intention (and is likely an ordinary intention) or Holton is wrong that alternative intentions are required for an intention to be partial.

Thanks to Catherine Rioux for discussion.

36 Regarding the first possibility, Julia Staffel suggested to me that it is natural to say things like:

(13) I sort of intend to go to the store.

(14) I strongly intend to go to the store.

This appears easily explicable if we accept partial intentions. But we can also explain utterances of (13) and (14) without referring to partial intentions. First, "sort of" is not typically used to introduce degrees but to characterize situations where it is indeterminate whether a qualitative concept applies (e.g., "He is sort of bald," "I sort of understand what you are saying"). Second, though "strongly" often is a degree modifier, it can also signal stability or robustness. So when I say that I strongly believe my mother's testimony, I might mean that I believe it and would continue to do so even if evidence to the contrary were mounted against it; not only do I believe it, but I conditionally believe it, given all sorts of countervailing evidence. Likewise, strongly intending may simply be an intention that I am committed to in a way that is stable even were I to encounter strong countervailing reasons. This can be so even if the intention does not come in degrees.

37 This contrasts with Chan, "A Not-so-Simple View of Intentional Action"; Goldstein, "A Preface Paradox for Intention"; Beddor, "Fallibility for Expressivists."

38 Marušić, *Evidence and Agency Norms of Belief for Promising and Resolving*, 60; see also Archer, "Do We Need Partial Intentions?"

Another option says that a partial intention is distinguished by having a mere partial belief in its success.[39] But according to this view, the partial intention diagnosis of Friday Eggs seems indistinguishable in substance from the ordinary intention diagnosis. Both claim that (i) you directly express a high credence that you will go to the store and (ii) you indirectly express an attitude directed toward going to the store that can be carried out or not, structures your deliberation, and makes you subject to characteristic intention-like norms. They merely disagree on whether to call this an ordinary intention because of these features, or a partial intention because it is accompanied by a credence.[40]

4. The Relationship between Intention-Like and Belief-Like Attitudes

In section 3, we saw reasons to prefer the partial intention diagnosis of what you indirectly convey in uttering (2) in Friday Eggs to the ordinary intention diagnosis, and reasons for the opposite conclusion. Both are still open possibilities. Either way, we will have to reject at least one of the principles from the introduction, *Intention Implies Belief* and *Intention Implies Commitment*.

4.1. Consequences of the Ordinary Intention Diagnosis

Suppose the ordinary intention diagnosis of what you indirectly express in uttering (2) in Friday Eggs is correct. Let us look in more depth at what you communicate. Typically, if you directly express some degree of credence in a proposition, you imply that you do not believe that proposition (even though high confidence is compatible with belief). This is an implicature that arises due to Grice's Maxim of Quantity: "Make your contribution as informative as is required."[41] For example, it would be strange to say that most people came to the party when they all did. A hearer can reason that if everyone came to the party, you would have said so, so in saying that most people came, you imply that not all of them did. Likewise, it would be strange to express that you are pretty confident of a proposition when you believe it. In Friday Eggs, a hearer can reason that if you believed you would go to the store, you would have just

39 Chan develops a view like this in "A Not-so-Simple View of Intentional Action." Other authors consider and reject this view. See Holton, "Partial Belief, Partial Intention," 41–42; Shpall, "The Calendar Paradox," 822.

40 Chan entertains a similar objection to his view and gives a response in "A Not-so-Simple View of Intentional Action," 7–8. A full evaluation of this move is beyond the scope of this paper.

41 Grice, "Logic and Conversation," 26.

said (1). Given this, we should say the following about Friday Eggs: in uttering (2), you directly express high credence that you will go to the store, implicate that you do not believe that you will go to the store, and indirectly express an intention to go to the store.

You have thereby communicated a state of mind that is obviously inconsistent with one of our theses from the introduction. If you are sincere, then the following thesis must be false:

Intention Implies Belief: If *S* intends to ϕ, then *S* believes *S* will ϕ.

Some people who accept Intention Implies Belief do so because they think that intending to ϕ is or involves believing that you will ϕ.[42] But the ordinary intention diagnosis creates problems, even for those who reject this metaphysical claim, as long as they accept Intention Implies Belief.

Some might think that intention does not imply belief, but *rational* intention does:

Rational Intention Implies Belief: If *S* is rational and intends to ϕ, then *S* believes *S* will ϕ.

This would be shown to be false by the ordinary intention diagnosis of Friday Eggs provided that you were sincerely expressing a *rational* combination of attitudes in uttering (2). This seems reasonable, given the ease with which we utter and accept others' utterances of sentences of the form "I will probably ϕ."[43]

This does not mean that there are no connections between intention and belief-like attitudes. The ordinary intention diagnosis is consistent with (though it does not entail or even strongly support) the following three views:

Intention Implies High Confidence: If *S* intends to ϕ, then *S* has high confidence that *S* will ϕ.[44]

42 Marušić and Schwenkler, "Intending Is Believing"; Harman, "Practical Reasoning."

43 Those with cognitivist sympathies may wonder how we are to explain the rational norms on intention if we reject these theses. My project here is not to vindicate any explanation of those norms, but they are right that we would to need to avail ourselves of a non-cognitivist explanation as in Bratman, *Intention, Plans, and Practical Reason*. This sort of challenge to noncognitivism is commonly made in the metaethics literature. See Hale, "Can There Be a Logic of Attitudes?"; van Roojen, "Expressivism and Irrationality"; Schroeder, *Being For*. And there is a similar advantage for cognitivist accounts of confidence or credence in formal epistemology. See Moss, *Probabilistic Knowledge*; Lennertz, "Noncognitivism and the Frege-Geach Problem in Formal Epistemology." Thanks to Catherine Rioux for discussion.

44 Holton discusses, with some pessimism, a similar thesis involving his notion of partial belief in "Partial Belief, Partial Intention." In "Instrumental Rationality," Wedgwood

> *Intention Implies Lack of Disbelief*: If *S* intends to ϕ, then *S* does not believe *S* will not ϕ.[45]
>
> *Intention Implies Belief in Possibility*: If *S* intends to ϕ, then *S* believes it is possible that *S* will ϕ.[46]

Whether and which of these is correct is not determined by our investigation of utterances of sentences like (2) in situations like Friday Eggs. Perhaps there are other data about communication that can cast light on these theses, or perhaps their status must be decided independently of data about how we express our intentions.

4.2. Consequences of the Partial Intention Diagnosis

Now suppose, instead, that the partial intention diagnosis is correct. On one reading of Intention Implies Belief, where "intends" picks out any intention, partial or full, but "believes" picks out only ordinary beliefs, it is false for the reasons discussed in the previous section. Nonetheless, we might think that a graded version of this thesis is more plausible:

> *Intention Implies Belief**: If *S* intends, to degree *n*, to ϕ, then *S* has credence, of degree *n*, that *S* will ϕ.

This may be a plausible principle that is in line with the motivations of at least some advocates of the original principle (though I am skeptical it would satisfy some, for the reasons discussed below).[47] Because of this complication, I want to spend this section discussing the other principle from the introduction:

---

endorses a principle that stands in relation to Intention Implies High Confidence as Rational Intention Implies Belief stands to Intention Implies Belief:

> *Rational Intention Implies High Confidence*: If *S* is rational and intends to ϕ, then *S* has high confidence that *S* will ϕ.

And Setiya proposes a different relationship between intention and confidence: "In doing ϕ intentionally, one is more confident that one is doing it than one would otherwise be" ("Practical Knowledge," 391).

45 Bratman, *Intention, Plans, and Practical Reason*; Mele, "Intention, Belief, and Intentional Action."

46 Wallace, "Normativity, Commitment and Instrumental Reason."

47 "Intention Implies Belief*" is a misleading name for this principle. The claim that belief comes in degrees or there are partial beliefs, despite being a popular way of speaking, is not *strictly speaking* plausible. Moon mounts convincing arguments against it in "Beliefs Do Not Come in Degrees." We, of course, have states of confidence or credences that are like beliefs in some ways (though different in others). Indeed, the huge literature on the relationship between credence and belief signals tacit acceptance that credences are not mere partial versions of ordinary beliefs. See, for instance, Jackson, "The Relationship

Intention Implies Commitment: If *S* intends to ϕ, then *S* is committed to ϕ-ing.

If we accept the partial intention diagnosis, then a traditional reading of this principle is false.

To see why, let us think about what it is for *S* to be committed to ϕ-ing. Marušić and Schwenkler suggest it is a truth commitment to the proposition that *S* will ϕ.[48] Those who accept this notion of the commitment implied by intention often also accept Intention Implies Belief. Indeed, on this conception Intention Implies Commitment does not tell us much more than Intention Implies Belief. But, as Velleman notes, for theorists of a cognitivist persuasion, this sense of commitment plays the role of intention.[49] Because it is a truth commitment to something that is up to the agent, it has the functional role of a commitment to an action.

We can follow Bratman in characterizing a commitment to an action, ϕ, as having a certain functional role.[50] This role includes, at least, structuring our actions as we approach the time to ϕ and guiding our practical reasoning about whether to ϕ, and introducing norms on both of these. If we are committed to ϕ-ing and we do not change our minds, then as it becomes time to ϕ, we should and will tend to do so. And if we are committed to ϕ-ing, we should not and will tend not to reconsider whether to ϕ.[51]

Let us suppose that the partial intention diagnosis is correct—that you indirectly express a partial intention to go to the store in uttering (2) in Friday Eggs. Then it seems that you are not committed to (or settled on) going to the store in the sense just discussed. For instance, it may become time to go to the store, but still, you might not do it. And you may reconsider whether to go to the store in the interim. Furthermore, a defender of the partial intention diagnosis will claim in some cases that not going, or reconsidering, does not

---

Between Belief and Credence." We can, of course, use terms in ways we want, provided that we do not confuse an ordinary notion like belief with a technical one like partial belief. So we should realize that Intention Implies Belief* is not an obvious or innocuous generalization from Intention Implies Belief but is a substantive principle that requires independent motivation.

48 They say, "When we intend to do something, just as when we believe something, we have made a commitment: we have settled a question or reached a conclusion" ("Intending Is Believing," 321). See also Velleman, "What Good Is a Will?," 209–10.

49 Velleman, "What Good Is a Will?," 210.

50 Bratman, *Intention, Plans, and Practical Reason*, 107–10.

51 These features of commitment may also be shared by Mele's notion that an intention to ϕ implies that one is *settled* on ϕ-ing. See Mele, "Intention, Belief, and Intentional Action"; Hieronymi, "Controlling Attitudes."

directly violate any norms.[52] So we have a scenario where you have an intention to go to the store but are not committed to doing so. Intention Implies Commitment is false.

Here are two natural replies that result. First, those who avow Intention Implies Commitment might not mean to be talking about partial intentions, but only ordinary ones. So they might retain the thesis in this way. This is fine, but it then is difficult to make sense of how partial intentions help with our original problem. In what sense are partial intentions *intentions* at all if they do not have some of the central features of intentions, which come from intentions being commitments?

The second reply accepts that what makes intentions partial is that the corresponding commitments are partial.[53] More precisely:

> *Intention Implies Commitment**: If *S* intends, to degree *n*, to ϕ, then *S* is committed, to degree *n*, to ϕ-ing.

Degree of intention mirrors degree of commitment. I want to note, however, that our ordinary notion of commitment does not come in degrees. When we think about ordinary commitments, like one to pick up the kids from practice or to give up chocolate for Lent or to love and cherish until death do us part, we are thinking of ordinary, all-or-nothing states. It does not make sense to talk about a partial commitment to pick up the kids or to refrain from eating chocolate or to marry.[54] Goldstein attempts to generalize Bratman's notion of commitment so that it can be partial.[55] Such a picture might be a right, but it is quite far from the standard one in which intentions are attitudes that commit you, in the ordinary sense, to reasoning and acting in certain ways.

52 An anonymous reviewer insightfully notes that this can depend on the reason that one's intention is partial. In many cases a partial intention will rule out reconsideration. The partialness of the intention in these cases comes not from openness to reconsideration, but from uncertainty about how the world is—in our case about whether the store will be open when I am able to go. I think the reviewer has hit on an extremely important distinction between different ways that an intention might be partial—one that they note has normative consequences. I do not have space here to fully explore this, though I hope that it will be taken up in future work—both my own and others'.

53 Advocates of partial intentions often talk about them as partial commitments. See Shpall, "The Calendar Paradox"; Goldstein, "A Preface Paradox for Intention"; Beddor, "Fallibility for Expressivists," 771; Jian, "Rational Norms for Degreed Intention (and the Discrepancy between Theoretical and Practical Reason)."

54 There are commitments we might describe as weaker or more measured, like one to give up chocolate for Lent, except on Fridays, or to love and cherish until death—or substantive and important differences in our life goals—do us part. But these differ not in the strength or degree of commitment but in what we are committed to.

55 Goldstein, "A Preface Paradox for Intention," 6.

I suspect that many would agree on this count with Wedgwood's remark that "while there are degrees of belief, there are no degrees of intention."[56] What we have seen here is that accepting the partial intention diagnosis of what you express in uttering (2) in Friday Eggs requires accepting partial intentions in a robust sense—where they lack the commitment or settledness of ordinary intentions. (For instance, the account we discussed above where partial and ordinary intentions were distinguished merely by the strength of the accompanying belief/credence would not do here.) This requires a richer and more fine-grained picture of intentions and their features.[57]

For simplicity, I will continue to talk as if Intention Implies Commitment is false if the partial intention diagnosis is right. But the reader should keep the preceding caveats in mind.

4.3. Shared Consequences

One consequence of either diagnosis of what you convey in uttering (2) in Friday Eggs is a plausible response to a linguistically based argument in favor of Intention Implies Belief. Velleman argues that if we deny a tight connection between belief and intention, we are left with a puzzle about why the natural way to express an intention to ϕ is to say we will or are going to ϕ. For if intention did not imply belief, then we should be able to felicitously utter sentences which come out to be Moore paradoxical:[58]

(15) I will go to the store today, and (but) I do not believe that I will.

Velleman is partly right in that sentences of the form "I will *F*" are natural expressions of intentions. But this is only the natural form of expressing an intention to ϕ when the speaker also believes that they will ϕ. Regardless of whether one accepts Intention Implies Belief, they should say, as we did in section 1, that if you use a sentence like (1), you directly express a belief to go to the store:

(1) I will go to the store today.

56 Wedgwood, "Instrumental Rationality," 302.

57 Some authors who advocate for partial intentions explicitly develop such a picture and note that it might seem "radical." See Shpall, "The Calendar Paradox," 802–3. Shpall calls such partially committed states *inclinations*, where being inclined contrasts with being *settled*. For arguments that the picture for partial intentions will have to be quite different than the one for partial beliefs or credences, see Jian, "Rational Norms for Degreed Intention (and the Discrepancy between Theoretical and Practical Reason)." This is in contrast to Goldstein, "A Preface Paradox for Intention"; Shpall, "The Calendar Paradox."

58 Velleman, "What Good Is a Will?," 206–7; Grice, "Intention and Uncertainty," 269.

This explains why (15) sounds Moore paradoxical; it *is* Moore paradoxical. But there is a different natural way of expressing an intention, whether ordinary or partial, toward ϕ, when the speaker does not believe that they will ϕ, but merely has confidence that they will. A natural form for doing so is "I will probably ϕ."[59]

Nonetheless, we might think that the following also does not sound felicitous, though it seems predicted to be so if either of our diagnoses is true:

(16) I'll probably go to the store today, and (but) I don't believe that I will.

I agree that this is strange, but it is because of the tendency to read "I do not believe *p*" as "I believe not-*p*" (e.g., "I do not believe you are telling the truth" is usually read as "I believe you are not telling the truth"). We can avoid this complication by eschewing the "do not believe" construction and relying on the idea that belief rules out alternative possibilities and chances, as well as uncertainty, while mere credence does not. The contrast is then stark between the cases the objectors think are bad and the cases where intention is expressed along with uncertainty:

(15′) I will go to the store today, and (but) I might not/there is a chance I will not/I am not sure I will.

(16′) I will probably go to the store today, and (but) I might not/there is a chance I will not/I am not sure I will.

In either diagnosis, the felicity of (16′) suggests that Velleman's argument for Intention Implies Belief should not be convincing.[60]

5. Conclusion

A quick look at uses of first personal future-tensed sentences like (1) suggests that intending to do something implies believing that you will do it and being committed to (or settled upon) doing it. But we have seen that similar but overlooked sentences like (2) are used in ordinary scenarios in ways that are inconsistent with at least one of Intention Implies Belief and Intention Implies

59 Holton gives a similar defense but does not recognize the naturalness of the "I will probably ϕ" construction: "Where that belief is lacking, intention is more naturally reported by saying that one intends to act (or that one will try to act, if the act of trying can be separated out), often with a qualification that one is unsure of success" ("Partial Belief, Partial Intention," 52). See also Williams "Deciding to Believe," 138.

60 Thanks to Justin Snedegar for discussion of these nuances. And thanks to Aness Webster for leading me to realize that both diagnoses can explain the range of data related to Velleman's argument.

Commitment. In one plausible diagnosis, the speaker of (2) expresses an intention to go to the store but does not believe that they will. In the other plausible diagnosis, the speaker's intention to go to the store is partial and, so, does not commit them to going. Either way, some piece of the natural, first-pass picture of intentions is incorrect.[61]

Colgate University
blennertz@colgate.edu

REFERENCES

Anscombe, G. E.M. *Intention*. Cambridge, MA: Harvard University Press, 1957.

Archer, Avery. "Do We Need Partial Intentions?" *Philosophia* 45, no. 3 (April 2017): 995–1005.

Asarnow, Samuel. "Noncognitivism in Metaethics and the Philosophy of Action." *Erkenntnis* 88, no. 1 (January 2023): 95–115.

Bach, Kent, and Robert M. Harnish. *Linguistic Communication and Speech Acts*. Cambridge, MA: MIT Press, 1979.

Beddor, Bob. "Fallibility for Expressivists." *Australasian Journal of Philosophy* 98, no. 4 (2020): 763–77.

Bratman, Michael E. "Intention, Belief, Practical, Theoretical." In *Spheres of Reason*: *New Essays in the Philosophy of Normativity*, edited by Simon Robertson, 29–61. Oxford: Oxford University Press, 2009.

———. *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press, 1987.

———. "Simple Intention." *Philosophical Studies* 36, no. 3 (October 1979): 245–59.

Brunero, John. "Against Cognitivism about Practical Rationality." *Philosophical Studies* 146, no. 3 (December 2009): 311–25.

Buchanan, Ray. "A Puzzle about Meaning and Communication." *Noûs* 44, no. 2 (June 2010): 340–71.

Chan, David K. "A Not-so-Simple View of Intentional Action." *Pacific Philosophical Quarterly* 80, no. 1 (March 1999): 1–16.

61 Thanks to Jay Jian, Justin Snedegar, Laura Tomlinson Makin, Aness Webster, and Aaron Wolf for comments and discussion. An ancestor of this paper was presented at a session of the 2021 Central APA. Thanks to my commenters, Nathaniel Baron-Schmitt and Luis Rosa, and to the attendees of that session, particularly David Braun, David Chan, Jay Jian, Catherine Rioux, and Julia Staffel. The discussions springing from that session led to a complete reframing of the paper.

Christensen, David. *Putting Logic in Its Place: Formal Constraints on Rational Belief*. Oxford: Oxford University Press, 2004.

Davidson, Donald. "Intending." In *Philosophy of History and Action*, edited by Yirmiahu Yovel, 41–60. London: D. Reidel Publishing Company, 1974.

Dogramaci, Sinan. "Rational Credence through Reasoning." *Philosophers' Imprint* 18, no. 11 (May 2018): 1–25.

Dowell, J. L. "A Flexible Contextualist Account of Epistemic Modals." *Philosophers' Imprint* 11, no. 14 (November 2011): 1–25.

Eriksson, Lina, and Alan Hájek. "What Are Degrees of Belief?" *Studia Logica* 86, no. 2 (July 2007): 183–213.

Ferrero, Luca. "Conditional Intentions." *Noûs* 43, no. 4 (December 2009): 700–741.

Frankish, Keith. "Partial Belief and Flat-out Belief." In *Degrees of Belief*, edited by Franz Huber and Christoph Schmidt-Petri, 75–93. Berlin: Springer, 2009.

Goldstein, Simon. "A Preface Paradox for Intention." *Philosophers' Imprint* 16, no. 14 (July 2016): 1–20.

Grice, H. P. "Intention and Uncertainty." *Proceedings of the British Academy* 57 (1971): 263–79.

———. "Logic and Conversation." In *Studies in the Way of Words*, 22–40. Cambridge, MA: Harvard University Press, 1989.

Hale, Bob. "Can There Be a Logic of Attitudes?" In *Reality, Representation, and Projection*, edited by John Haldane and Crispin Wright, 337–63. Oxford: Oxford University Press, 1993.

Harman, Gilbert. "Practical Reasoning." *The Review of Metaphysics* 29, no. 3 (March 1976): 431–63.

Hawthorne, John, and Jason Stanley. "Knowledge and Action." *Journal of Philosophy* 105, no. 10 (October 2008): 571–90.

Hieronymi, Pamela. "Controlling Attitudes." *Pacific Philosophical Quarterly* 87, no. 1 (March 2006): 45–74.

Holton, Richard. "Partial Belief, Partial Intention." *Mind* 117, no. 465 (January 2008): 27–58.

Jackson, Elizabeth. "The Relationship between Belief and Credence." *Philosophy Compass* 15, no. 6 (June 2020): 1–13.

Jian, Jay. "Rational Norms for Degreed Intention (and the Discrepancy between Theoretical and Practical Reason)." *Australasian Journal of Philosophy* 101, no. 2 (2023): 360–74.

Kolodny, Niko. "The Myth of Practical Consistency." *European Journal of Philosophy* 16, no. 3 (December 2008): 366–402.

Konek, Jason. "Probabilistic Knowledge and Cognitive Ability." *Philosophical Review* 125, no. 4 (October 2016): 509–87.

Lance, Mark Norris. "Subjective Probability and Acceptance." *Philosophical Studies* 77, no. 1 (January 1995): 147–79.

Lennertz, Benjamin. "Noncognitivism and the Frege-Geach Problem in Formal Epistemology." *Philosophy and Phenomenological Research* 102, no. 1 (2021): 184–208.

———. "Quantificational Attitudes." *Journal of Philosophy* 118, no. 11 (November 2021): 585–613.

———. "Reasoning with Uncertainty and Epistemic Modals." PhD diss., University of Southern California, 2014.

Ludwig, Kirk. "What Are Conditional Intentions?" *Methode: Analytic Perspectives* 4, no. 6 (2015): 30–60.

Maher, Patrick. "The Irrelevance of Belief to Rational Action." *Erkenntnis* 24, no. 3 (May 1986): 363–84.

Marušić, Berislav. *Evidence and Agency: Norms of Belief for Promising and Resolving*. Oxford: Oxford University Press, 2015.

Marušić, Berislav, and John Schwenkler. "Intending Is Believing: A Defense of Strong Cognitivism." *Analytic Philosophy* 59, no. 3 (September 2018): 309–40.

Mele, Alfred R. "Intention, Belief, and Intentional Action." *American Philosophical Quarterly* 26, no. 1 (January 1989): 19–30.

Moon, Andrew. "Beliefs Do Not Come in Degrees." *Canadian Journal of Philosophy* 47, no. 6 (2017): 760–78.

Moon, Andrew, and Elizabeth Jackson. "Credence: A Belief-First Approach." *Canadian Journal of Philosophy* 50, no. 5 (2020): 652–69.

Moss, Sarah. "On the Semantics and Pragmatics of Epistemic Vocabulary." *Semantics and Pragmatics* 8 (March 2015): 1-81.

———. *Probabilistic Knowledge*. Oxford: Oxford University Press, 2018.

Muñoz, Daniel. "Thinking, Acting, Considering." *Australasian Journal of Philosophy* 96, no. 2 (2018): 255–70.

Price, Huw. "Conditional Credence." *Mind* 95, no. 377 (January 1986): 18–36.

———. "Does 'Probably' Modify Sense?" *Australasian Journal of Philosophy* 61, no. 4 (December 1983): 396–408.

Van Roojen, Mark. "Expressivism and Irrationality." *The Philosophical Review* 105, no. 3 (July 1996): 311–35.

Ross, Jacob. "Acceptance and Practical Reason." PhD diss., Rutgers University, 2006.

Rothschild, Daniel. "Expressing Credences." *Proceedings of the Aristotelian Society* 112, no. 1 (April 2012): 99–114.

Schroeder, Mark. *Being For: Evaluating the Semantic Program of Expressivism*. Oxford: Oxford University Press, 2008.

Setiya, Kieran. "Cognitivism about Instrumental Reason." *Ethics* 117, no. 4 (July 2007): 649–73.

———. "Practical Knowledge." *Ethics* 118, no. 3 (April 2008): 388–409.

Shpall, Sam. "The Calendar Paradox." *Philosophical Studies* 173, no. 3 (March 2016): 801–25.

Staffel, Julia. "Can There Be Reasoning with Degrees of Belief?" *Synthese* 190, no. 16 (November 2013): 3535–51.

Swanson, Eric. "The Application of Constraint Semantics to the Language of Subjective Uncertainty." *Journal of Philosophical Logic* 45, no. 2 (April 2016): 121–46.

Travis, Charles. *Occasion-Sensitivity: Selected Essays*. Oxford: Oxford University Press, 2008.

Velleman, J. David. "What Good Is a Will?" In *Rational and Social Agency: The Philosophy of Michael Bratman*, edited by Manuel Vargas and Gideon Yaffe, 83–105. Oxford: Oxford University Press, 2014.

Wallace, R. Jay. "Normativity, Commitment, and Instrumental Reason." *Philosophers' Imprint* 1, no. 4 (December 2001): 1–26.

Wedgwood, Ralph. "Instrumental Rationality." In *Oxford Studies in Metaethics*, vol. 6, edited by Russ Shafer-Landau, 280–309. Oxford: Oxford University Press, 2011.

Williams, Bernard. "Deciding to Believe." In *Problems of the Self: Philosophical Papers 1956–1972*, 136–51. Cambridge: Cambridge University Press, 1973.

Yalcin, Seth. "Bayesian Expressivism." *Proceedings of the Aristotelian Society* 112, no. 2 (July 2012): 123–60.

———. "Context Probabilism." In *Logic, Language and Meaning*, edited by Maria Aloni, Vadim Kimmelman, Floris Roelofsen, Galit W. Sassoon, Katrin Schulz, and Matthijs Westera, 12–21. Berlin: Springer, 2012.

———. "Epistemic Modals." *Mind* 116, no. 464 (October 2007): 983–1026.

———. "Nonfactualism about Epistemic Modality." In *Epistemic Modality*, edited by Andy Egan and Brian Weatherson, 295–332. Oxford: Oxford University Press, 2011.

Journal of Ethics and Social Philosophy
VOL. 25 NO. 3 · SEPTEMBER 2023
https://doi.org/10.26556/jesp.v25i3.2429

NATURALIZING MORAL NATURALISM

Jessica Isserow

One of the most pressing tasks for metaethicists is that of solving the *location problem*: finding a home for morality in the natural world. It goes without saying that some have risen to the occasion more enthusiastically than others, and it is one enthusiast in particular that shall occupy my attention here. The naturalist moral realist affirms continuity between ethics and the empirical sciences, striving to integrate her metaethics with the outputs of scientific theorizing. To her mind, moral epistemology does well to take science as its guide; moral facts are ripe for empirical investigation.[1]

Unfortunately, the naturalist canon does not always reflect these noble ambitions.[2] The naturalist is committed to letting the world do (much of) the talking. But so far, she has scarcely given it the chance to speak. My aim here is to set us back on course. The organizing theme of this paper is that the outputs of empirical investigations are of underrecognized significance for the moral naturalist. Its more specific contention is that these empirical resources help her to address two fundamental challenges that she faces.

Moral naturalists are often said to have trouble accommodating the *intensional* and *extensional* character of morality.[3] A metaethical position accommodates morality's *intensional* character just in case it is in keeping with (what are commonly regarded as) important conceptual commitments of moral thought and talk. Moral naturalism seems to fail dismally in this regard, for it is famously unfaithful to what many take to be a core conceptual commitment of moral discourse: that all agents have reason to act as morality requires independently of their contingent ends. Indeed, naturalists usually take an agent's reasons

1 For representative declarations of these commitments, see Boyd, "How to Be a Moral Realist"; Railton, "Naturalism and Prescriptivity." Different naturalists will admittedly embrace these commitments to different degrees (see section 2 below).

2 Hereafter, I substitute "(moral) naturalist" and "(moral) realist" for the more cumbersome "naturalist moral realist." There are obviously other sorts of naturalists and other sorts of realists, but they are not my focus here.

3 I borrow the distinction from Southwood (*Contractualism and the Foundations of Morality*), who uses it to assess different varieties of moral contractualism.

to be moral to be *hostage to* such ends. Naturalists' critics allege that this outlook fails to take morality seriously as a normative phenomenon. Call this the *intensional challenge*.[4]

A metaethical position accommodates morality's *extensional* character just in case it (largely) accords with substantive judgments concerning the extension of terms such as "morally required" and "morally impermissible"—for example, the judgment that it is morally impermissible to subject people to inhumane treatment on account of their skin color. Here again, the naturalist seems to come up short. This is because (as I explain below) her method for identifying which natural properties are (or constitute) the moral ones is fairly *permissive*; it seems to allow for moral truths that conflict with our substantive moral judgments. Of course, no metaethical theory can plausibly be expected to take *none* of these judgments to be mistaken or confused. Properly understood, then, the concern is not simply that the naturalist allows for moral truths that conflict with these judgments, but that she allows for moral truths that conflict with them in rather *striking* ways. Call this the *extensional challenge*.

As I conceive of these challenges, their upshot is as follows: the naturalist has incurred significant explanatory debt to date, and it is imperative that she either pay off this debt or discharge it. The naturalist could pay off her debt by demonstrating that she can indeed accommodate the intensional and extensional dimensions of morality. Alternatively, she could discharge her debt by establishing that the phenomena she fails to accommodate are not properly viewed as central to either dimension. To my mind, the naturalist has not exercised her full potential in either regard, for she is yet to fully avail herself of the resources at her disposal—insights from evolutionary theory, psychology, and ethnography in particular. I will argue that these resources help her to address both challenges in a more satisfying way. This is not to peddle the radical thesis that metaethics is a battle best fought on empirical ground. But it does, I think, demonstrate what we stand to gain by covering multiple terrains in our philosophical pursuits.

My first order of business will be to spell out the commitments of moral naturalism (section 1). I will then turn my attention to the intensional challenge (section 2). Here, naturalists have traditionally responded with an optimistic prognosis: given widespread human interests and concerns, many of us do as a matter of fact have reason to be moral—enough of us to vindicate intuitions concerning morality's normative credentials. No doubt, this prognosis has high

4 My locution (*the* intensional challenge) is admittedly misleading; this is certainly not the *only* sort of intensional challenge that naturalists face. Naturalists have also been pressed for not securing the right kind of objectivity, for example. See Loeb, "Gastronomic Realism"; Kurth, "What Do Our Critical Practices Say about the Nature of Morality?" I am picking my battles here.

intuitive plausibility. But it is only a start. Claims about human interests are, after all, *empirical claims*—and the naturalist's claim still remains in need of sustained empirical attention. My own objective will be to defend this claim as a robust empirical generalization. In section 3, I tackle the extensional challenge. Many worry that the naturalist's method for identifying which natural properties are moral properties commits her to classifying as morally permissible a range of behaviors that we regard as morally perverse. I shall argue that the naturalist can distinguish perverse moral frameworks from legitimate ones on principled grounds. Nothing I say entails that moral naturalism is home and dry. But my arguments do suggest that the position has far more going for it than many have thought.

1. MORAL REALISM, NATURALLY

A moral naturalist takes moral properties to be natural properties.[5] My discussion will be restricted to *realist* varieties of moral naturalism, according to which moral judgments are beliefs, there are moral properties, and these properties are constituted by or identical to mind-independent natural properties. The notion of mind independence can be tricky to tie down. For my part, I take it to be best captured by the thought that moral truths hold independently of our attitudes in the sense that they are not constituted by our beliefs or opinions about them.[6] It is trickier still to tie down the notion of a natural property. I will work with an understanding in which natural properties are those susceptible to empirical investigation.[7] Nothing of great import will hang on this understanding. I provide it in the interest of situating the discussion.

There are two distinguishing commitments of moral naturalism that will be relevant to my purposes. I do not claim that each is necessary to qualify as a member of the camp. In the absence of both, however, I think it is safe to say that the position in question would be one that parted ways from the majority of the naturalist canon. The first commitment is rooted in a particular approach to normativity. Traditionally, the naturalist takes an agent's (intrinsic) desires to be the ultimate source of her reasons for action.[8] What fundamentally grounds

5 I intend this characterization to include nonreductive naturalists, who take moral properties to be natural properties that supervene on but are not reducible to other (nonmoral) natural properties.

6 See Brink, *Moral Realism and the Foundations of Ethics*, 18–20.

7 Railton, "Naturalism and Prescriptivity," 154; Copp, "Why Naturalism?" 185. Cf. Sturgeon, "Moore on Ethical Naturalism," 538.

8 I am here skirting around different (and no doubt more sophisticated) articulations of this commitment. Rather than speak of an agent's (intrinsic) desires, the naturalist may prefer

an agent's normative reasons—and, in turn, her (ir)rationality—is what she fundamentally cares about. When paired with the naturalist's characterization of moral facts themselves, this approach yields a contingent relationship between moral requirements and an agent's normative reasons. Moral facts are mind-independent natural facts that bear no essential connection to an agent's desires. Whether or not morality provides her with normative reasons is thus completely hostage to whether she cares about the ends at which morality aims, or the moral good as such.[9]

A word of caution here. To claim that morality is not necessarily a source of normative reasons is not to claim that morality is not necessarily a source of *moral* reasons. The naturalist can readily admit that there are moral reasons in the sense that there is a system of requirements that sanctions morally good behavior. But this is not saying much. (There are also reasons of etiquette in this sense.) In denying that morality is necessarily a source of normative reasons, the naturalist is denying that moral requirements have intrinsic reason-giving force—denying that any agent would be irrational, or at least guilty of a normative mistake, were she unresponsive to them.[10] Obviously, none of this entails that the naturalist will look upon moral failures approvingly. She will, however, resist describing them as *rational* failures. Henceforth, I shall refer to this commitment as *no necessary irrationality in immorality*.

Second, the naturalist is committed to a certain permissiveness concerning the matter of which natural properties are the moral ones—though some care must be taken in spelling out the precise sense of permissiveness at issue. To this end, it is helpful to consider the naturalist's recipe for identifying which natural properties are moral properties. Often, she begins with the observation that morality has one or more distinctive functions in human life—making flourishing societies possible or stabilizing cooperation, for example.[11] She then draws upon these in sorting the moral from the nonmoral. Moral codes are said to be those that best serve a "society's needs and non-moral values."[12]

---

to say that it is her needs and values or her ends that ground her normative reasons. (See Copp, *Morality, Normativity, and Society*, ch. 9; Railton, "Moral Realism.") Either could be substituted for "intrinsic desires" without obscuring the basic point.

9 Railton, "Moral Realism," 166; Brink, *Moral Realism and the Foundation of Ethics*, ch. 3; Boyd, "How to Be a Moral Realist," 340–42.

10 These remarks reflect the well-known distinction between normativity in the rule-involving and reasons-providing senses. See Parfit, *On What Matters*, 2:267–68.

11 See Copp, *Morality, Normativity, and Society*; Sterelny and Fraser, "Evolution and Moral Realism."

12 Copp, *Morality, Normativity, and Society*, 159–60.

Moral facts are said to be facts about "human cooperation and the social practices that support [it]."[13]

Claims such as these reflect claims about which natural properties the moral properties are (or are likely to be) given the roles they are usually taken to play. Sometimes, such claims are presented as the product of conceptual analysis.[14] Other times, they are put forward in the spirit of an empirical hypothesis. This is a choice point that marks a divide between *a priori* and *a posteriori* varieties of moral naturalism. For my purposes, I need not take sides. What it is important to appreciate is that on neither outlook do these claims decide all first-order moral issues in advance. It remains an open question what the moral truths actually are—which moral code does as a matter of fact best promote human flourishing or cooperation, say. What the moral truths actually are remains an open empirical question, one that ultimately hinges upon what the world turns out to be like—let the empirical chips fall where they may. I will refer to this latter commitment as *the open-endedness of morality*.[15]

Although both *no necessary irrationality in immorality* and *the open-endedness of morality* are widely embraced among moral naturalists, they lead to two

13 Sterelny and Fraser, "Evolution and Moral Realism," 984. Cf. Boyd, "How to Be a Moral Realist," 329.

14 Jackson, *From Metaphysics to Ethics*.

15 Some may question my suggestion that *a priori* naturalists take morality to be open ended in this way. Jackson takes a conceptually competent, idealized reasoner with full information about the world to be in a position to know the moral truths ("From Metaphysics to Ethics," 31–42). ("Full information" must be qualified lest the reasoner's knowledge be trivial. We might restrict it to information about the worldly supervenience base described in some semantically neutral language.) Given this, it may be difficult to see how *a priori* naturalists such as him are faithfully characterized as viewing moral truths as open ended. Surely our moral concepts, together with the state of the world, *already* decide what the moral truths actually are? But concerns such as these merely require that we precisify the sense of open-endedness at issue. For Jackson, the moral truths depend upon our network of implicit or explicit moral beliefs and opinions ("folk morality") following critical reflection ("mature folk morality") *together with* the state of the world ("From Metaphysics to Ethics," 133). The conceptually competent idealized reasoner can determine which natural facts the moral facts *would be* at each world that is presented to her for consideration. But she cannot know which natural facts the moral facts *actually are* until she knows precisely which world she is in—until, that is, she knows (at least some, perhaps all) empirical facts as well. Moral truths are therefore open ended even for the *a priori* naturalist, in the sense that their content crucially hangs upon the nature of a particular class of nonmoral facts. Both *a priori* and *a posteriori* naturalists are thus committed to the kind of open-endedness that derives from leaving the fate of moral truths in the hands of some yet-to-be-determined, nonmorally specified state of the world. Crucially, as we will see in section 3, it is this sense of the open-endedness of morality that naturalists' opponents capitalize upon.

well-known challenges. The naturalist's commitment to *no necessary irrationality in immorality* invites the accusation that she fails to do justice to morality's normative credentials. Given her commitment to *the open-endedness of morality*, she has also been taken to task for overgenerating admissible moralities. These are, respectively, the intensional and extensional challenges for moral naturalism. I take each in turn.

2. THE INTENSIONAL CHALLENGE

The naturalist who occupies my attention is committed to *no necessary irrationality in immorality*. Her critics allege that this outlook fails to take morality seriously as a normative phenomenon. Taking morality seriously (so this line goes) requires taking it to be intrinsically reason giving. The naturalist must, however, view this as an exercise in vaulting ambition. She simply cannot take morality seriously if this is what taking morality seriously requires.

My aim in what follows will be to show that the intensional challenge loses much of its sting once we turn our attention to deep-seated features of human sociality and psychology. I begin by distinguishing different faces of the problem (section 2.1). Ultimately, I think the naturalist should concede to her opponents that there is a deep connection between moral requirements and reasons for action. She should, however, deny that it is deep in the way they think it is. More specifically, the naturalist should take "human beings have reason to be moral" to be true when construed as a *robust empirical generalization*.[16] Though not all agents take morality seriously, *we* certainly do—and it turns out to be surprisingly difficult for us not to. The task for section 2.2 will be to defend the empirical plausibility of this generalization. Defending it as a solution to the intensional challenge will be the business of section 2.3.

A caveat will be helpful before proceeding. Some philosophers may find themselves puzzled by *normativity*—not just by *moral normativity*. Some variation of the intensional (and indeed, extensional) challenge that I address here likely arises for naturalism about the normative as well as moral naturalism. And while certain lessons may carry over, I do not pretend to be offering a comprehensive response to the broader suite of challenges that have been brought to bear against naturalist approaches in metaethics. The paper is addressed, then, most directly to those philosophers who find themselves especially puzzled by normativity in the moral sphere—though I do hope

16 As I have noted, many naturalists accept (something like) this empirical claim. Yet they seldom if ever build a convincing empirical case for it.

those whose puzzlement extends wider still will be able to extract something useful from it as well.

2.1. Interpreting the Challenge

The intensional challenge to moral naturalism has many faces. In the interests of focusing the discussion, I want to isolate two strands of thought within this family and concentrate my critical attention there. The first takes the naturalist to task for identifying morality in something *external* to human agents.[17] Insofar as the naturalist regards moral properties as mind independent—something "out there" waiting to be discovered—her proposal gives rise to the potential for a gaping distance between the moral facts and the ends with which we identify. It remains an open question whether anyone *cares* about these facts. Yet it does not seem to be an open question whether we care about morality. Moral considerations have a deep practical hold upon us. The naturalist, then, appears to be looking for morality in all the wrong places. If we are to close the normative gap that she leaves wide open, then we would likely do better by understanding morality in terms of something *internal* to human agency—perhaps even in terms of human agency itself. This face of the intensional challenge reflects a concern about *normative distance*.

A second face draws upon deep-seated intuitions concerning morality's normative reach. Many find it highly intuitive that *all* agents have reason to be moral—not merely those who have particular desires or ends that would be served by being moral. This intuition seems to favor the proposal that moral reasons are *categorical* in character: that moral requirements provide an agent with normative reasons *independent* of whatever desires or preferences she happens to have.[18] More often than not, the point is driven home by calling upon a variety of morally dubious characters ranging from opportunists to outsiders. One well-known opportunist is Hume's sensible knave, who only acts morally well when doing so is to his benefit.[19] Outsiders are different: their psychological architecture differs in rather fundamental ways from our own. They and their ilk lie at the outskirts of the human community. (Extreme sadists are an exemplary case.) What opportunists and outsiders have in common is a desire set that is best served by acting contrary to moral requirements. Given this, the naturalist seems committed to saying that they lack reason to be moral. Yet that assessment fails to respect the intuition that these individuals *do* have reason to be moral. Opportunists and outsiders seem guilty of a normative mistake;

17 Korsgaard, *The Sources of Normativity,* 112.

18 Shafer-Landau, "A Defence of Categorical Reasons."

19 Hume, *Enquiry Concerning the Principles of Morals.*

they do not merely appear to ignore rules that it is optional to take seriously. Morality's normative reach thus seems more extensive than the naturalist can allow inasmuch as she constructs our reasons to be moral upon contingent foundations such as human preferences. This second face of the intensional challenge reflects a concern about *normative jurisdiction*.

In addressing the intensional challenge, I propose to draw inspiration from Hume. When reflecting upon morally dubious characters, Hume concedes that such persons may lack the sorts of concerns that yield reasons to be moral. Speaking of his sensible knave, he remarks: "if his heart rebel not against such pernicious maxims, if he feel no reluctance to the thoughts of villainy or baseness, he has indeed lost a considerable motive to virtue."[20] But to this, Hume adds an important qualification. Though the knave's heart may not rebel against pernicious principles, *ours* certainly do. Most people *are* emotionally constituted and socially situated such that they have reason to be moral.

Hume's strategy takes us some way in addressing concerns about morality's normative jurisdiction; if *most* human agents have reasons to be moral, then morality's jurisdiction is fairly respectable.[21] Though Hume's reasoning may seem simple and straightforward, I think it bears more fruit than one might expect. Indeed, a little empirical digging reveals it to be capable of mitigating concerns about normative distance as well. I do not doubt that there are a number of respectable interpretations of Hume's argumentative strategy. But I am going to propose that we develop his insights along the following lines: "human beings have reason to be moral" is true when construed as a *robust empirical generalization*. Let me explain what I mean by this using an analogy.

Consider the following generalization *E*: "elephants care for their young." This is true when interpreted as a statistical claim, for *most* elephants do care for their young. But it can also aspire to be more than a statistical claim, for there is a deeper explanation for this statistic; it is something about *the nature of elephants* that *explains why* they care for their young. Elephants mature slowly, relying on maternal (and allomaternal) milk for nutrition during their early years. Mothers also usually give birth to one calf at a time, a conservative reproductive strategy that favors high investment in individual offspring. Thus, *E* is also true when interpreted as a *characterizing generic*, a claim that—as I am understanding it here—tells us something about what is *normal* for members of a kind.[22] For our purposes, we can understand normality in terms of deeply

20 Hume, *Enquiry Concerning the Principles of Morals*, sec. 9, pt. 2.

21 It remains to be shown whether it is *respectable enough*. I take up this concern in section 3.3.

22 See Nickel, "Generics and the Ways of Normality." Cf. Cavedon and Glasbey, "Outline of an Information-Flow Model of Generics."

entrenched features of a kind's members. These are features on which many others depend, and which, if changed, would result in a wholesale change in the kind. Such features need not be strictly speaking intrinsic; elephants' high degree of sociality, for instance, is plausibly counted among them.

To be sure, *E* admits of exceptions. Elephant mothers in circuses and zoos, for example, reject or kill their calves surprisingly often—something that is partly accounted for by many of these mothers having been deprived of close contact with older females themselves when they were young.[23] But such exceptions are no threat to the truth of *E* as an empirical generalization. Insofar as they are exceptions, they seem to be principled ones: they are exceptions that prove the rule. If we wanted to transpose this idea into a framework of characterizing generics, then we might build upon Nickel's suggestion that "normality" always depends to some degree upon our "inductive target" as well as features of the kind in question.[24] If the rearing habits of elephants are our inductive target, then our interest presumably concerns what their typical caring practices look like. Yet caring practices—like much else in biology—reflect "a complex interaction" between a range of factors.[25] Some of these factors will be less relevant than others given our theoretical aim of arriving at useful and informative empirical generalizations about elephant rearing. Facts about elephants' maturation cycle and reproductive strategy, for instance, seem relevant given this aim, whereas influences introduced by human circuses and zoos do not.

My goal in what follows will be to argue that "human beings have reason to be moral" (henceforth, I will call this *H*) is on a par with *E* in these respects. *H* is plausibly true when interpreted as a statistical claim, a claim about what holds true for *most* of us. But it is *also* true when interpreted as a characterizing generic, a claim that reflects what is normal for us and holds true in virtue of deeply entrenched features of our psychology and sociality. (As we shall see, this is an instance where the evidence that favors the generic claim favors the statistical claim as well.) As I hope these qualifications make clear, the ensuing discussion is simply intended to illuminate phylogenetically ancient and important features of human beings that ground their reasons to be moral—no suspicious normative or teleological assumptions are being smuggled in. My arguments neither presuppose nor are intended to support any notions of natural human goodness and defect or standards of human excellence.[26]

23 Kurt and Mar, "Neonate Mortality in Captive Asian Elephants (*Elephas maximus*)."

24 Nickel, "Generics and the Ways of Normality," 643–45.

25 Nickel, "Generics and the Ways of Normality," 643.

26 Cf. Foot, *Natural Goodness*; Hurka, *Perfectionism*.

To my mind, much of the foregoing has to some extent been lost in previous responses to the intensional challenge. To be sure, it has not been lost on naturalists that humans typically have reason to be moral.[27] Yet it has not been sufficiently emphasized that this claim can aspire to be more than a mere statistic. Naturalists do, to their credit, sometimes appeal to facts about the human condition—to our sympathy or sociality, for example.[28] But that is only a start. It is no substitute for engagement with the rich empirical literature bearing on the matter. This literature reveals that our reasons to be moral have deep psychological roots.

Let me add one final clarification prior to proceeding. One may wonder whether, given *H*, humans generally have *decisive* reason to be moral (such that being moral is always rationally required) or merely *sufficient* reason to be moral (such that being moral is always rationally permissible). Speaking of "decisive" reasons here may seem a little ambitious. I would be content if my arguments at least supported the conclusion that humans often have decisive reason and very often sufficient reason to be moral. (Indeed, some may well find this less ambitious conclusion closer to the truth!) I would not be content if my arguments merely suggested that humans generally have some (perhaps vanishingly small) reason to be moral. Though I wager that the latter claim is true, it is not enough to address the intensional challenge.

Those skeptical of my strategy will fall into two camps. Some will suspect that *H* is false. Others will insist that, even if *H* is true, it does little to mitigate the intensional challenge. I will tackle each skeptic in turn.

2.2. Is the Empirical Generalization Plausible?

The naturalist who occupies my attention is, recall, committed to a particular conception of normative reasons: what grounds an agent's normative reasons is (roughly) what she cares about. Insofar as the naturalist takes normative reasons to be rooted in an agent's conative psychology, then, the task of supporting *H* is effectively the task of showing that moral edicts have strong resonance with us in a manner that reflects entrenched features of human psychology. I will focus upon three features in particular: our prosocial emotions, our sociality, and our need for good repute. As will become clear, there is a great deal of explanatory overlap here, for these features are mutually supporting.

Beginning with the prosocial emotions, humans clearly care about the welfare of others. Even human infants exhibit strong other-regarding concerns.[29]

27 See Railton, "Moral Realism," 170; Copp, *Morality, Normativity and Society*, 244.

28 Boyd, "How to Be a Moral Realist," 341; Copp, *Morality, Normativity and Society*.

29 Liszkowski et al., "12- and 18-Month-Olds Point to Provide Information for Others"; Warneken and Tomasello, "Altruistic Helping in Human Infants and Young Chimpanzees."

This is unsurprising. The survival of our species has long been predicated upon successful cooperation, and there has long been biological and cultural selection for emotional responses that support it. We feel sympathy in response to others' suffering, anger in response to their transgressions, and guilt in response to transgressions of our own.

Each of these experiences has attentional and motivational import. Prosocial emotions focus our attention on others—on their needs, their actions, and their situations. Empathetic emotions ground other-regarding concerns and motivate helping behavior.[30] Anger is a formidable motivator too: as its intensity increases, so too do the costs we are willing to incur to penalize mistreatment.[31] Guilt plays a central role in maintaining interpersonal relationships, urging us to repair social bonds that are threatened by our misdemeanors.[32] It should be emphasized that these experiences are not merely influential—they are typically quite powerful. Even *proactive* guilt can rein in a temptation to renege on social commitments.[33]

Some may complain that the behavior these emotions (dis)incentivize is merely *prosocial*—that it is not yet *moral*. But prosocial emotions need not fly solo. In human social worlds they are governed by shared standards and expectations. Norms *direct* our feeling; they tell us where to focus our sympathy, how much anger is warranted, and whether and when guilt is appropriate. There is a fundamental social need to coordinate our behavior, and a tried and trusted way of doing so is to direct our prosocial emotions toward similar action classes; to feel anger and guilt (in appropriate measure) in response to the same transgressions, and to reserve empathy for the same sorts of people. Moral norms, then, *direct* our prosocial responses in ways that build upon our emotional architecture as well as our capacities for rational reflection—not just any instance of guilt or empathy is sanctioned. Moral education is in part education in how to feel.[34]

With that said, it is worth observing that a healthy correlation between prosocial and moral (that is, morally sanctioned) behavior is precisely what many varieties of naturalism predict. Recall that for (many) naturalists, what

30 Eisenberg et al., "Relation of Sympathy and Personal Distress to Prosocial Behavior"; Findlay et al., "Links between Empathy, Social Behavior, and Social Understanding in Early Childhood"; Toi and Batson, "More Evidence That Empathy Is a Source of Altruistic Motivation."

31 Bosman and van Winden, "Emotional Hazard in a Power-to-Take Experiment"; de Quervain et al., "The Neural Basis of Altruistic Punishment."

32 Tangney and Dearing, *Shame and Guilt*, 124–25, 184–85.

33 See Frank, *Passions within Reason*.

34 Mameli, "Meat Made Us Moral."

distinguishes morality from other normative domains is precisely the *kinds of* natural phenomena that it makes its business: "society's needs," "human cooperation," or facts about "harm" and "benefit."[35] Again, this is not to suggest that being prosocial is *coextensive* with being moral. It is only to point out that a healthy degree of overlap here is to be expected. Reflection, deliberation, and negotiation have significant roles to play in the formation of moral norms as well.[36]

Importantly, the relationship between norms and emotion cuts both ways. Moral norms not only direct human feeling but have been molded by it. Every culture's moral package builds upon our affective architecture.[37] Norms backed by feeling gain more traction; emotion makes particular standards more salient and memorable.[38] Our emotional configuration also constrains the norms that we can get behind. Norms that align more closely with our affective predispositions tend to be more learnable. None of this is to deny cross-cultural variation. Different packages of norms build upon different features of human psychology, and they do so in different ways.

The prosocial emotions therefore enable coordination and promote cooperative response. Both features are important. The satisfaction of most of our human needs (such as subsistence and security) depends in some way upon our social group. Members of human societies have long been interdependent, and as a result their survival has long depended upon effective cooperation.[39] Our psychology reflects this history. We are adapted for interactive and collaborative living, cognitively as well as emotionally. Humans value joint activities intrinsically. Human children not only value cooperative games that lack an instrumental rationale but often transform tasks with an instrumental aim into cooperative interactions.[40]

35 See, respectively, Copp, *Morality, Normativity, and Society*, 159–60; Sterelny and Fraser, "Evolution and Moral Realism," 984; Foot, "Moral Arguments," 510.

36 See Kitcher, *The Ethical Project*. A related worry about my arguments is that they only establish reasons to conform to social conventions, not reasons to be moral. In fact, however, they establish both. The considerations discussed here explain why people often have strong incentives to conform to group norms. In addition, however—and as the considerations introduced in section 4 will help to drive home—they also explain why people often have strong incentives to defy them when they are judged or felt to be wrong, or when they conflict with powerful prosocial impulses that favor kindness or mercy. Any view that denied that both incentives are typically present would have a hard time accounting for social change born of moral resistance—and why such change often meets resistance in turn.

37 Haidt, *The Righteous Mind*.

38 Nichols, *Sentimental Rules*.

39 Tomasello, *A Natural History of Human Morality*.

40 Tomasello, *Why We Cooperate*, 63–65.

In a social world predicated upon cooperation, it helps to have someone to cooperate *with*. It is especially helpful to have a *good* cooperative partner—one who will not leave you for dead as soon as the cooperative labor has borne its fruit. This insight features prominently in evolutionary accounts of morality, where theories of partner choice often play a central role.[41] But the basic lesson carries over to the present. Nice guys do not necessarily finish last; *ceteris paribus*, they tend to do rather well on the cooperative market. We want to mix in the same circles as the reliable, the trustworthy. Even human infants prefer those with helpful dispositions.[42]

Reputation matters in a cooperative market. Having a bad name means having less social capital—no one wants to be paired up with a knave. And being short on collaborators carries real costs. The price of being unpopular is high, ranging from lower job prospects to lower life expectancy.[43] There are internal costs as well. We feel shame when others think ill of us. And shame is highly punitive.[44] It is bad—and it *feels bad*—to get a bad rap. Indeed, social disapproval is often regarded as an especially toxic form of punishment. Many report preferring pain, incarceration, amputation, or even *death* to a heavily tarnished reputation.[45] Humans care about how they fare in the court of public opinion. We take active steps to shape our reputations, and not only by acting in socially sanctioned ways but also "by joining in the conversation" about our actions and justifying them to others.[46]

The foregoing strongly suggests to me that human beings are generally emotionally and socially situated such that they have reason to be moral. Strong other-regarding concerns, a high degree of interdependence, and a need for others' good opinion are widespread and deeply entrenched features of our psychology and ways of life. Given this, characteristically moral behavior (helping others, say) tends to be to our benefit. Importantly, these insights would seem to apply to Hume's knave as well. Our social preferences favor authentically helpful dispositions—not opportunism. And social acceptance is not a luxury the average person can afford to forgo. The satisfaction of many fundamental

41 Tomasello, *A Natural History of Human Morality*; Stanford, "The Difference between Ice Cream and Nazis."

42 Kuhlmeier, Wynn, and Bloom, "Attribution of Dispositional States by 12-Month-Olds"; Hamlin, Wynn, and Bloom, "Social Evaluation by Preverbal Infants."

43 Western, Kling, and Weiman, "The Labor Market Consequences of Incarceration"; House, Landis, and Umberson, "Social Relationships and Health."

44 Tangney and Dearing, *Shame and Guilt,* 137–38.

45 Vonasch et al., "Death before Dishonor."

46 Sperber and Baumard, "Moral Reputation," 511.

human needs—alliances, romantic partnerships, careers—depends upon the accumulation of social capital.

Of course, the knave has a response at the ready here: no one need *know* that he is only virtuous for show. So long as he *appears* virtuous, he can avoid the costs that accompany a bad reputation. Yet this seems to reflect a naïve optimism on his part. The proposal that we are capable of systematically fooling others about our moral caliber lacks empirical plausibility. People are rather good at predicting others' cooperative intentions, especially those of acquaintances.[47] All in all, pretending to be good is a dangerous game.

Simply put, the best way to earn a good reputation is to deserve it—to actually *be* good.[48] At a minimum, being good requires internalizing standards: developing dispositions to feel anger when such standards are violated, and guilt when one falls short. We have seen that these prosocial emotions have motivational value. But they confer *signaling* value as well; emotional response is hard to fake, making it a fairly reliable sign of moral commitment.[49] These assurances are important, for we do not only choose collaborators with an eye to their track record—we try to make reasonable inferences about their mental states as well.[50] It is for this reason that the policy of behaving morally only when morality pays is not an especially promising policy. The kind of moral behavior that tends to pay is the kind that stems from sincere commitment.[51]

So much for the knave. What of the outsider, though? The outsider, recall, lacks typical human motives. She is (let us suppose) indifferent to others' opinions, impervious to guilt, and perfectly content to go it alone. Some will be inclined to regard outsiders as unassailable counterexamples to *H*. On reflection, however, I think it is more plausible to view them as principled exceptions

47 See Brosig, "Identifying Cooperative Behavior"; Frank, Gilovich, and Regan, "The Evolution of One-Shot Cooperation"; Pradel, Euler, and Fetchenhauer, "Spotting Altruistic Dictator Game Players and Mingling with Them."

48 Frank, *Passions within Reason*; Sterelny, *The Evolved Apprentice*, ch. 5; Sperber and Baumard, "Moral Reputation."

49 Frank, *Passions within Reason*.

50 See Sperber and Baumard, "Moral Reputation," 507. I should add that this is nothing approaching the whole story. Moral emotions likely differ in their signaling value; some may be less difficult to fake than others. See O'Connor, "The Evolution of Guilt." And I do not mean to claim that we make judgments about others *purely* on the basis of their emotional profile. More plausibly, we collate and draw upon different sources of evidence—reputation, behavior, emotional response—in arriving at a judgment.

51 At this stage, some readers will want to object that these observations only demonstrate (at best) that we have reasons to *be* good. To this, they will be quick to add that the reasons to be moral that the intensional challenge demands are reasons to *do* good. Rest assured, I take up this challenge in section 3.3.

to it. As should now be clear, a callous disregard for others is hardly characteristic of human beings. Indeed, this is among the key diagnostic criteria for a range of human *pathologies*.[52] Thus, psychopaths and other populations with systematic deficiencies in affective response pose no threat to our empirical generalization. These are the easy cases, the exceptions that prove the rule.

Or are they? Perhaps the easy cases are not *quite* so easy. One complicating factor is that disorders such as psychopathy lie within a spectrum, and that not all members of this population lie at its extreme end. Those members who do will invite the response above. But what of those who merely show psychopathic tendencies? The further away someone lies from the extreme, the *less* likely it becomes that the person will *lack* characteristic other-regarding concerns or the need for social support networks. But then, it also becomes *more* likely that the person will have reasons to be moral—in which case, we can simply apply the same reasoning we applied to the knave. The "principled exception" response, then, primarily concerns extreme outsiders; I do not deny that this phenomenon is graded in important ways.[53]

Let me now consider what I take to be the hard cases. There have long been drastic inequalities in wealth and power within human societies. Those who enjoy disproportionate shares of these resources (whom I will simply refer to as "elites") are not uncharacteristic human beings in the manner that outsiders are; they presumably share the same characteristic human concerns as the rest of us. But this would seem to spell trouble for *H*, for it is not clear that elites *do* have reasons to be moral. The features of our psychology and social lives that I have emphasized—other-regarding concerns, reputation, interdependence—seem far less pronounced in those occupying the upper echelons of society. Elites surely need not invest so much in their reputation; they seem less entrenched in our networks of interdependence and less beholden to others than the rest of us.

For my part, I think that elites *do* have reasons to be moral. It is true that Elon Musk does not have to cozy up to his boss to get a promotion. However, he is not practically isolated from the rest of society either. Stock prices in Musk's companies have been known to plummet as a result of his careless words (he once called an analyst a "boring bonehead"). Of course, such interdependence is plausibly more common in modern liberal societies. (Louis XIV certainly did not have to

52 American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*.

53 Yet another complication is that not all naturalists may agree with my suggestion that psychopaths in particular qualify as outsiders. Some, of course, will: my treatment above is importantly similar to Boyd's, for instance (though unlike him, I do not trace the psychological abnormality back to a "cognitive deficit.") See Boyd, "How to Be a Moral Realist," 340–42. But other naturalists seem to favor a contrary perspective. See, for instance, Brink, "Responsibility, Incompetence, and Psychopathy."

worry about his stock prices falling.) Bullying may therefore be a safer strategy in certain kinds of environments. Still, it is not without risk; radically discounting others' interests carries dangers of its own.[54] (I elaborate upon these in section 3.)

Though the above perspective strikes me as reasonable, let me offer a further possibility for those unconvinced. Elites may very well qualify as highly *uncharacteristic* human beings in a perfectly good sense. Here, we might recall Nickel's suggestion that "normality" is determined not merely by features of a kind but by our inductive target as well—something that, in turn, is informed by our interest in arriving at useful empirical generalizations.[55] If it is the factors underlying humans beings' reasons to be moral that are our inductive target, then our theoretical interest presumably concerns what human beings' typical ways of relating to one another look like. To this end, reputational effects, emotional dispositions, and social interdependence are relevant factors to consider. But the peculiar position of elites—who are socially situated in strikingly different ways to just about everyone else given their comparative lack of interdependence—arguably are not.[56]

My treatment of the hard cases thus has two prongs. I myself believe there is a reasonable case to be made that elites *do* have reason to be moral. That said, there is also an alternative outlook according to which such persons simply fall outside the scope of our inductive target and lack reason to be moral. Either way, there is no trouble for *H*.

2.3. Is It Enough?

I have argued that human beings have reason to be moral, where that is understood as a robust empirical generalization. The task of the previous section was to motivate the empirical plausibility of that generalization. The task of this section will be to motivate its metaethical serviceability. I now address both faces of the intensional challenge.

Let me begin with the normative distance worry, which takes the naturalist to task for locating morality in something *external* to human agents. Insofar as

54 Elites such as Musk arguably have further interests that are served by being moral. A referee helpfully observes that this seems especially true when we turn our attention to moral norms of a broadly Kantian kind and focus on personal relationships. If Musk does not keep his promises or treat people with respect, for instance, then he presumably will not have many (true) friends.

55 Nickel, "Generics and the Ways of Normality," 642.

56 Of course, I have maintained throughout that *H* is true *both* as a characterizing generic *and* as a statistical claim about most human beings. But the latter interpretation is easily accommodated in the case of elites. It is immensely difficult to reduce practical dependence upon others without certain resources (including vast sums of money and human capital). Clearly, not everyone has such resources, for not everyone can be in the top 1 percent.

moral facts are mind independent, there is thought to be a real potential for a gaping distance between them and the ends with which we identify. Upon reflection, however, this is not quite right. In light of the considerations raised in section 2.2, it is more perspicacious to view the naturalist as providing an answer that is at once internal *and* external. Moral facts are not simply facts about our wills or attitudes; they are objective facts about the practices that effectively support our cooperative endeavors. In this sense, they are external. But *which* practices effectively support our cooperative endeavors depends heavily upon deep-rooted features of human psychology. As we have seen, not all norms are created equal; those that have greater motivational uptake are more likely to be preserved and passed on. The prosocial emotions thus establish a harmony between motivation and moral response. Moral facts, then, have a crucial *internal* element, for they depend in crucial ways upon our emotional constitutions.

I turn now to the normative jurisdiction worry: On my account, is morality's normative jurisdiction extensive enough? I have conceded that outsiders lack reason to be moral, and have argued that this need not undermine *H*. Still, some may worry about outsiders escaping our *criticism*. Insofar as we concede that outsiders lack reason to be moral, we seem to have rendered inadmissible any normative complaint we might have had against them. We can no longer charge such individuals with having failed to acknowledge the reasons within their normative landscape. If they care not for moral matters, and we acknowledge that there is no reason for them to do so, then what is left for us to say to them?

I am inclined to view this question as premised upon a faulty assumption—namely, that we must have something *to say to* the outsider. We will certainly want something *to do about* them. And we will certainly have a lot *to say about* them. Yet it is difficult to see why they are properly viewed as a target of moral conversation. One can understand why opportunists meet this condition. (Here, the call is coming from inside the house!) But the outsider is incapable of authentic participation in moral life. Perhaps she is someone to be controlled or contained—but she is surely not someone to be *convinced*. For a genuine conversation to proceed, there must be common ground—something we clearly lack with the outsider.[57]

It is worth saying something more in defense of this position, especially since some may take the treatment of outsiders to crosscut the intensional and extensional challenges. (If we deny that outsiders have reason to be moral, then we may risk not getting the content of morality quite right either.) In response, it bears mentioning, following Sharon Street, that the naturalist can

57 Cf. Manne, "Internalism about Reasons," 96–97, 103; Woods, "Footing the Cost (of Normative Subjectivism)."

easily accommodate our *morally unfavorable opinion* of the outsider.[58] We are indeed saying something true when we describe the outsider as a cad, an evildoer, a villain, or a malefactor (or various other things that should probably be omitted from an academic philosophy paper). Of course, such accusations are unlikely to strike fear into the heart of the outsider. But they certainly hit a nerve *with us*. Given the enormous importance that we attach to our terms of interaction with one another, being branded a moral pariah is just about the deepest insult we have at our disposal.

The opponent of naturalism's worry, then, cannot be that the naturalist is incapable of charging the outsider with an important normative failing. The worry is that the naturalist cannot charge the outsider with *a particular kind* of normative failing: namely, a failure to recognize and respond to reasons that she does indeed have. But the foregoing considerations should, I think, leave us feeling less confident that the latter charge is truly needed. Failing to recognize or respond to your own reasons is indeed a kind of shortcoming, one that signals some sort of normative defect, such as irrationality. But it is often regarded an even *greater* shortcoming to fail to live up to moral standards. Being a fool might be bad, but being a jerk is arguably worse.

Still, it would be helpful if the naturalist were capable of explaining away the intuition that outsiders *are* guilty of the normative failing that her opponent has in mind: Why is there the temptation to think that outsiders *do* have reason to be moral—when in fact they do not? My own suspicion is that this is likely owing to a common but understandable error in our thinking about them. Ordinarily, when we want to interpret others' behavior, we proceed on the assumption that they are fundamentally *like us*.[59] Outsiders, however, are *not* fundamentally like us. It is, then, rather misguided to import our own psychology into our efforts to interpret them—indeed, outsiders may well be unintelligible from our perspective.[60] But it is also understandable that we make such an error. Outsiders do not, after all, tend to announce their presence. And insofar as *H* holds true as a statistical claim as well, the default assumption that those outsiders we encounter are likely to have reasons to be moral seems justified, even if ultimately false.[61]

It follows from what I have said that morality's normative jurisdiction is not *limitless*—some human agents lie beyond its reach. But it does not necessarily

58 Street, "In Defense of Future Tuesday Indifference," 293.

59 See Davidson, "Radical Interpretation."

60 Notice that unlike outsiders, opportunists do seem intelligible, for they *are* fundamentally like us. And it is precisely *because* they are sufficiently like us that they typically *do* have reason to be moral.

61 Cf. Street, "In Defense of Future Tuesday Indifference," 293–94.

follow that I have failed to address the intensional challenge. Given the above arguments, it strikes me that we should now be far less confident that the folk view of morality finely distinguishes—let alone unambiguously decides—between the possibilities that reasons to be moral are:

1. Somewhat common among human agents, given their contingent preferences
2. Widespread among human agents, given their contingent preferences
3. Widespread among human agents in contingent albeit robust ways
4. Possessed by all human agents
5. Necessarily possessed by all human agents

Given my arguments, the naturalist can establish that morality's jurisdiction is respectable (possibility 3), even if it is not quite as expansive as some of us believed it to be (as supporters of possibilities 4 or 5 would have it).

But have I really *addressed* the opponents' concerns, or have I simply *dismissed* them? That opponent, recall, is not only concerned with capturing *a sizable contingent* of agents in our moral-reasons net; she is also concerned with *the manner* in which we do so. Following a well-known naturalist tradition, we have responded to the intensional challenge by appealing to widespread but *contingent* facts about human preferences. But what truly lies at the heart of the intensional challenge, it seems, is the concern that naturalism fails to capture the way in which moral reasons are *special*. And according to those who raise the challenge, this specialness is best spelled out in terms of such reasons being inescapable or categorical—that is to say, in terms of their *not* being premised upon contingent human preferences. Recall that it is precisely in virtue of making this move that the naturalist's opponent seems to do better justice to intuitions concerning morality's normative reach; they are able to say that its jurisdiction reflects something more in the order of possibility 5 in the list above.

In response, I think that the naturalist can readily concede that moral reasons *are* special. What she will deny is that they are special in the particular way her opponents take them to be. For the naturalist, what makes moral reasons special is in essence the special role they occupy in our hearts; as an empirical matter, human beings attach immense importance to being good and doing good. It is for precisely this reason, moreover, that we understandably (though, if I am right, mistakenly) take outsiders to have reasons to be moral: in our efforts to interpret and relate to them, we assume that they are fundamentally like us in this way.

Nevertheless, the opponent will understandably press: our reasons to be moral are still left hostage to ordinary human preferences on this approach. Are they not then rather *un-special*? Well, yes—*if* we have already decided in advance that being special requires being categorical. But one point that

needs emphasizing in this context is that we are not forced to choose between possibility 1 and possibility 5; something in between may well be capable of explaining the special importance we attach to the moral dimensions of our lives *and* vindicating our expectation that many (even if not all) others will tend to attach special weight to these dimensions as well. The naturalist, then, need not rest content with the claim that our reasons to be moral stand or fall with some *flimsy* alliance of human preferences. She can and should go further than this, emphasizing that many of the social and emotional factors that ground our reasons to be moral reflect robust features of our social environments and psychology. To put the point in a slogan: it need not follow from the fact that our reasons to be moral are contingent that they are *precarious*; some things are both contingent *and robust*.[62]

Let me conclude my discussion of the intensional challenge by considering a further feature of my arguments that some may take issue with. It may be objected that I have not shown that it is rational *to act* as morality recommends. At best, I have shown that it is rational to be (or become) *a person* who is intrinsically motivated to act as morality recommends. But surely it is the first of these conclusions that is needed to address the intensional challenge—not the second.

There is of course a respectable strategy for responding to this concern: forge a connection between the two conclusions. David Gauthier is well known for proposing that insofar as it is rational to be a virtuous person, it is also rational to perform acts that are the output of a virtuous disposition—even when doing so is not to one's immediate benefit.[63] While I think this response is on the right track, I do not wish to borrow from it too uncritically. Gauthier's project is driven by an ambition to account for morality's normative reach in a way that makes little if any appeal to other-regarding concerns. I do not share this ambition. Gauthier's strategy also comes dangerously close to simply redefining "rational action." Far from establishing a nexus between doing good and being good, one may worry whether the strategy is not too quick to simply assume or stipulate that there is one. Let me, then, do a little more to motivate the idea that there is such a nexus. I will proceed on the assumption that my arguments have established reasons to *be* good. The present question is how that could establish reasons to *do* good as well.

Notice first that becoming a good person usually involves reordering one's priorities. As a morally mediocre individual, I might value my career above

62 Though she does not quite put the point in these terms, I take it that something like this is what Foot was getting at in her discussion of the volunteers in the siege of Leningrad. See Foot, "Morality as a System of Hypothetical Imperatives," 310–11.

63 Gauthier, *Morals by Agreement,* 170–77.

all else, with my wardrobe coming in at a close second, and my family a distant third. After having undergone a process of moral character development, however, my preference ordering would likely be different. (A good person presumably puts their family before their wardrobe.) Part and parcel of being a morally good person, then, is having particular priorities—priorities that plausibly favor acting as morality requires on any (or many a) choice occasion. Although the morally mediocre may have preference orderings that justify neglecting their family for their careers, morally good people usually will not.

Further, it is not implausible that doing good may be important for *remaining* good. Moral action may be habit forming. Good behavior cultivates good character, as moral motives are reinforced by positive feedback from the social environment. Conversely, acting immorally often involves setting aside human feeling (pangs of guilt, say), as well as characteristic human concerns ("What will others think?"). The more we set aside such concerns, the more adept we are likely to become at overcoming the prosocial impulses that promote moral response. Overcoming these impulses is difficult, but it is by no means impossible. Thus, patterns of bad behavior likewise seem habit forming; over time, we risk breeding insensitivity. Doing bad may *make* us bad. At the very least, it seems apt to make us *less good*.

3. THE EXTENSIONAL CHALLENGE

Recall that a metaethical position accommodates morality's *extensional* character just in case it (largely) accords with substantive judgments regarding the extension of moral terms such as "morally impermissible." To see why the naturalist has trouble meeting this requirement, we can begin by revisiting her recipe for discovering which natural properties are the moral ones: she identifies some *nonmoral* purpose that morality serves and proposes that the moral facts are those that fit the bill. Consider, for example, Kim Sterelny and Ben Fraser's contention that

> a natural notion of moral truth falls out of the picture that moral belief evolved (in part) to recognize, respond to, promote, and expand the practices that make stable cooperation possible. For there are objective facts about the conditions and patterns of interaction that make cooperation profitable, and about those that erode those profits.[64]

On this approach, we are empirically corrigible when it comes to what the moral facts are—the naturalist lets the world do much of the talking. There is,

64 Sterelny and Fraser, "Evolution and Moral Realism," 985.

however, a problem with letting the world do the talking: we might not like what it has to say. The naturalist proposes to single out moral properties by the nonmoral purposes they serve. Yet when the empirical dust has settled it may well turn out that these purposes are served in unsettling ways. There may be moral systems that fulfil morality's function (to promote cooperation, say) but sanction a range of behaviors that seem obviously morally impermissible. Insofar as the naturalist is committed to viewing these systems as comprising moral *truths*, she cannot plausibly accommodate morality's extensional character. It is this challenge that will occupy my attention in the remainder of this paper.

Two clarifications will be useful before proceeding. First, I am going to restrict myself in what follows to the variety of moral naturalism defended by Sterelny and Fraser.[65] This restriction is purely for illustrative purposes; the basic strategy could be enlisted by other naturalists as well. With that said, Sterelny and Fraser's framework seems especially likely to raise the following concern: Am I in the business of defending *realism* or *relativism* here? What if moral system$_1$ turns out to best promote cooperation for society$_1$, whereas moral system$_2$ best promotes cooperation for society$_2$? Many naturalists are, as it turns out, prepared for this eventuality. Frank Jackson argues that there is reason to expect convergence on a particular human morality but admits that this cannot be known in advance; if divergence truly is in our stars, then we should be willing to retreat into relativism.[66] Likewise, Richard Boyd thinks it is "pessimistic" to expect more than one human morality to emerge but concedes the possibility. (He nevertheless maintains that were it to eventuate, that would only "refute moral realism as that doctrine is ordinarily construed" and "would not undermine a generally realistic conception of moral language.")[67] But is this really all the relativism-realism divide boils down to: a mere empirical conjecture or a potentially misplaced hope? Perhaps so, perhaps not. I certainly do not want to pick *that* meta-metaethical battle here. The point is simply this: if the illustrative example raises concerns about realism retreating into relativism, it is arguably not unrepresentative of naturalism in this respect.

Moving on to our second clarification, it seems optimistic to expect that any specific naturalist identification of moral properties with natural properties is correct *as it currently stands*. Given this, we should not expect any present variety of naturalism to escape the extensional challenge *completely* unscathed. My ambition, then, is not to show that one promising implementation of naturalism is *completely immune* to the extensional challenge. It is rather to demonstrate that

65 Sterelny and Fraser, "Evolution and Moral Realism."

66 Jackson, *From Metaphysics to Ethics*, 137.

67 Boyd, "How to Be a Moral Realist," 351–52.

it is far more *resistant* to the challenge than is commonly thought. This will, I hope, make us less inclined to think the naturalist project is doomed to fail and more confident in its future. Although we are yet to develop a foolproof variety of moral naturalism, we are perhaps not quite so far off as some may fear.

Now to the demonstration. Sterelny and Fraser propose to understand moral truths as "maxims that are members of near-optimal normative packages—sets of norms that if adopted, would help generate high levels of appropriately distributed, and hence stable, cooperation profits."[68] Their proposal is premised upon a particular empirically well-motivated picture of the evolution of human cooperation. To summarize, Sterelny and Fraser maintain that cooperative arrangements are more likely to be stable when the distribution of cooperative profits is *fair*—roughly, when there is not a huge disparity between any individual's investment and her returns. When everyone is guaranteed roughly proportionate returns, everyone has a stake in the venture being successful. On this account, our moral psychology evolved to support effective cooperative arrangements such as these. We are adapted to "recognize, respond to, promote, and expand the practices that make stable cooperation possible."[69]

A notable worry with this picture is that there seems to be nothing to prevent normative packages of the kind that interest Sterelny and Fraser from being morally *perverse*. Sterelny and Fraser may be forced to embrace the uncomfortable conclusion that what turn out to be the moral truths—for them, maxims that are members of a near-optimal normative package—conflict in striking ways with a swathe of substantive moral judgments. The worry is not baseless, especially given the details of Sterelny and Fraser's proposal. Moral norms have a bad track record. The catalog of morally prescribed behaviors in human societies is dreadful, ranging from honor killings to foot-binding, female genital mutilation, and slavery.

At this juncture, a naturalist seems to find herself in a double bind. She cannot simply *define* the problem away. Building substantive moral premises into the conditions for effective cooperation amounts to abandoning her purely descriptive recipe for identifying the moral among the natural. Nor, it seems, can she hope to dismiss aberrant moralities on purely empirical grounds. If maxims that permit enslavement serve a society's cooperative purposes, then the naturalist seems committed to viewing them as moral truths. Yet that seems wrong. As Max Barkhausen observes, "Most of us are deeply opposed to the idea that any way of coordinating on mutually beneficial behavior that our

68 Sterelny and Fraser, "Evolution and Moral Realism," 985.

69 Sterelny and Fraser, "Evolution and Moral Realism," 985.

moral evolution might have led us to endorse is as good as any other."[70] As it turns out, however, the naturalist is *not* committed to the idea that any way of coordinating is as good as any other. On closer inspection, perverse moral norms are not particularly effective in promoting stable, efficient cooperation.

The thought that the naturalist is forced to accept whatever evolution throws her way seems to be premised upon an unreasonable optimality assumption. The assumption seems to be that a norm's very existence as a facilitator of cooperation entails that it is part of a near-optimal normative package. This assumption is empirically suspect; in practice, many factors make moral optima difficult to reach. In what follows, I draw attention to some of the mechanisms that lead to the establishment and entrenchment of suboptimal normative packages. I will then explain why these packages are properly viewed as suboptimal—why they plausibly fail to promote stable cooperation. The mechanisms that I explore overlap to some degree. But each raises considerations distinct enough to deserve mention.

One such consideration is raised by Sterelny and Fraser themselves.[71] Norms are not only tools of cooperation—they are tools of *coordination*. Norms establish shared expectations for behavior. These expectations are internalized, and deviations are heavily punished. This feature incentivizes conformity, but it also carries a danger, for punishment can stabilize destructive behaviors as well as cooperative ones.[72] Even when the status quo runs *against* profitable forms of cooperation, then, agents can still have strong incentives for compliance.

Norms are also tools of identification. Cultures differentiate themselves through "ethnic markers" such as patterns of speech, dress, and dietary preferences. Patterns of normative response are differentia as well: members of a culture dress, dine, *and* moralize like one another. Importantly, moralizing is not just a matter of paying lip service to social mores. Talk is cheap. (One need not be committed to the cause to denounce the freeloader who spent the afternoon slacking off.) Thus, groups often demand costly signals of commitment to their way of life. Though these costly displays can promote group cohesion, they do not always promote stable and profitable forms of cooperation. Many signals of religious commitment, for instance, impose nontrivial opportunity costs.[73]

These considerations caution against taking a norm's existence as a facilitator of cooperation as a reliable sign that it forms part of a near-optimal

70 Barkhausen, "Reductionist Moral Realism and the Contingency of Moral Evolution," 677.

71 Sterelny and Fraser, "Evolution and Moral Realism," 1001.

72 Boyd and Richerson, "Punishment Allows the Evolution of Cooperation (or Anything Else) in Sizable Groups"; Abbink et al., "Peer Punishment Promotes Enforcement of Bad Social Norms."

73 See Bulbulia, "Religious Costs as Adaptations That Signal Altruistic Intention."

normative package. Moral norms have a lot of work to do. They must support stable cooperative ventures. But they must coordinate and delineate as well, all while remaining sensitive to changes in the social-environmental landscape. Yet all of this merely explains why nonoptimal packages *arise*; it does not explain their stubbornness. Just why do suboptimal moral packages *persist*?

One part of the explanation for the entrenchment of nonoptimal packages is that norms are not modular.[74] Norms form part of interconnected webs of cultural standards and expectations; it can be hard to modify one element of that web without making drastic changes to the rest. Human groups also tend to be normatively homogenous.[75] This leaves less room for suboptimal norms to be selected against in favor of the superior ones on offer (there simply *are not* any superior ones on offer).

Moreover, escaping from suboptimal packages often requires solving difficult collective action problems. A useful illustrative example is the practice of female foot binding.[76] Initially, foot binding functioned as a high-cost signal of status: only the wealthy could afford to immobilize potential workers. But it lost this signatory value when it became universal practice. At this stage, *everyone* was worse off than they were without the practice. Yet unilateral defection was no longer an option in a world where unbound feet meant poor marital prospects. A similar lesson applies to inegalitarian social arrangements. Just about everyone is worse off when the distribution of the cooperative surplus is radically unequal. But unilateral revolt is not a viable strategy; a successful revolution requires a critical mass of dissenters.

Finally, humans engineer their social worlds. If you are at the top, then you will presumably want to stay there. A legitimizing ideology is a wise investment—power tends to have a longer shelf life once you have convinced others that you have the celestial tick of approval.[77] The elite need not be swindlers, to be sure. Over time, they may well come to believe their own propaganda. Moral conviction is sadly not immune to the influence of self-interest. It is not in the least bit surprising that the institution of slavery was favored by those who stood to gain economically, nor that the elite often have a penchant for social stratification.[78]

74 Sterelny, "SNAFUS," 325; Buchanan and Powell, "De-moralization as Emancipation," 121.

75 Sterelny, "SNAFUS," 325.

76 See Sterelny, "SNAFUS."

77 Sterelny, *The Evolved Apprentice*, 111.

78 See Enoch, *Taking Morality Seriously,* 192–93; Buchanan and Powell, "De-moralization as Emancipation," 122–23. As I hope the discussion here makes plain, it is possible for a social arrangement to be "entrenched" without being "stable" (though I admit this sounds strange). To take a more familiar example: a business might be excessively micromanaging, bureaucratic, and procrustean in its handling of employees—and these arrangements may

So far, I have argued that we should be wary of inferring from the presence and perseverance of a norm as a facilitator of cooperation that it forms part of a near-optimal package. But one may still demand a positive rationale for thinking that dreadful norms *do not* form part of such packages. On the face of it, honor killings and slavery support cooperation rather well. It certainly is not *obvious* that societies that endorse these behaviors are less stable than those that do not. A fully responsible treatment of this issue would require consideration of normative packages on a case-by-case basis—a Herculean task that I cannot hope to undertake here. Let me, however, provide some principled grounds for thinking that norms that heavily discount the interests of large subsections of the population are unlikely to be near-optimal.

The first thing to note is that unfairness breeds resentment, and that resentment breeds instability in turn. Following Phillip Kitcher, the "technological possibilities for violent retaliation now increasingly available to the poor" mean that radically inegalitarian societies often have a strong potential for collapse.[79] Though certain forces can and do entrench unfair arrangements, it does not follow that such arrangements are robust to any changes. Indeed, many factors threaten to bring these systems crumbling down.

For one thing, oppressive norms can often be difficult for large subsections of the population to internalize. Elliot Turiel documents astounding resistance to sexist mores. Among these is the example of women in Saudi Arabia, who protested laws refusing them the right to drive by driving a convoy of cars through the city of Riyadh.[80] Kristen D. Neff also found that lower-middle-class Hindu women in India are highly critical of their lack of independence.[81] Gerry Mackie reports that many women in cultures that practice female circumcision strongly disapprove of it.[82]

Norms that are not fully internalized have a strong potential for disintegration. When agents do not value compliance for its own sake, they must be provided with strong incentives to play along—usually, heavy penalties for

---

very well come to be entrenched as the result of market forces. But insofar as such policies breed contempt, they are unlikely to be stable; employees will, for instance, likely engage in efforts to undermine the system through strikes or other forms of protest. Even if these bureaucratic arrangements come to be entrenched, then, they seem unlikely to be stable in the long run. At the very least, they seem far *less* stable (and thus, further from optimal) than salient alternative systems.

79 Kitcher, *The Ethical Project,* 311. See also Railton, "Moral Realism," 191–94.

80 Turiel, *The Culture of Morality*, ch. 9.

81 Neff, "Judgments of Personal Autonomy and Interpersonal Responsibility in the Context of Indian Spousal Relationships."

82 Mackie, "Female Genital Cutting," 143.

noncompliance. Yet this introduces a danger: as soon as the penalties break down, so too does the compliance. Antifornication norms are a nice illustration.[83] Sexual impulses are strong motivators. Reining them in requires penalizing promiscuity, through stigmas attached to illegitimate children, for example. As soon as contraception and urbanization appeared on the scene, these disincentives were not nearly as powerful. (Contraception removes the threat of illegitimate children. Urbanization affords greater privacy.) Moreover, such mechanisms of policing *are themselves* a significant social cost. Norms that must be enforced on a sizable portion of the population are *ipso facto* expensive norms to have.

In summary, then, near-optimal packages require stability (among other things), and stability is difficult to achieve when the interests of large subsections of the population are heavily discounted. This basic insight goes a considerable way toward helping the variety of naturalism under investigation avoid countenancing perverse normative packages. Insofar as the norms that make up these packages are not plausibly viewed as near-optimal, the naturalist need not say that they reflect the moral truths. This is not to deny that naturalists still have their work cut out for them. Any particular recipe for identifying the moral among the natural will still need to contend with the full suite of available empirical data. A more thorough defense of the idea that perverse norms are (relatively) ineffective at promoting stable, efficient, cooperation would require getting into the weeds to a greater extent than I have here.

On that note—and in the spirit of inspiring further optimism—let me edge closer to a conclusion with a more concrete case study. According to Christopher Boehm's well-known work on hunter-gatherer groups, capital punishment is not an uncommon response to reciprocity violations.[84] One may want to claim that these excessive punishment norms are morally perverse. And yet, they seem to have proven effective in stabilizing hunter-gather societies and their broadly egalitarian social arrangements for quite some time. Is the naturalist not then forced to view these punishment norms as reflecting moral truths?

Everything here will, of course, depend upon the details. To begin with, we should not overstate the extent of capital punishment in response to reciprocity violations; fewer than half of the groups Boehm studied (24 of 50) reported it, and ostracism and shaming are far more frequent reactions than moralistic killing.[85] It is also important to consider *which sorts of* reciprocity violations are typically met with capital punishment. Among the most common are those

83 Buchanan and Powell, "De-moralization as Emancipation," 113–14.

84 Boehm, *Moral Origins.*

85 Boehm, *Moral Origins* and "The Moral Consequences of Social Selection," 172, 174.

that involve an individual intimidating other group members—for instance, through "psychotic aggression" or "repeated murder." When the bully is only perceived as a moderate threat to group functioning, nonlethal measures are often used instead.[86] These details may lead us to question whether hunter-gatherer punishment norms are *obviously* perverse (in the manner that, say, genocide or slavery are—the permissibility of capital punishment for murder is, after all, still a matter of live debate). They might also lead us to question *which* capital punishment norms really do the work of stabilizing cooperation. Given its relative frequency, it may well be that capital punishment in response to repeated murder bears the bulk of the explanatory burden here—as opposed to say, capital punishment in response to theft or taboo violations. And this latter point, of course, feeds into a more general question: namely, whether these normative packages are reasonably viewed as *near-optimal*. As Bohem notes, capital punishment is incredibly costly in the hunter-gatherer context insofar as it cuts off reproductive opportunities and limits social and familial support networks.[87] Even if excessive punishment norms stabilize cooperation to some degree, then, it is not unlikely that cooperation could be rendered more stable and effective still under alternative, less excessive arrangements.

In general, empirical questions such as these will clearly remain important to any recipe for identifying the moral among the natural of the sort that has occupied my attention here—that is, to the sort of recipe that singles out moral properties by appealing to certain *nonmoral* facts, such as the norms that stabilize efficient cooperation, or how language users would apply moral terms following negotiation and reflection.[88] What I have sought to show is that this sort of recipe turns out to be far more promising than it initially appears; it is far from being a foregone conclusion that all implementations of it will send us plummeting headfirst into the realm of perverse moral norms. *Some* implementations might, of course. But this just seems like a reason for thinking that some naturalists have gotten their particular recipe wrong rather than an indication of a problem with having such a recipe.[89]

86 Boehm, *Moral Origins* and "The Moral Consequences of Social Selection," 173.

87 Boehm, "The Moral Consequences of Social Selection," 174.

88 Sterelny and Fraser, "Evolution and Moral Realism"; Jackson, *From Metaphysics to Ethics.*

89 For a point of comparison: a particular functionalist analysis of mental states may end up counting the wrong kinds of things as desires. But this would not necessarily be a problem with functionalism or its naturalist ambitions; it would instead be a problem with that particular way of using functionalism (or, more perspicaciously, with the particular background theory being put to use). As with just about any method or schema, what we get out depends upon what we put in.

Naturalists who do not share my optimism may prefer a different sort of recipe. The most salient alternative would be one that singles out moral properties by appealing to certain *moral facts*. One might tie moral truths to the judgments of a morally reasonable person, for example, or to judgments that are interpersonally justifiable.[90] One benefit of this approach is that it seems well suited to fending off perverse norms; perhaps we simply cannot hope to get the right moral results *out of* our naturalistic recipe without putting the right moral ingredients *in*. However, I am inclined to agree with Bart Streumer that this alternative approach faces insurmountable difficulties (most notably, a problem of vicious regress).[91] But naturalists who disagree will, I hope, still be able to extract an important lesson from this paper: we may be able to go much further with a purely descriptive recipe than has previously been thought.

4. CONCLUSION

My organizing focus in this paper has been the naturalist's prospects for accommodating the intensional and extensional character of morality. My organizing ambition has been to build a case for an optimistic prognosis. I do not pretend that these are the *only* challenges that moral naturalism faces. One not-too-distant cousin of the extensional challenge, for instance, appeals to the unsettling *arbitrariness* that the naturalist seems content to tolerate. In the naturalist's way of seeing things, the only thing to recommend our own package of norms over other possible contenders seems to be that—owing to idiosyncratic features of our history and psychology—such norms promote profitable, stable forms of cooperation among us. Given this outlook, it is difficult to see what *justifies* our norms over alternatives that achieve the same ends for other possible versions of ourselves.

I cannot hope to offer a response to this additional challenge here. But my arguments do suggest a natural line of reply. Should we ever arrive at a near-optimal normative package, we will have arrived at a way of getting along that is well suited to the creatures that we are. Contrary to what initial appearances may suggest, the fact that this normative package will be well suited to us seems *far from* arbitrary, for its suitability will be explained by deep-seated and relatively inflexible features of our social existence and psychology. "It works for us" might

90 See Brink "Realism, Naturalism, and Moral Semantics," 175–76.

91 See Streumer, *Unbelievable Errors,* 55–57. Streumer is also skeptical about the prospects of the recipe that I favor, which he takes to fall prey to what he calls "the false guarantee objection." See Streumer, *Unbelievable Errors,* 47–55. This objection is similar to what I have called the extensional challenge—hence my disagreement with Streumer's grim assessment of it.

sound shallow. But it sounds far less shallow once we remind ourselves just what is required for a moral package to work for us: it must resonate with us, coordinate us, and promote profitable cooperative enterprise among us. As I have been concerned to emphasize, not just any mode of moral interaction fits this bill. Any that does will have to build upon the very features that make us human.[92]

University of Leeds
j.m.isserow@leeds.ac.uk

REFERENCES

Abbink, Klaus, Lata Gangadharan, Toby Handfield, and John Thrasher. "Peer Punishment Promotes Enforcement of Bad Social Norms." *Nature Communications* 8, no. 1 (September 2017): 1–8.

American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. 5th ed. Washington, DC: American Psychiatric Association, 2013.

Barkhausen, Max. "Reductionist Moral Realism and the Contingency of Moral Evolution." *Ethics* 126, no. 3 (April 2016): 662–89.

Boehm, Christopher. *Moral Origins: The Evolution of Altruism, Virtue, and Shame*. New York: Basic Books, 2012.

———. "The Moral Consequences of Social Selection." *Behaviour* 151, no. 2–3, (November 2014): 167–83.

Bosman, Ronald, and Frans Van Winden. "Emotional Hazard in a Power-to-Take Experiment." *The Economic Journal* 112, no. 476 (January 2002): 147–69.

Boyd, Richard N. "How to Be a Moral Realist." *Contemporary Materialism: A Reader*, edited by Paul K. Moser and J. D. Trout, 307–70. New York: Routledge, 2005.

Boyd, Robert, and Peter J. Richerson. "Punishment Allows the Evolution of Cooperation (or Anything Else) in Sizable Groups." *Ethology and*

92 I am very grateful for the feedback that I received on this paper at several workshops and conferences over the years. I thank audiences and prereaders at the Origins of Social Inequality, Hierarchy, and Social Complexity Workshop at the Australian National University; the Philosophy of Biology Conference at Dolphin Beach; the Meta-Ethics Workshop at the Frankfurt Business School; and MetaEssen VI at the University of Duisburg-Essen. For comments on earlier drafts, special thanks are owed to Ellen Clarke, Edward Elliott, Guy Fletcher, Ben Fraser, Josef Holden, R. A. Rowland, Kim Sterelny, and Pekka Väyrynen. For comments on the latest version, I thank the two anonymous referees for this journal who took the time to provide constructive feedback. This work was supported by an Australian Research Council Grant for the project "The Origins of Inequality, Hierarchy, and Social Complexity" [grant number FL130100141].

Sociobiology 13, no. 3 (May 1992): 171–95.
Brink, David O. *Moral Realism and the Foundation of Ethics*. Cambridge: Cambridge University Press, 1989.
———. "Realism, Naturalism, and Moral Semantics." *Social Philosophy and Policy* 18 (Summer 2001): 154–76.
———. *Responsibility, Incompetence, and Psychopathy: The Lindley Lecture*. Lawrence: University of Kansas, 2013.
Brosig, Jeannette. "Identifying Cooperative Behavior: Some Experimental Results in a Prisoner's Dilemma Game." *Journal of Economic Behavior and Organization* 47, no. 3 (March 2002): 275–90.
Buchanan, Allen, and Russell Powell. "De-moralization as Emancipation: Liberty, Progress, and the Evolution of Invalid Moral Norms." *Social Philosophy and Policy* 34, no. 2 (Winter 2017): 108–35.
Bulbulia, Joseph. "Religious Costs as Adaptations That Signal Altruistic Intention." *Evolution and Cognition* 10, no. 1 (2004): 19–38.
Cavedon, Lawrence, and Sheila Glasbey. "Outline of an Information-Flow Model of Generics." *Acta Linguistica Hungarica* 42, nos. 3/4 (1994): 227–45.
Copp, David. *Morality, Normativity, and Society*. New York: Oxford University Press, 2001.
———. "Why Naturalism?" *Ethical Theory and Moral Practice* 6, no. 2 (June 2003): 179–200.
Davidson, Donald. "Radical Interpretation." *Dialectica* 27, nos. 3/4 (1973): 313–28.
De Quervain, Dominique J.-F., Urs Fischbacher, Valerie Treyer, Melanie Schellhammer, Ulrich Schnyder, Alfred Buck, and Ernst Fehr. "The Neural Basis of Altruistic Punishment." *Science* 305, no. 5688 (August 2004): 1254–258.
Eisenberg, Nancy, Richard A. Fabes, Paul A. Miller, Jim Fultz, Rita Shell, Robin M. Mathy, and Ray R. Reno. "Relation of Sympathy and Personal Distress to Prosocial Behavior: A Multimethod Study." *Journal of Personality and Social Psychology* 57, no. 1 (July 1989): 55–66.
Enoch, David. *Taking Morality Seriously: A Defense of Robust Realism*. Oxford: Oxford University Press, 2011.
Findlay, Leanne C., Alberta Girardi, and Robert J. Coplan. "Links between Empathy, Social Behavior, and Social Understanding in Early Childhood." *Early Childhood Research Quarterly* 21, no. 3 (2006): 347–59.
Foot, Philippa. "Moral Arguments." *Mind* 67, no. 268 (October 1958): 502–13.
———. "Morality as a System of Hypothetical Imperatives." *The Philosophical Review* 81, no. 3 (July 1972): 305–16.
———. *Natural Goodness*. Oxford: Clarendon Press, 2003.
Frank, Robert H. *Passions within Reason: The Strategic Role of the Emotions*. New

York: W. W. Norton, 1988.
Frank, Robert H., Thomas Gilovich, and Dennis T. Regan. "The Evolution of One-Shot Cooperation: An Experiment." *Ethology and Sociobiology* 14, no. 4 (July 1993): 247–56.
Gauthier, David. *Morals by Agreement*. Oxford: Clarendon Press, 1986.
Haidt, Jonathan. *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. London: Penguin, 2012.
Hamlin, J. Kiley, Karen Wynn, and Paul Bloom. "Social Evaluation by Preverbal Infants." *Nature* 450, no. 7169 (November 2007): 557–59.
House, James S., Karl R. Landis, and Debra Umberson. "Social Relationships and Health." *Science* 241, no. 4865 (July 1988): 540–45.
Hume, David. *Enquiry Concerning the Principles of Morals*. Edited by Tom Beauchamp. Oxford: Oxford University Press, 2005.
Hurka, Thomas. *Perfectionism*. New York: Oxford University Press, 1993.
Jackson, Frank. *From Metaphysics to Ethics*. Oxford: Oxford University Press, 1998.
Korsgaard, Christine Marion. *The Sources of Normativity*. Cambridge: Cambridge University Press, 1996.
Kitcher, Phillip. *The Ethical Project*. Cambridge, MA: Harvard University Press, 2011.
Kuhlmeier, Valerie, Karen Wynn, and Paul Bloom. "Attribution of Dispositional States by 12-Month-Olds." *Psychological Science* 14, no. 5 (September 2003): 402–8.
Kurt, Fred, and Khyne U. Mar. "Neonate Mortality in Captive Asian Elephants (*Elephas maximus*)." *Zeitschrift für Säugertierkunde* 61, no. 3 (1996): 155–64.
Kurth, Charlie. "What Do Our Critical Practices Say About the Nature of Morality?" *Philosophical Studies* 166, no. 1 (September 2013): 45–64.
Liszkowski, Ulf, Malinda Carpenter, Tricia Striano, and Michael Tomasello. "12- and 18-Month-Olds Point to Provide Information for Others." *Journal of Cognition and Development* 7, no. 2 (November 2006): 173–87.
Loeb, Don. "Gastronomic Realism—A Cautionary Tale." *Journal of Theoretical and Philosophical Psychology* 23, no. 1 (Spring 2003): 30–49.
Mackie, Gerry. "Female Genital Cutting: A Harmless Practice?" *Medical Anthropology Quarterly* 17, no. 2 (June 2003): 135–58.
Mameli, Matteo. "Meat Made Us Moral: A Hypothesis on the Nature and Evolution of Moral Judgment." *Biology and Philosophy* 28, no. 6 (September 2013): 903–31.
Manne, Kate. "Internalism about Reasons: Sad but True?" *Philosophical Studies* 167, no. 1 (January 2014): 89–117.
Neff, Kristin D. "Judgments of Personal Autonomy and Interpersonal

Responsibility in the Context of Indian Spousal Relationships: An Examination of Young People's Reasoning in Mysore, India." *British Journal of Developmental Psychology* 19, no. 2 (June 2001): 233–57.
Nichols, Shaun. *Sentimental Rules: On the Natural Foundations of Moral Judgment*. Oxford: Oxford University Press, 2004.
Nickel, Bernhard. "Generics and the Ways of Normality." *Linguistics and Philosophy* 31, no. 6 (2008): 629–48.
O'Connor, Cailin. "The Evolution of Guilt: A Model-Based Approach." *Philosophy of Science* 83, no. 5 (January 2016): 897–908.
Parfit, Derek. *On What Matters*. Vol. 2. Oxford: Oxford University Press, 2011.
Pradel, Julia, Harald A. Euler, and Detlef Fetchenhauer. "Spotting Altruistic Dictator Game Players and Mingling with Them: The Elective Assortation of Classmates." *Evolution and Human Behavior* 30, no. 2 (March 2009): 103–13.
Railton, Peter. "Moral Realism." *The Philosophical Review* 95, no. 2 (April 1986): 163–207.
———. "Naturalism and Prescriptivity." *Social Philosophy and Policy* 7, no. 1 (Autumn 1989): 151–74.
Shafer-Landau, Russ. "A Defence of Categorical Reasons." *Proceedings of the Aristotelian Society* 109, no. 1 (August 2009): 189–206.
Southwood, Nicholas. *Contractualism and the Foundations of Morality*. Oxford: Oxford University Press, 2013.
Sperber, Dan, and Nicolas Baumard. "Moral Reputation: An Evolutionary and Cognitive Perspective." *Mind and Language* 27, no. 5 (November 2012): 495–518.
Stanford, P. Kyle. "The Difference between Ice Cream and Nazis: Moral Externalization and the Evolution of Human Cooperation." *Behavioral and Brain Sciences* 41 (2018): e95.
Sterelny, Kim. *The Evolved Apprentice*. Cambridge, MA: MIT Press, 2012.
———. "SNAFUS: An Evolutionary Perspective." *Biological Theory* 2, no. 3 (September 2007): 317–28.
Sterelny, Kim, and Ben Fraser. "Evolution and Moral Realism." *British Journal for the Philosophy of Science* 68 (December 2017): 981–1006.
Street, Sharon. "In Defense of Future Tuesday Indifference: Ideally Coherent Eccentrics and the Contingency of What Matters." *Philosophical Issues* 19 (2009): 273–98.
Streumer, Bart. *Unbelievable Errors*. Oxford: Oxford University Press, 2017.
Sturgeon, Nicholas L. "Moore on Ethical Naturalism." *Ethics* 113, no. 3 (April 2003): 528–56.
Tangney, June Price, and Ronda L. Dearing. *Shame and Guilt*. New York:

Guilford Press, 2003.
Toi, Miho, and C. Daniel Batson. "More Evidence That Empathy Is a Source of Altruistic Motivation." *Journal of Personality and Social Psychology* 43, no. 2 (August 1982): 281–92.
Tomasello, Michael. *A Natural History of Human Morality*. Cambridge, MA: Harvard University Press, 2016.
———. *Why We Cooperate*. Cambridge, MA: MIT Press, 2009.
Turiel, Elliot. *The Culture of Morality: Social Development, Context, and Conflict*. Cambridge: Cambridge University Press, 2002.
Vonasch, Andrew J., Tania Reynolds, Bo M. Winegard, and Roy F. Baumeister. "Death before Dishonor: Incurring Costs to Protect Moral Reputation." *Social Psychological and Personality Science* 9, no. 5 (July 2018): 604–13.
Warneken, Felix, and Michael Tomasello. "Altruistic Helping in Human Infants and Young Chimpanzees." *Science* 311, no. 5765 (March 2006): 1301–3.
Western, Bruce, Jeffrey R. Kling, and David F. Weiman. "The Labor Market Consequences of Incarceration." *Crime and Delinquency* 47, no. 3 (July 2001): 410–27.
Woods, Jack. "Footing the Cost (of Normative Subjectivism)." In *Methodology and Moral Philosophy*, edited by Jussi Suikkanen and Antti Kauppinen, 166–190. London: Routledge, 2018.

Journal of Ethics and Social Philosophy
VOL. 25, NO. 3 · SEPTEMBER 2023
https://doi.org/10.26556/jesp.v25i3.1483

THRESHOLD CONSTITUTIVISM AND SOCIAL KINDS

Mary Clayton Coleman

In "Constitutivism without Normative Thresholds," Kathryn Lindeman raises two objections to what she aptly calls *Threshold Constitutivism*. My aim in this short discussion is to respond to her first objection.[1] Although I will argue that this objection fails, I will also argue that thinking through how to respond to it reminds us of something important—namely, that many of the norm-governed kinds that are directly related to intentional action are social kinds, that is, kinds whose existence conditions we ourselves collectively write.

Constitutivism is an attractive position, on my view, because it aims to show that claims about which actions we ought to perform are true (at least in part) in virtue of the nature of intentional action rather than in virtue of the supposed existence of realist truthmakers that many of us find metaphysically, epistemically, and motivationally puzzling. According to what Lindeman calls naïve constitutivism, the norms that are constitutive of a kind *K* are those norms that an individual must fully satisfy in order to be a *K*. So understood, naïve constitutivism leaves no room for defective kind-members and, thus, no room for it to be true that an individual *K* ought to become better than she is, *qua K*. This means that naïve constitutivism cannot give us a (nonrealist) account of what makes it true that some *K* ought to become better than she is, *qua K*.

As Lindeman explains, many "constitutivists make room for defective kind-members" by rejecting naïve constitutivism and accepting what she calls the *Threshold Commitment*, which says, "For norm-governed kinds, an individual must *at least partially satisfy* the constitutive norms of a kind . . . in order to be a member of that kind."[2] Lindeman then argues that constitutivists who accept this commitment—that is, Threshold Constitutivists—face an insurmountable problem about what I will call *nonthreshold norm-governed kinds*. This is the objection I aim to answer in this discussion.

1 Lindeman's second objection depends on her view, which I do not share, that "Normative Constitutivism has ambitions to be an explanatory strategy for norms *in general*" ("Constitutivism without Normative Thresholds," 238, emphasis added).

2 Lindeman, "Constitutivism without Normative Thresholds," 235–36, emphasis added.

First, of course, I need to explain the objection. *K* is a nonthreshold norm-governed kind if and only if:

1. *K* is a norm-governed kind: *K* is a "goodness-fixing kind the [goodness-fixing] norms of which come from its constitutive nature"; but
2. *K* is not a threshold kind: membership in *K* is not a matter of even minimally meeting the norms that are good-making for members of *K*; instead, the conditions for being a *K* are nonnormative.[3]

From here on, I will refer to nonthreshold norm-governed kinds simply as *nonthreshold kinds* since almost all of the kinds I will discuss are norm governed.

It may seem as if the idea of a nonthreshold kind is incoherent and, thus, as if Lindeman's objection cannot get off the ground. After all, if there are standards about what makes something a good or bad *K that follow from what it is to be a K*, how can something count as a *K* if it fails to meet those standards to any degree at all?

Lindeman does not explicitly consider this challenge, but she does argue that "social kinds like Spouse appear to provide ready cases of [norm-governed] kinds that lack normative thresholds."[4] According to Lindeman, Spouse is a nonthreshold kind because "in contemporary practice, we take legal recognition to be at least partially determinant of *becoming* a spouse, and recognized removal of legal recognition to be a sufficient (and, along with the death of one's spouse, exhaustive) condition on *ceasing* to be a spouse."[5] Furthermore, "one does not become a spouse by being a good enough one, and one cannot cease being a spouse *merely in virtue of* being a bad enough one."[6] In other words, all one has to do to be someone's spouse is to become and remain legally married to that person, and becoming and remaining legally married to someone has nothing to do with whether one is even a minimally good spouse to that person.

Lindeman's account of Spouse mistakenly conflates two different kinds: Spouse and Life Partner. Life Partner is, I contend, a threshold kind. To be someone's life partner is, very roughly, to maintain at least some of the following relationships with her over a long period of time: cohabitation, a very close economic relationship, a very close emotional relationship, a sexual relationship. If you do poorly enough at maintaining these relationships with someone, you fall below the threshold required to count as her life partner.

3 Lindeman, "Constitutivism without Normative Thresholds," 239.

4 Lindeman, "Constitutivism without Normative Thresholds," 238–39. I will follow Lindeman and "use the singular, capitalized noun to pick out the kind, and lowercase uses to pick out instances in the singular or plural" (239n26).

5 Lindeman, "Constitutivism without Normative Thresholds," 239.

6 Lindeman, "Constitutivism without Normative Thresholds," 239.

Contrast that with Spousehood. To become someone's spouse, you and another person (who you are legally eligible to marry) take whatever steps are required to become legally married. In principle, one or both of you could decide to enter into this marriage without any consideration of whether the other person will be even a minimally good life partner. However, most people decide to marry someone only when they think she will be a good (enough) life partner. What it takes to remain a spouse is simply that you and the person you are married to stay alive and neither of you divorces the other. For simplicity's sake, let us focus on divorce.[7] If you live in a state with no-fault divorce laws, then, in principle, either of you could dissolve your marriage without any consideration of whether the other is a good life partner. However, I take it that at least part of the point of no-fault divorce laws is that they allow a married person to end her marriage if she judges that her spouse is no longer a minimally good life partner without having to prove to the state that her judgment is correct.

In sum, I agree with Lindeman that Spouse is a nonthreshold kind. However, *contra* Lindeman, Spouse is not a norm-governed kind. Spouse *seems* like a norm-governed kind because Life Partner is a norm-governed kind, and we have linked Spousehood closely to Life Partnerhood, both in terms of our individual decisions about whether to become someone's spouse and in terms of our societal decisions about what laws should govern Spousehood.

It may now seem even more tempting to argue that the idea of a nonthreshold kind is incoherent, but we should resist that temptation. It would be perfectly coherent for there to be a kind *K* such that the goodness-fixing norms for *K*s follow from the nature of *K*-hood, and yet the conditions by which an individual becomes (and remains) a *K* have nothing to do with whether that individual complies with those norms. In principle, we can write the existence conditions for our social kinds—the kinds we collectively create—so that they say whatever we want them to say.

Let us return, then, to Lindeman's claim that nonthreshold kinds pose a fatal problem for Threshold Constitutivists. She argues as follows: Threshold Constitutivists cannot explain the constitutive norms of nonthreshold kinds because their explanation of the norms of a kind *K* depends on the idea that in order for an individual to be a *K*, that individual has to meet those norms, at least minimally, and that is not true for nonthreshold kinds.

My response to Lindeman, very briefly, is this. There are two different questions we need to answer in order to explain the goodness-fixing norms of any

7 I assume it is obvious that a dead person cannot maintain any of the relevant relationships and thus cannot be even a minimally good life partner.

norm-governed kind that is directly related to intentional action. Threshold Constitutivists can answer the first question exactly the same way any constitutivist would, and they can give an entirely satisfactory answer to the second question as well.

The first question we have to answer to explain the norms of a norm-governed kind *K* is what I call the *content question*: What are the goodness-fixing norms of *K*s? In other words, what is the content of those norms? Threshold Constitutivists can answer this question exactly the same way any constitutivist would—namely, by deriving the content of the norms from the nature of the kind. (A good house keeps the weather out. A good doctor helps her patients become and remain healthy. Etc.)

If a kind *K* is directly related to intentional action, then in order to explain the norms of *K*, we also have to answer what I call the *compliance question*: Why should an individual *K* comply with the goodness-fixing norms of *K*-hood? For example, why should a doctor help her patients become and remain healthy? Why should someone building a house build one that keeps the weather out?

Threshold Constitutivists begin their answer to the compliance question this way: if you do not comply with the goodness-fixing norms of *K*, you will not be (or will not be making) a *K* at all. For example, "According to Korsgaard, what should make you interested in building a *good* house is the risk that if you do not do it well enough, you will not end up with a house at all."[8] (It is crucial to notice that this answer is not complete unless it also addresses the question, Why should one be [or make] a *K*? How best to address this question is not Lindeman's focus in her paper, and it will not be mine here, but I will return to it briefly below.)

Lindeman then argues, quite rightly, that Threshold Constitutivists cannot say the same thing about why someone should comply with the norms of a nonthreshold kind, since someone can completely fail to comply with those norms and still be (making) a member of that kind.

Here—I submit—is how Threshold Constitutivists should answer the compliance question about nonthreshold kinds, in three steps.

Step One. Simplifying greatly, but in a way that will not matter for my argument, we can divide all conceivable norm-governed kinds into two types:

1. *Good kinds*: those where it would be good if at least some individuals complied with the goodness-fixing norms of the kind; and
2. *Bad kinds*: those where it would be bad if anyone complied with the goodness-fixing norms of the kind.

8 Lindeman, "Constitutivism without Normative Thresholds," 238.

Step Two. If there are any nonthreshold kinds that are good kinds, then we ought to change their existence conditions so that they become *threshold* kinds. Once those changes are made, the Threshold Constitutivist will have no trouble answering the compliance question about the kinds in question since they will have become threshold kinds.

Take Doctor, for example. This is a norm-governed kind since someone is a better doctor the more effective she is at helping her patients become and remain healthy. However, suppose Doctor were a nonthreshold kind. That is, suppose that in order to become (and remain) a doctor, a person did not have to be even minimally good at helping her patients become and remain healthy. For example, suppose we made medical licenses available to anyone who wanted one and could pay the fee—the way we make fishing licenses available—and suppose we allowed people to keep their medical licenses no matter how ineffective they were at helping their patients. In that case, if someone were a doctor but did not see any reason to comply with the goodness-fixing norms of Doctorhood, we would not be able to convince her to comply by saying, "If you do not, you will not be a doctor at all." However, the problem with this awful scenario is not Threshold Constitutivism. The problem is that we should not allow someone to become or remain a doctor if she is not at least a minimally good doctor. In short, Doctor should be a threshold kind.

I hear an objection: someone with a medical license who is wholly ineffective at helping people become and remain healthy is not really a doctor. My reply is that one of the following is true:

1. Such a person *is* a doctor because she is legally permitted to treat patients.
2. You are right about *this* attempted example of a nonthreshold kind. Doctor is (already) a threshold kind because having a license is not sufficient (or perhaps even necessary) for being a doctor. But some other example would work in its place.
3. No other example would work in its place, in which case the idea of a nonthreshold norm-governed kind is incoherent after all, and Lindeman's objection, which is the focus of my discussion, never gets off the ground.

Step Three. If the compliance question were asked about a nonthreshold kind that is a bad kind, Threshold Constitutivists would have no trouble giving the right answer. Imagine a kind we might call Gratuitous Tormentor. The goodness-fixing norm of this kind says, "Cause other people as much pain as you can in such a way that no one is benefited." Now imagine that Gratuitous Tormentor is a nonthreshold kind. For example, imagine that someone can

become a gratuitous tormentor simply by obtaining a Gratuitous Tormentor license, regardless of whether she is even a minimally "good" gratuitous tormentor—that is, regardless of whether she is even minimally "good" at causing others gratuitous pain. (A Gratuitous Tormentor license would make it legal for the licensee to cause others gratuitous pain.) Finally, imagine that the compliance question were asked about this kind: Why should someone who is a gratuitous tormentor comply with the goodness-fixing norm(s) of this kind? The correct answer to this question is, "She shouldn't. No one should." A Threshold Constitutivist can give this answer just as well as anyone else.

I hear another objection. The three-step answer I have just given depends on a distinction between good kinds and bad kinds, but it seems impossible to give a constitutivist account of this distinction. After all, how could the goodness (or badness) of someone's following the norms of a kind *K* be grounded in the goodness-fixing norms of some other kind *K** (much less in the goodness-fixing norms of *K* itself), as it would have to be for the account to be constitutivist?

Giving a constitutivist account of the distinction between good kinds and bad kinds would be challenging. However, if Threshold Constitutivists can answer the compliance question about any threshold kinds, then they can also give an account of the good kind/bad kind distinction. Why? Because the ultimate question they have to address to answer the compliance question is also the ultimate question they have to address to give an account of this distinction. Let me explain.

The compliance question asks, Why should an individual *K* comply with the norms of *K*-hood? The Threshold Constitutivist answer is "because if she does not, she will not be a *K*." Thus, as I mentioned above, to address the compliance question fully, Threshold Constitutivists have to address the question, Why should an individual *K* be (or remain) a *K*? The most straightforward Threshold Constitutivist answer to that is "because, given this individual's situation, the goodness-fixing norms of Intentional Agenthood require her to be a *K*."[9] And once Threshold Constitutivists give that answer, they need to address what I call the *intentional agent question*, which asks, Why should an individual be (or remain) an intentional agent?[10]

9 This is the *most straightforward* Threshold Constitutivist answer because it does not involve assuming that the goodness-fixing norms of any kinds other than Intentional Agent and *K* are normative for the individual in question.

10 I say Threshold Constitutivists need to *address* this question—rather than *answer* it—because Hille Paakkunainen argues that "the nature of agency can in principle ground authoritative reasons for agents to act, even if there isn't, in addition, a reason to be an

Return now to the good kind/bad kind distinction. Any account of this distinction has to answer the question, For any norm-governed kind *K*, would it be good if someone complied with the goodness-fixing norms of *K*? If *K* is a social kind, as it has to be in order to be a nonthreshold kind, the question can be put this way: Should the existence condition(s) for *K*s be rewritten in such a way that in order for someone to be (or remain) a *K*, she must comply with the norms of *K*? The most straightforward Threshold Constitutivist answer to that question will be: "Yes, if in order for us to comply with the norms of Intentional Agenthood, we need to live in a society in which at least some people comply with the norms of *K*; no, if in order for us to comply with the norms of Agenthood, we need to live in a society in which no one complies with the norms of *K*."[11] And once Threshold Constitutivists give this answer, they then need to address the intentional agent question, which asks: Why should an individual agent be (or remain) an intentional agent?

In sum, the objection from Lindeman that I have focused on does not undermine Threshold Constitutivism. However, by thinking through how to respond to that objection, we are reminded of something important: many of the norm-governed kinds that are directly related to intentional action are social kinds—that is, kinds whose existence conditions we ourselves collectively write. Everyone, whether constitutivist or not, needs to think seriously about what those existence conditions should be because what they are is up to us.

Illinois Wesleyan University
mary.clayton.coleman@gmail.com

REFERENCES

Lindeman, Kathryn. "Constitutivism without Normative Thresholds." *Journal of Ethics and Social Philosophy* 12, no. 3 (December 2017): 231–58.

agent" and I want to remain neutral about whether Paakkunainen is right about this ("Doing Away with the 'Shmagency Objection' to Constitutivism," 433).

11 For many—perhaps most—potential kinds, the answer to this question will not be so straightforward. Some kinds will be entirely optional with respect to agenthood; that is to say, whether anyone in our society complies with their goodness-fixing norms will make no difference to our ability to comply with the norms of Intentional Agenthood. The instantiation of other kinds by someone in our society will make it easier or harder for us to comply with the norms of Agenthood, but their instantiation will not be strictly necessary for nor strictly incompatible with our complying with those norms.

Paakkunainen, Hille. "Doing Away with the 'Shmagency Objection' to Constitutivism." *Manuscrito* 41, no. 4 (October–December 2018): 431–80.

Journal of Ethics and Social Philosophy
VOL. 25, NO. 3 · SEPTEMBER 2023
https://doi.org/10.26556/jesp.v25i3.2742

IS CONTRASTIVE CONSENT NECESSARY FOR SECONDARY PERMISSIBILITY?

Peter A. Graham

In nonconsequentialist ethics there is a phenomenon of secondary permissibility whereby an otherwise morally impermissible option is rendered permissible by the presence of another option. There is some controversy, however, about what the deontic mechanics of this moral phenomenon are. In this paper, I critique a recent approach, that of Theron Pummer, to the deontic mechanics of secondary permissibility.

The phenomenon of secondary permissibility is evident by way of a comparison of the moral data in three cases:

> *Turn*: A trolley is about to kill five innocent strangers. You can turn the trolley onto me, thereby saving the five and killing me.
>
> *Hurl*: A trolley is about to kill five innocent strangers. You can hurl me at the trolley, thereby stopping the trolley and saving the five, but also paralyzing me.
>
> *TurnHurl*: A trolley is about to kill five innocent strangers. You can turn the trolley onto me, saving the five and killing me. You can instead hurl me at the trolley, saving the five and paralyzing me.

Intuitively, it is permissible for you to turn the trolley onto me in Turn but impermissible for you to hurl me at the trolley in Hurl. And this is so, even though being paralyzed is less of a harm to me than is being killed. Interestingly—and here is where the phenomenon of secondary permissibility enters—intuitively, it is permissible for you to hurl me at the trolley and impermissible for you to turn it onto me in TurnHurl. Kamm, who introduced us to this phenomenon and dubbed it "secondary permissibility," would say that in TurnHurl, your hurling me at the trolley is "secondarily permissible."[1]

1 Kamm, *Morality, Mortality* and *Intricate Ethics*.

Pummer argues that the correct explanation of the secondary permissibility data in TurnHurl must make appeal to a notion of contrastive consent.[2] According to Pummer, for it to be permissible for you to hurl me in TurnHurl, I must give you my contrastive consent to being hurled by saying to you something like, "You may hurl me at the trolley rather than turn it onto me." Absent such an avowal on my part, he claims, it would not be permissible for you to hurl me in TurnHurl.

Crucially for Pummer, in giving my contrastive consent to being hurled rather than having the trolley turned onto me, I need not give any ordinary noncontrastive consent either to being hurled or to having the trolley turned onto me. It is consistent with my giving you my contrastive consent to being hurled rather than having the trolley turned onto me, according to Pummer, for me to also say, "You may not hurl me at the trolley, nor may you even turn it onto me." Essential to contrastive consent is that "rather" construction—"you may hurl me at the trolley *rather* than turn it onto me."

The details of Pummer's account of how contrastive consent supposedly explains the moral data in TurnHurl are not important for my purposes. My question is simply: is my avowing anything like that which Pummer says constitutes the giving of contrastive consent necessary for it to be permissible for you to hurl me (and impermissible for you to turn the trolley onto me) in TurnHurl? It seems not. I need not utter (or even think) anything like "you may hurl me at the trolley rather than turn it onto me" for it to be permissible for you to hurl me in TurnHurl. Rather, it seems that it just *is* permissible for you to hurl me in TurnHurl (and that it just *is* impermissible for you to turn the trolley onto me in that case).[3]

As a matter of fact, people do not ordinarily say such strange things as "You may not *X* and you may not *Y*, but you may *X* rather than *Y*." Though people do often say things like "You may *X*," thereby giving plain old consent to *X*-ing,

2 Pummer, "Contrastive Consent."

3 Pummer claims that it would be odd for me to refuse contrastive consent to being hurled rather than having the trolley turned onto me in TurnHurl (Pummer, "Contrastive Consent," 683–84). But even if it would be odd for me to refuse contrastive consent in such circumstances—i.e., even if it would be odd, after having been both introduced to the concept of contrastive consent and given the explanation of why such consent is needed for your hurling me to be permissible—for me to reply no to the query "Do you contrastively consent to being hurled rather than having the trolley turned onto you?"—that is neither here nor there. Were I neither introduced to the concept of contrastive consent nor explicitly asked the question, it just would not even occur to me to say the kind of thing Pummer thinks is necessary for me to say to make it permissible for you to hurl me in TurnHurl. As I go on to note in the text, Pummer's contrastive consent is a weird bird; it is not anything we see in common parlance at all.

and they do also sometimes say "You may *X* if *Y*," thereby giving conditional consent to *X*-ing given *Y*, they simply do not ever say things like that which, according to Pummer, is the giving of contrastive consent.

Here, however, is something we might well imagine me saying in TurnHurl:

> Do not hurl me at the trolley! And do not turn the trolley onto me! But *if* you are going to harm me to save those five people, then hurl me at the trolley rather than turn it onto me!

But in this case, it seems, I am not so much giving contrastive consent to being hurled rather than having the trolley turned onto me (after refusing to noncontrastively consent to either) as *commanding* you to hurl me rather than turn the trolley onto me if you are going to do one of them (after commanding you to do neither). So, if there is anything contrastive in the vicinity of what people plausibly might actually say in cases like TurnHurl, it is the issuance of contrastive commands rather than the giving of contrastive consent.

But is my issuing a contrastive command like that above even necessary for it to be permissible for you to hurl me in TurnHurl? Again, it seems not. Imagine for instance that I did not issue any such contrastive command but that you knew of me that, though I most preferred not being harmed in any way, I did prefer being hurled and thus paralyzed to having the trolley turned onto me and thus killed. (Though I am skeptical that there is any such thing as contrastive consent, preferences most certainly are contrastive, and essentially so.) I cannot see how it would be impermissible for you to hurl me in such a case. In fact, it does not even seem necessary for me to have a preference for being hurled for it to be permissible for you to hurl me in TurnHurl. For if I had no preferences between your hurling me and your turning the trolley onto me—if I were just indifferent between your hurling me, thereby paralyzing me, and your turning the trolley onto me, thereby killing me—then surely, in that case, it would be permissible for you to hurl me—thereby saving the five and causing me less harm.

At most, it seems, when it comes to preferences, what is necessary for it to be permissible for you to hurl me is that I *not* have a preference for having the trolley lethally turned onto me rather than being paralyzingly hurled at it. One might think that it would be wrong of you to hurl me at the trolley in TurnHurl if I actively preferred having it lethally turned onto me to my being paralyzingly hurled at it. But if I have no preference between the two, surely it would be permissible for you to take the option which harms me less to achieve the good of saving the five.

Pummer would disagree with all of this. He would say that the foregoing fails to take seriously the fact that harming me *as a means* to achieving the good of

saving the five is morally worse than harming me *as a side effect* of achieving it. It is more *pro tanto* morally wrong to harm someone as a means than it is to harm them as a side effect, and it is that greater *pro tanto* moral wrongness that makes my issuing some sort of contrastive consent (or contrastive command) required for it to be permissible for you to hurl me in TurnHurl.[4] Now, whether harming someone as a means to some good is more *pro tanto* morally wrong than is harming them as a side effect of achieving that good, as I say, it just does not seem true that my saying any such thing is necessary for your hurling me to be permissible in TurnHurl. But put that to one side. Is harming someone as a means more *pro tanto* morally wrong than harming them as a side effect? No. It is not.

If I am a morally conscientious person (one who prefers acting less [*pro tanto*] morally wrongly to acting more [*pro tanto*] morally wrongly), will I be more concerned not to harm people as a means to achieving my ends than I will be not to harm them as a side effect of achieving them? I do not think so. Imagine that a trolley is about to hit a single person to whom it would cause a broken toe if allowed to strike them. Now suppose I can press one of two buttons, either of which will save the one from having his toe broken by the trolley. Pressing the first will do so by redirecting the trolley onto a separate track where it will kill another innocent person, *A*. Pressing the second will do so by hurling a different innocent person, *B*, into the trolley, thereby killing him. Pressing either button would be impermissible, of course. But would I as a morally conscientious agent be *more* inclined to press the first button than the second just because pressing the second would cause the harm to *B* as a means, whereas pressing the first would cause the harm to *A* as a side effect? I cannot see that I would. As a matter of fact, if pressing the second button would not kill *B* but merely paralyze him, I am sure, as a morally conscientious person, I would prefer pressing the second button to pressing the first (though, as a morally conscientious person I would of course also prefer not pressing either button to pressing either of them).[5]

4 The thought that means-harming is morally worse (i.e., more *pro tanto* morally wrong) than side-effect-harming does indeed seem to be Pummer's ground for thinking that something like contrastive consent is necessary for the permissibility of your hurling me in TurnHurl. He writes: "In TurnHurl, in the absence of contrastive consent, the barrier against being hurled (harmed as a means) is stronger than the barrier against being turned onto (harmed as a side effect)…. My contrastive consent makes it the case that the barrier against being hurled is weaker than the barrier against being turned onto" (Pummer, "Contrastive Consent," 682). This talk of barriers, I believe, is just another way of talking about *pro tanto* moral wrongness; the higher the barrier against performing a certain action, the more *pro tanto* morally wrong is that action.

5 Here, and throughout, I am presupposing for the sake of argument that whether an action counts as harming as a means or harming as a side effect is solely a function of the causal

Or suppose that pressing either of the two buttons has a 100 percent chance of preventing the broken toe but only has a very small chance, *n* percent, of either killing *A* as a side effect (by pressing the first button) or paralyzing *B* as a means (by pressing the second button). (Suppose pressing each button has a 100–*n* percent chance of causing the trolley to just stop, and whereas pressing the first button has an *n* percent chance of lethally turning the trolley onto *A*, pressing the second has an *n* percent chance of paralyzingly hurling *B* at the trolley, thereby causing it to stop.) Surely, if I were a morally conscientious person, I would prefer pressing the second button to pressing the first. (And I would most certainly have that preference if the chance that pressing the second button has of paralyzing *B* as a means is even the tiniest bit less than the chance that pressing the first button has of killing *A* as a side effect.)

What is more, if harming someone as a means were more *pro tanto* morally wrong than harming them as a side effect, it could only be subjectively permissible to press the first button. (Were it more *pro tanto* morally wrong to means-paralyze *B* than to side-effect-kill *A*, the expected wrongness of pressing the first button should be lower than that of pressing the second.) But that is simply implausible. There is no chance, *n* percent, such that pressing the first button with an *n* percent chance of killing *A* as a side effect of preventing the broken toe would be subjectively permissible while pressing the second button with an *n* percent chance of paralyzing *B* as a means to preventing it is subjectively impermissible. (It is clear that there must be some minuscule chances of causing someone else to die [or be paralyzed], whether as a means or as a side effect, that it is subjectively permissible to take in order to prevent a broken toe. We routinely subjectively permissibly subject others to minuscule chances of death [and paralyzation] in order to alleviate harms less severe than a broken toe, as when we drive to the pharmacy, thereby risking killing [paralyzing] others with our car, to purchase headache medicine.) So, harming someone as a means to some good just is not more *pro tanto* morally wrong than is harming them as a side effect of achieving it.

In response to these arguments, Pummer might contend that the greater *pro tanto* moral wrongness of means-harming as opposed to side-effect-harming only holds when the side-effect-harming is relevantly proportionate.[6] (The

relations between the action, the harm caused, and the good achieved, and *not* a function of the intentions or mental states of the agent. Views according to which whether an action counts as means-harming or as side-effect-harming, and thus potentially the permissibility of that action, is a function of the intentions or mental states of the agent are notoriously fraught. Even were it allowed that whether an action counts as means-harming or as side-effect-harming depends on the intentions or mental states of the agent, my arguments could be suitably modified to reach the very same conclusions for which I argue in the text.

6 I am grateful to an anonymous referee for this suggestion on Pummer's behalf.

side-effect harming's being relevantly proportionate just means that it would ordinarily be permissible, given the good it would bring about, were it the only option, other than doing nothing, that one had.) This maneuver might accommodate the datum that it is not more *pro tanto* morally wrong to paralyzingly hurl *B* at the trolley to prevent the broken toe than it is to lethally turn the trolley onto *A* to prevent it. And that is because lethally turning the trolley would not be relevantly proportionate when the good that would be achieved by doing so is just the prevention of a broken toe.

This maneuver will not work. Not only does it seem *ad hoc*, but the view that means-harming is, in general, morally worse than proportionate side-effect-harming is mistaken. To see this, just note that it is not more *pro tanto* morally wrong to proportionately means-paralyze someone to bring about a good than it is to proportionately side-effect-kill someone to bring about that good. Consider:

> *Three Option Trolley*: A trolley is about to kill *n* strangers. There are two ways one can save them: one can either divert the trolley onto a side track, thereby killing *A*, or one can hurl *B* at the trolley, thereby paralyzing him and stopping the trolley.

If it would be permissible to hurl *B* were the turning option not available (that is, were *n* such that hurling *B* would be proportionate in Three Option Trolley), then surely not only would hurling *B* be permissible (and turning the trolley onto *A* impermissible) in Three Option Trolley, it would most certainly be less *pro tanto* morally wrong than turning the trolley onto *A* would be. So, means-paralyzing is *not*, in general, morally worse than proportionate side-effect-killing.

Now perhaps Pummer might suggest that it is only disproportionate means-paralyzing that is morally worse than proportionate side-effect-killing. Again, this too seems to be an *ad hoc* maneuver. But, even despite that, it as well seems to be a mistaken view. By stipulation, means-paralyzing someone to save five people from death is disproportionate, whereas side-effect-killing someone to save five people from death is proportionate. But if one could press a button (button 1) that has a 100 percent chance of saving the five and a minuscule chance, *n* percent, of saving them by lethally turning the trolley onto *A*, or press another button (button 2) that has a 100 percent chance of saving the five and the same minuscule chance, *n* percent, of saving them by paralyzingly hurling *B* into the trolley, surely it would only be subjectively permissible to press button 2 (assuming, that is, that *n* is small enough that in a version of the case in which pressing button 2 was the only option one had, aside from doing nothing, pressing it would be subjectively permissible). (Suppose pressing each button

has a 100–*n* percent chance of causing the trolley to just stop, and whereas pressing the first button has an *n* percent chance of lethally turning the trolley onto *A*, pressing the second button has an *n* percent chance of paralyzingly hurling *B* at the trolley, thereby causing it to stop.) This shows that it is not more *pro tanto* morally wrong to means-paralyze *B* than it is to side-effect-kill *A*. (Were it more *pro tanto* morally wrong to means-paralyze *B* than to side-effect-kill *A*, the expected wrongness of pressing button 1 should be lower than that of pressing button 2; but, as it is intuitive that only pressing button 2 would be subjectively permissible in such a case, it seems that the expected wrongness of pressing button 2 must, in fact, be lower than that of pressing button 1. And if that is right, then it must be that means-paralyzing *B* to save the five is less, not more, *pro tanto* morally wrong than is side-effect-killing *A* to save them.) And that is true even though side-effect-killing someone to save five from death is proportionate, whereas means-paralyzing someone to save five from death is not. So, it does not save the proposal that means-harming is morally worse than side-effect-harming to restrict the claim to disproportionate means-harmings and proportionate side-effect-harmings. Harming someone as a means to some good just is not more *pro tanto* morally wrong than is harming them as a side effect of achieving it.

Now you might think that the moral data in Turn and Hurl show that harming someone as a means is more *pro tanto* morally wrong than is harming someone as a side effect. That fact, you might think, is what explains why it is permissible for you to lethally turn the trolley onto me in Turn but impermissible for you to paralyzingly hurl me at the trolley in Hurl. Whereas the good of saving five people is enough to outweigh the *pro tanto* moral wrongness of killing me as a side effect in Turn, it is not enough to outweigh the greater *pro tanto* moral wrongness of paralyzing me as a means in Hurl. But that overly simple explanation, according to which the good of saving the five is in some way *weighed against* the *pro tanto* moral wrongness of paralyzing as a means and killing as a side effect, need not be the correct explanation of the moral data in Turn and Hurl. Rather, it might just be that harming people in the course of saving others is permissible via certain causal mechanisms but not via others. That fact, however, need not be cashed out in terms of some greater *pro tanto* moral wrongness of harming as a means rather than harming as a side effect.[7] As a matter of fact, as my arguments above have shown, it should not be so cashed out.

So where does that leave us? It does not seem that my issuing anything like contrastive consent to being hurled rather than having the trolley turned onto

7 For instance, the solution to the trolley problem offered in Kamm, *Intricate Ethics*, makes no mention of a greater *pro tanto* moral wrongness of harming as a means as opposed to harming as a side effect.

me is necessary for its being permissible for you to hurl me in TurnHurl. And the thought that such an utterance is necessary because harming someone as a means is more *pro tanto* morally wrong than is harming them as a side effect is simply mistaken. At most, what is necessary for its being permissible for you to hurl me at the trolley in TurnHurl is that I *not* have a preference for having the trolley lethally turned onto me rather than my being paralyzingly hurled at it. (Even here, though, the moral power of my preferences is not absolute: if hurling me would only break my arm, instead of paralyzing me, my preference for having the trolley lethally turned onto me rather than having my arm broken by being hurled at it would not make it permissible for you to lethally turn the trolley onto me. In such a case, of turning the trolley onto me and hurling me, only hurling me would be permissible.)

If nothing like contrastive consent is necessary for the permissibility of your hurling me in TurnHurl, and only my lacking a preference for having the trolley turned onto me rather than my being hurled at it is, then we need an account of secondary permissibility that is not sensitive to facts about contrastive consent but is sensitive to the preferences of the potential victim.[8]

University of Massachusetts Amherst
pgraham@umass.edu

REFERENCES

Graham, Peter A. "'Secondary Permissibility' and the Ethics of Harming." *Journal of Moral Philosophy* 18, no. 2 (April 2021): 156–77.

Kamm, F. M. *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. New York: Oxford University Press, 2007.

———. *Morality, Mortality*. Vol. 2, *Rights, Duties, and Status*. New York: Oxford University Press, 1996.

Pummer, Theron. "Contrastive Consent and Secondary Permissibility." *Philosophy and Phenomenological Research* 106, no. 3 (May 2023): 677–91.

8 In Graham, "'Secondary Permissibility' and the Ethics of Harming," I offer an account of secondary permissibility which is not sensitive to facts about contrastive consent. Though that account is not sensitive to the preferences of the potential victim, I think it could be easily amended so that an active preference of the potential victim for having the trolley lethally turned onto her rather than her being paralyzingly hurled at it can act as a defeater for the permissibility of hurling her (and the impermissibility of turning the trolley onto her) in TurnHurl.

I am grateful to Theron Pummer and two anonymous referees for this journal for their helpful comments on previous drafts of this paper.