JOURNAL *of* ETHICS & SOCIAL PHILOSOPHY

VOLUME XX · NUMBER 3
November 2021

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. 20, NO. 3 · NOVEMBER 2021
https://doi.org/10.26556/jesp.v20i3.965

OVERRIDING ADOLESCENT REFUSALS OF TREATMENT

Anthony Skelton, Lisa Forsberg, and Isra Black

ADOLESCENTS are routinely treated differently from adults, even when they possess agential capacities that are not dissimilar. Some instances of differential treatment rely on the assumption that responsible adults or institutions are better placed to direct an adolescent's life. In this article we attempt to make philosophical sense of one notable case of differential treatment of adolescents: the concurrent consents doctrine in the law of England and Wales (and other jurisdictions).[1] Our discussion of this doctrine may shed light on the justification for treating adolescents differently from (and paternalistically compared to) adults in medical and other domains.

According to the concurrent consents doctrine, adolescents found to have decision-making capacity have the power to consent to—and thereby, all else being equal, permit—their own medical treatment. However, adolescent refusals of treatment do not have the power to always render treatment impermissible; other parties—that is, individuals who exercise parental responsibility, or a court—retain the authority to consent on their behalf.

The concurrent consents doctrine is puzzling. The adolescents of interest to us possess the minimum rationality considered necessary for agency. When adults possess the same, their decisions in respect of medical treatment are normatively determinative. Yet under the concurrent consents doctrine, the consents of adolescents who possess the same threshold degree of rationality are treated as normatively determinative, but their refusals are not always so treated. At the same time, the concurrent consents doctrine seems intuitively plausible. It attempts to strike a balance between protecting adolescent well-being and respecting burgeoning autonomy.

How might we justify the asymmetry in the normative power of consent to

1 See, e.g., Child and Family Services Act, CCSM c. C80 (1985) (Manitoba); A.C. v. Manitoba (Director of Child and Family Services) [2009] 2 SCR 181 (Can.); Law Reform Commission (Ireland), *Children and the Law*, para. 2.160; Children (Scotland) Act, 1995 c. 36, sec. 2; Minister for Health v. AS, [2004] WASC 286.

and refusal of medical treatment posited by the concurrent consents doctrine?[2] In this article, we develop a view supporting some instances of differential treatment of adult and adolescent agents, including possibly the concurrent consents doctrine. Our account harnesses the strengths of rival defenses of differential treatment, while avoiding their infelicities.

In section 1, we briefly outline the legal regime for concurrent consents in England and Wales. In sections 2 and 3, we discuss and reject two attempts to defend the asymmetry in consent to and refusal of medical treatment by reference to transitional paternalism. In section 4, we consider and reject a stage-of-life justification for differential treatment. In section 5, building on the critical insights of the previous sections, we articulate a new rival justification for differential treatment based on a conception of adolescent well-being that is distinct from that of adults and younger minors. This seems to offer the most promising support for the concurrent consents doctrine. We then defend our view against three objections.

By way of preliminaries, it is important to clarify our focus. There seem to be at least two general strategies for justifying concurrent consents. The first strategy focuses on adolescent decision-making capacity—for example, by relying on a risk-relative standard of capacity, according to which refusal with likely very poor outcomes requires greater competence.[3] The second strategy attempts to justify concurrent consents, even on the assumption that adolescents possess decision-making capacity in respect of the choice to consent to or to refuse treatment. Our paper engages with justificatory strategies of the second kind.

We stipulate that the cases with which we are concerned are those in which the treatment is (at least) in the adolescent's clinical best interests, the treatment is standard with a high probability of success, and refusal carries a high probability of severe harm or death. For simplicity, we will often refer to such cases as relating to serious medical treatment.

1. THE LEGAL REGIME FOR CONCURRENT CONSENTS IN ENGLAND AND WALES

In England and Wales, health professionals must, as a general matter, gain valid

2 For legal consideration of this issue, see Eekelaar, "White Coats or Flak Jackets?"; Elliston, "If You Know What's Good for You"; Harmon, "Body Blow"; Gilmore and Herring, "'No' Is the Hardest Word"; and Lowe and Juss, "Medical Treatment."

3 See, e.g., Buchanan and Brock, *Deciding for Others*; Wicclair, "Patient Decision-Making Capacity and Risk"; Wilks, "The Debate over Risk-Related Standards of Competence." If one inclines toward the risk-relative approach to decision-making capacity, our discussion potentially supplements that argumentative strategy. If one rejects the risk-relative view but finds the asymmetry between adolescent consents and refusals intuitively plausible, our discussion explores alternative routes to justification of the concurrent consents doctrine.

consent for medical treatment to be lawful.[4] Adults eighteen years of age and over generally possess the power to determine whether to undergo medical treatment; no other party has the power to validly consent to or refuse their treatment.[5] Children aged under sixteen years generally have no power to make decisions (that is, consent or refuse) in respect of their own medical treatment; rather, any such decisions are to be made by individuals who exercise parental responsibility over the child.[6]

Notwithstanding the above, all else being equal (that is, assuming adequate information provision and the absence of undue influence), any minor under sixteen years of age may gain the power to consent to her own medical treatment when she satisfies the requirements of the test for decision-making capacity established in *Gillick v. West Norfolk and Wisbech* AHA—that is, when she "achieves sufficient understanding and intelligence to enable … her to understand fully what is proposed."[7] However, the acquisition of *Gillick* competence does not entail the disappearance of the power to consent to treatment on the adolescent's behalf by the individuals who exercise parental authority or by the courts.[8]

4 Aintree University Hospitals NHS Foundation Trust v. James, [2013] UKSC 67. Consent will be valid when *P* possesses adequate information about the intervention offered, per Chatterton v. Gerson, [1981] QB 432; *P* possesses decision-making capacity, per Mental Capacity Act, 2005, c. 9 (hereafter cited as MCA 2005); and autonomy-undermining external influence is absent, per Re T (Adult: Refusal of Treatment), [1993] Fam 95 (CA Civ). Treatment may be provided without consent to some individuals detained under the Mental Health Act 1983. This is special, rather than general, law. Treatment may also be provided without consent to individuals aged sixteen or over who lack decision-making capacity, subject to the requirement that the intervention is in the patient's best interests. In such cases, consent is deemed by operation of law; see MCA 2005, secs. 4 and 5.

5 See, e.g., Ms B v. An NHS Hospital Trust, [2002] EWHC 429 (Fam).

6 Children Act, 1989 c. 41, sec. 3(1); see, e.g., Gillick v. West Norfolk and Wisbech AHA, [1986] AC 112 (HL) (hereafter cited as *Gillick*).

7 *Gillick*, 189 (Lord Scarman).

8 In *Gillick*, Lord Scarman holds that "the parental right to determine whether or not their minor child below the age of 16 will have medical treatment *terminates*" upon the acquisition of *Gillick* competence (188–89, emphasis added). Some commentators interpret Lord Scarman's dictum as authority for the proposition that the legal power to consent to and refuse medical treatment transfers from individuals who exercise parental responsibility to adolescents upon the acquisition of *Gillick* competence by the latter—e.g., Bainham, "The Judge and the Competent Minor." Subsequent legal decisions reject this view, holding that "the parental right to determine" refers only to the ability to veto valid consent provided by competent minors; see Re R (A Minor) (Wardship: Consent to Treatment), [1992] Fam 11 (CA) (hereafter cited as *Re R*); and Re W (A Minor) (Medical Treatment: Court's Jurisdiction), [1993] Fam 64 (CA) (hereafter cited as *Re W*).

Rather, parents and the courts retain the power to consent concurrently with the adolescent.[9]

The position with regard to concurrent consents is similar for adolescents aged sixteen and seventeen. In virtue of the Family Law Reform Act 1969, section 8(1), consent to "surgical, medical or dental treatment" by these minors is legally effective upon meeting the conditions for valid consent applicable to adults. As such, adolescents sixteen and seventeen years of age benefit from a rebuttable presumption of capacity to consent to medical treatment.[10] However, the Family Law Reform Act 1969 section 8(3) preserves "any consent which would have been effective if [section 8(1)] had not been enacted." The courts have interpreted section 8(3) as preserving concurrent consents by individuals exercising parental responsibility or the courts on behalf of sixteen- and seventeen-year-olds.[11]

For competent adolescents there exists, then, an asymmetry in the normative power of consent and refusal. All else being equal, an adolescent may give legally effective consent to treatment (unlike minors lacking capacity), but their valid refusal of treatment may not (unlike adults) be legally effective if the individual(s) exercising parental responsibility or the court consent and thereby render medical treatment lawful.[12] In what follows, we consider how this asymmetry might be supported philosophically.

9 *Re R*, 23–24.

10 MCA 2005, secs. 1(2) and 2(5). See secs. 2(1) and 3(1) for the test for capacity.

11 *Re W*, 84. In the recent case of NHS Trust v. X, [2021] EWHC 65 (Fam), Munby expressed the view that both *Re R* and *Re W* remain good law.

12 There is some doctrinal uncertainty about the scope of the concurrent consents doctrine. On the one hand, in *Re W*, Lord Donaldson holds that "the inherent powers of the court under its parens patriae jurisdiction are theoretically limitless. . . . There can therefore be no doubt that it has power to override the refusal of a minor," which would suggest that the concurrent consents doctrine is applicable to all refusals of treatment (81). On the other hand, Lord Donaldson himself states that "prudence does not involve avoiding all risk, but it does involve avoiding taking risks which, if they eventuate, may have irreparable consequences or which are disproportionate to the benefits which could accrue from taking them," which suggests restriction of the (practical) scope of the concurrent consents doctrine to medical treatment decisions with potentially serious consequences (81–82). This interpretation aligns with the dictum of Balcombe who holds that the court's override will operate when the child risks death or "severe permanent injury" (88). Nolan also holds that the court has a duty where the child runs the risk of death or "grave and irreversible mental or physical harm" (94). Since we focus on medical treatment decisions with potentially serious consequences, it is not necessary to engage further with the issue of the scope of the doctrine.

2. THE PARITY ARGUMENT FOR ASYMMETRICAL TRANSITIONAL PATERNALISM

Neil Manson defends the concurrent consents doctrine by appeal to what he terms transitional paternalism. On this account, the "normative power to permit treatment is shared between the adolescent and other parties (parents and courts)."[13] Accordingly, it is possible for one party to authorize treatment even when another party with whom the power is shared validly refuses it.[14] This distribution of normative powers is paternalistic because it involves one party possessing the power to consent to another's treatment against the latter's expressed wishes, for her benefit. The paternalism is transitional insofar as adolescents gain, once competent, normative powers that are shared until they become adults.

What justifies transitional paternalism (generally)? Manson appeals to a parity argument:

P1. If we justifiably accept "paternalistic restrictions for adolescents … in areas where any harm is unlikely to be fatal … we should not reject paternalistic restrictions in cases where the risk of serious harm to the adolescent is clear and imminent."
P2. We are justified in accepting paternalistic restrictions in areas in which harm is unlikely to be fatal.
C. Therefore, we ought not reject "paternalistic restrictions in cases where the risk of serious harm to the adolescent is clear and imminent."[15]

This argument does not alone justify the consent/refusal asymmetry in adolescent decision-making about medical treatment. This is because transitional paternalism can be instantiated in different ways.

Manson distinguishes between two forms of transitional paternalism:

> *Restricted-Scope Version*: In some domains, the adolescent has the power to consent to and to refuse treatment; in other domains, to do neither. In some domains, the adolescent is treated like an adult, and in some, like a child. On the restricted-scope view, it might be that in respect of decisions about serious medical treatment, neither consent nor refusal has power.

13 Manson, "Transitional Paternalism," 70. Manson uses the example of a joint bank account to illustrate an asymmetrical distribution of normative powers. On the terms of the arrangement, each account holder possesses the power to consent to certain transactions, even in the face of a valid objection by another account holder (69). Of course, the asymmetrical sharing of normative powers in this context is justified by the agreement between the account holders and the bank.

14 Manson, "Transitional Paternalism," 70.

15 Manson, "Transitional Paternalism," 71–72.

> *Constrained-Power Version*: Consent and refusal have normative power in all domains, but refusals are constrained by the consents of others in certain domains (for example, those in which serious harm might ensue).[16]

Only the constrained-power account instantiates the asymmetry between adolescent consent and refusal. Manson argues that we should prefer the constrained-power version of transitional paternalism to the restricted-scope version by invoking Suzanne Uniacke's distinction between compliance respect and consideration respect.[17] On the constrained-power version of transitional paternalism, adolescents' autonomous wishes are in every case at least considered (given *consideration respect*); in every case in which an adolescent consents, her consent is complied with (given *compliance respect*). Whereas on the restricted-scope form of transitional paternalism, at least in the cases in which we take an interest, adolescents' wishes may only ever receive consideration respect:

	Restricted-Scope View	Constrained-Power View
P consent to ϕ	Consideration respect	Compliance respect
P refusal of ϕ	Consideration respect	Consideration respect

Compliance respect is a more robust form of respect for autonomy.[18] Therefore, the constrained-power account of transitional paternalism offers a "higher grade of respect for the adolescent as an independent decision-maker" than the restricted-scope version.[19]

We have two worries about Manson's argument. First, Manson favors the constrained-power version of transitional paternalism over the restricted-scope version on the grounds that the former involves greater respect for adolescent autonomy. However, Manson does not justify the claim that more autonomy is better for adolescents. Without a justification for this claim, Manson lacks support for his position that the constrained-power view is superior to the restricted-scope view, which in turn is necessary to support the asymmetry in adolescent consents and refusals. In addition, an adequate justification of the asymmetry should provide an account of how autonomy relates to other, competing values, including those Manson thinks warrant constraining autonomy in the case of adolescent refusals.

Second, we have a worry about the first premise of Manson's parity argument. The reasons underpinning restrictions of autonomy in the case of nonfatal harm

16 Manson, "Transitional Paternalism," 72.

17 Uniacke, "Respect for Autonomy in Medical Ethics."

18 Manson, "Transitional Paternalism," 72.

19 Manson, "Transitional Paternalism," 72.

may not carry over to the restrictions of autonomy in the case of serious harm. For example, paternalism in respect of smoking, alcohol, some drug use, or seat belts is likely attributable to the fact that these activities involve weakness of will or irrationality. However, such a justification does not seem to work for refusals of medical treatment, especially when motivated by robust religious or moral views. We can see this in the case of adults, for whom there are paternalistic restrictions on various everyday activities but no such limitations for medical treatment decisions, including those involving potentially serious harm.

These criticisms impugn only Manson's version of transitional paternalism. The asymmetry between adolescent consents and refusals might be justified by a different account of the constrained-power version of transitional paternalism.

3. The Fundamental-Interests Argument for Asymmetrical Transitional Paternalism

Faye Tucker attempts an alternative defense of constrained-power transitional paternalism.[20] She offers the following argument:

P1. Children, including adolescents, have a set of fundamental interests, including in the development of self-governance and faring well.
P2. Adults have an obligation to advance these interests.
P3. The application of transitional paternalism best advances these interests in the medical setting.
C. Therefore, transitional paternalism is justified in the medical setting.

Tucker's defense of transitional paternalism relies on Tamar Schapiro's justification of paternalism toward children, including adolescents.[21] On Schapiro's view, an individual's beliefs and actions are attributable to her when she is self-governing. An individual is self-governing when she has a will, and she has a will when she possesses the capacity to assess her perceptions and motivational impulses (nature's authority) critically and to determine for herself what to do and believe. According to Schapiro, children's beliefs and actions are not attributable to them; they are determined (at least in part) by nature. Children lack the ability to stand back from their motivational impulses and perceptions to determine rationally and freely how to behave and what to think.[22] Children in this sense lack a will. Children are therefore not self-governing and not (fully) responsible for their actions and beliefs. Paternalism is then permissible: for

20 Tucker, "Developing Autonomy and Transitional Paternalism."

21 Schapiro, "What Is a Child?" and "Childhood and Personhood."

22 Schapiro, "Children and Personhood," 590–91.

paternalism is problematic only when it disregards another person's will or another's authority.[23]

Schapiro argues that we have both an obligation to assist children in becoming self-governing and an obligation of beneficence. Tucker thinks that the constrained-power version of transitional paternalism is the most suitable way of discharging these obligations in a clinical context, since it best cultivates an adolescent's capacity for self-governance while safeguarding her well-being.[24]

Tucker's view is vulnerable to three objections. The first objection is that it is unclear whether the constrained-power version of transitional paternalism better facilitates the interest in self-governance than the restricted-scope version. We might plausibly cultivate self-governance (and protect well-being) through the restricted-scope view.

Indeed, Schapiro suggests an account of this sort. On her view, as children enter adolescence they gain "adult status with respect to some domains of discretion, but not others."[25] The acquisition of discretion is based not only on whether the actions or beliefs in the relevant domain were attributable to adolescents but also on whether those adolescents could perform the relevant tasks proficiently.[26] Granting adolescents discretion in any one domain assists them in developing principled stances that might extend their authority to new domains.[27] The expansion of domains of discretion as the adolescent matures is a plausible route through which to arrive at full self-governance, because it involves developing a set of principles that eventually will extend to all domains. But this leaves open the possibility that the restricted-scope version of transitional paternalism better facilitates the development of self-governance.

It is not clear, therefore, that Tucker is able to construct a good defense of the constrained-power version of transitional paternalism based on Schapiro's view alone. And none of the reasons she gives for thinking otherwise are compelling. First, Tucker suggests that the restricted-scope account is less good at facilitating self-governance than the constrained-power account because "only the asymmetric sharing of normative powers enables young people to be involved in a

23 It is not clear that this is paternalism, because paternalism at least on some readings involves overriding the authentic rational ends of another individual.

24 Tucker, "Developing Autonomy and Transitional Paternalism," 762.

25 Schapiro, "What Is a Child?" 734.

26 Schapiro, "Children and Personhood," 591. The proficiency concern is there because on Schapiro's view, the allocation of discretion to adolescents is to be done responsibly.

27 We think this is a plausible rendition of Schapiro's view, but we are not certain of its accuracy. Schapiro changes her mind about the nature of the reasons for granting domains of discretion to adolescents. For an account of the changes, see Schapiro, "Children and Personhood," 591.

set of important decisions from which they would otherwise be excluded, and participation of this sort is *central* to the cultivation of their self-governance."[28]

In reply, one might contend that even if the adolescent's views are not normatively definitive in restricted domains of discretion, participation in the decision through consultation may take seriously the duty to promote self-governance, in addition to the duty of beneficence on the part of involved adults.[29] Participation or involvement in a decision in which one does not have the final say may nevertheless form the basis for the development of "provisional principles of deliberation."[30]

Tucker's point here relies on an inaccurate rendering of the restricted-scope version of transitional paternalism. It does not follow from the fact that an adolescent's views are not normatively determinative that her views would not bear on the decision-making process at all. Even Manson grants that the restricted-scope view affords consideration respect and therefore at least some—possibly quite robust—involvement in important decisions.[31]

Second, Tucker's erroneous characterization of the restricted-scope version of transitional paternalism infects another reason she offers for thinking the restricted-scope view is less good at facilitating self-governance than the constrained-power view. Tucker argues that only the asymmetrical version of transitional paternalism—that is, the constrained-power account—gives "consideration to young people's voices in respect of *all* clinical actions."[32]

The problem here is that Tucker frames the restricted-scope account as entailing that decisions are made on behalf of adolescents without their involvement. However, the restricted-scope account is able to accommodate a duty to consider young people's voices in all clinical actions—that is, by decision makers giving minors' wishes space in the deliberation about what, all things considered, is in their best interests. Indeed, a duty of this kind appears to exist in law and professional guidance for all minors.[33]

Third, Tucker claims that the restricted-scope account is less effective at facilitating self-governance than the constrained-power account because the latter is consistent with "the kind of social arrangements that best support autono-

28 Tucker, "Developing Autonomy and Transitional Paternalism," 765, emphasis in original.

29 See, e.g., *Re W*, 84; Re P (Medical Treatment: Best Interests), [2003] EWHC 2327 (Fam); General Medical Council, *0–18 Years*.

30 Schapiro, "What Is a Child?" 736.

31 Manson, "Transitional Paternalism," 72.

32 Tucker, "Developing Autonomy and Transitional Paternalism," 765, emphasis in original.

33 See, e.g., Re X (A Child) (Capacity to Consent to Termination), [2014] EWHC 1871 (Fam); and General Medical Council, *0–18 Years*.

my."[34] However, this amounts to no more than the rather weak claim that the constrained-power account does not conflict with such social arrangements. Rival views, including other forms of transitional paternalism, might also be consistent with such social arrangements. Whether this is the case will depend on what count as the social arrangements that support autonomy and on various empirical claims about what supports these institutions. Tucker refers to the social arrangements that foster the skills and attitudes associated with autonomy and the development of a deliberative perspective.[35] However, it is, for example, not evident that the restricted-scope account is inconsistent with the social arrangements that best support these kinds of skills and attitudes. Indeed, the sort of reasoning employed in the restricted-scope context by other rational agents—who, for Schapiro, might serve as good models for adolescents insofar as they are self-governing and insofar as they exercise their authority over children responsibly—might facilitate equivalent or greater self-governance in adolescents.[36] We now turn to the second and third objections that Tucker faces.

On the second objection, if Tucker wishes to base her transitional paternalism in part on the fact that the adolescent's will is insufficiently developed, it will be hard for her to justify the asymmetrical treatment of adolescent consent and refusal. If refusal is not always capable of rendering treatment impermissible because an adolescent lacks a fully developed will, why does the same deficiency in the will not cast doubt on consent? Schapiro seems not to allow for adult-like respect with regard to consent but childlike respect with regard to refusal in the same domain. Indeed, Tucker seems to admit this; she writes that "Schapiro argues her lack of reason means the child is unable to make her own choices, *whether good or bad*."[37] If one has authority with respect to a domain, one's deliberative perspective is for that domain authoritative, for the deliberative perspective involves a settled set of values or principles undergirding the decision. One is then authoritative in both one's deciding to do and deciding not to do something. This follows even if facts about proficiency are ultimately relied on to allocate a domain of discretion to an adolescent. Proficiency tests determine whether an adolescent is able to perform the relevant task competently, not whether that decision is attributable to her. So Tucker cannot rely on Schapiro's view to justify the asymmetry in consent and refusal for which she advocates.

The third objection to Tucker's position is that children plausibly have fundamental interests beyond that of becoming self-governing. In addition to the latter,

34 Tucker, "Developing Autonomy and Transitional Paternalism," 765.

35 Tucker, "Developing Autonomy and Transitional Paternalism," 762–63.

36 Schapiro, "Children and Personhood," 592–93, and "What Is a Child?" 734–37.

37 Tucker, "Developing Autonomy and Transitional Paternalism," 761–62, emphasis in original.

Tucker mentions welfare interests and other fundamental interests.[38] Children have a range of interests including a great range of prudential interests, the existence of which might justify limiting or permitting certain decisions that adolescents might make. Until we hear more from Tucker about what these interests are, we will lack insight into the way in which they might constrain cultivating the interest in self-governance. Tucker's version of transitional paternalism does not, then, improve on Manson's account of the view.

To this point, we have addressed views that explicitly seek to justify a constrained-power (asymmetrical) version of transitional paternalism in respect of serious medical treatment decisions involving adolescents. We now turn to two general views that might justify paternalism toward adolescents. Our aim is to determine the extent to which these might justify asymmetry in the respective power of adolescent consent to and refusal of medical treatment.

4. A STAGE-OF-LIFE DEFENSE OF CONCURRENT CONSENTS

Andrew Franklin-Hall attempts to justify paternalism with respect to adolescents in the domain of education.[39]

According to Franklin-Hall, adults have a right to autonomy, entailing a duty to respect their practical authority in deciding what to do. In order to justify equal standing among adults in this regard, the basis of this standing (a threshold degree of rationality or agency) cannot be too robust, for then adults would not possess equal autonomy rights and their practical authority would not be accorded equal respect. However, if the threshold for rationality or agency is set at a level that grounds the equal status of (most) adults, it would seem to ground similar status in adolescents, for typically they possess the minimum level of competence required for agency.[40] This generates a duty to respect their autonomy, and therefore their practical authority.

Yet, in education, adolescent autonomy is routinely restricted to allow adolescents, for their own good, to develop more than the minimum level of rationality or minimum capacity for agency. Here, adolescent autonomy is restricted in order to foster various robust autonomy-related capacities or skills (for example, imaginative reflection) and other character traits or virtues (for example, perseverance and moderation).[41] Franklin-Hall calls the tension between the

38 Tucker, "Developing Autonomy and Transitional Paternalism," 762.

39 Franklin-Hall, "On Becoming an Adult."

40 We take it that the threshold for rationality or agency would not be set so low so as to afford full practical authority for adults who suffer from severe cognitive impairments, etc.

41 It is important to note here that not all the aims of education to which Franklin-Hall al-

duty to respect adolescent autonomy and its restriction to promote other educational goods the "dilemma of liberal education."[42]

How is it possible to justify both respect for the autonomy of adults and paternalism toward adolescents in education? Franklin-Hall offers a stage-of-life solution to the dilemma. He argues that paternalism toward adolescents in education is justified because it takes place at the stage before which an individual has taken up full responsibility for her life. At this stage, an adolescent's values are provisional, and therefore they do not provide a stable and settled basis for her practical identity.[43] Such paternalism has a preparatory aim: it is "oriented toward preparing a person for full practical authority in adulthood."[44] It is conducted with the explicit and public understanding that the adolescent will at some point in the future assume full responsibility for her life.[45] Paternalism toward adolescents can be seen, then, as a normal—and temporary—part of an autonomous life, and so consistent with living a complete one. Finally, paternalism in education, according to Franklin-Hall, does not *interfere with* adolescents living their own lives according to their values; rather, paternalism *delays* the exercise of autonomy.[46]

Important to Franklin-Hall's story is the distinction between global and local autonomy. Global autonomy refers to life authorship, the power to determine one's "roles, projects, values, styles of living."[47] Local autonomy refers to an individual determining (or having the capacity to determine) what to do in a particular case at a particular time. Global autonomy is the more important of the two. Franklin-Hall argues that paternalistic limitations on autonomy in education relate only to local autonomy. The interventions interrupt local autonomy but merely delay the onset of global autonomy. This is consistent with living a "complete autonomous life."[48]

Franklin-Hall's solution to the dilemma of liberal education might be used to support paternalism toward adolescents in the medical setting, and, in particular, the asymmetrical authority of adolescent consents and refusals. The paternalis-

---

ludes are paternalistic. Some of the aims of education limit autonomy but for other than paternalistic reasons. It is unlikely, for example, that educating adolescents so as to foster "open-minded dialogue," "care," "toleration," and "mutual respect" is justified on paternalistic, rather than on moral, grounds (Franklin-Hall, "On Becoming an Adult," 234).

42 Franklin-Hall, "On Becoming an Adult," 235.

43 Franklin-Hall, "On Becoming an Adult," 229.

44 Franklin-Hall, "On Becoming an Adult," 240.

45 Franklin-Hall, "On Becoming an Adult," 239–40.

46 Franklin-Hall, "On Becoming an Adult," 239.

47 Franklin-Hall, "On Becoming an Adult," 237.

48 Franklin-Hall, "On Becoming an Adult," 241.

tic limitation of adolescent autonomy, despite the possession of the minimum capacity necessary for agency, occurs before an individual has assumed full responsibility for her life. At this stage of life, an adolescent's values are provisional—that is, the principles on which she acts do not constitute a stable and settled basis for her practical identity. The paternalism is developmental and temporary; its role is to prepare adolescents for the assumption of full practical authority in making medical decisions.[49] This provides a reason for giving adolescents some control over what happens to them in the medical setting. Moreover, paternalistic restrictions are instituted with the explicit and public understanding that the adolescent will at some point in the future assume full responsibility for her life. Paternalism toward adolescents in medicine can be seen, then, as a normal part of an autonomous life, and so consistent with living a complete one.

The power to give legally effective consent may play a role in preparing adolescents for the assumption of full practical authority, and may be useful from the point of view of developing a full inventory of capacities associated with autonomous choice. These and other preparatory reasons might warrant giving adolescents the authority to consent to treatment and permitting some role for refusals. But limitations on treatment refusal may be justified—in virtue of, in part, the provisional nature of an adolescent's values—in order to protect the adolescent from the full force of action on her principles. Finally, the limitation on refusal is imposed during the stage before control is important to shaping or authoring one's life. To wit, the paternalistic restriction on refusal interrupts local autonomy but merely delays the onset of global autonomy. The assumption is that the choice to determine whether to undergo medical treatment will be an adolescent's in the future—on passing the age of majority.

There is a potential complication with our attempt to extend Franklin-Hall's stage-of-life account to adolescent medical treatment. Franklin-Hall notes the existence of "forced, momentous" choices—choices both life shaping and in-

49 This claim is consistent with the dictum of Lord Donaldson in *Re W*:

> Adolescence is a period of progressive transition from childhood to adulthood and as experience of life is acquired and intelligence and understanding grow, so will the scope of the decision-making which should be left to the minor, for it is only by making decisions and experiencing the consequences that decision-making skills will be acquired. . . . "[G]ood parenting involves giving minors as much rope as they can handle without an unacceptable risk that they will hang themselves." I regard it as self-evident that [the paramountcy of children's welfare] involves giving them the maximum degree of decision-making which is prudent. Prudence does not involve avoiding all risk, but it does involve avoiding taking risks which, if they eventuate, may have irreparable consequences or which are disproportionate to the benefits which could accrue from taking them. (81–82)

capable of adjournment.[50] He argues that adolescents ought to be permitted to make decisions of this sort in some cases—for example, where "there is reason to believe it best for the adolescent to make her own decision," where making the choice for her "would violate her … conscience," or where being prevented from making the choice would unduly restrict the range of options open to her future adult self.[51] Some refusals of treatment might fall within these categories.

Consider a case in which an adolescent validly refuses an abortion. As Franklin-Hall notes, while in this case forcing a teenager to have an abortion might not deprive her of "a self-directed life—it would surely violate her … conscience."[52] It might be right, then, all things considered, to let her decide what to do. If, under the concurrent consents doctrine, another party has the power to consent to abortion, the refusal might nonetheless be honored because it would be wrong to exercise the power.[53] However, here the stage-of-life account buttresses the asymmetrical view, since in this case the refusal is not by itself presumed to be normatively determinative.[54] The refusal is permitted only because there are other factors present that make exercising the power of consent in some way problematic. So even when it is wrong not to let an adolescent decide, it does not follow that it is refusal alone that makes treatment impermissible. Thus the

50 Franklin-Hall, "On Becoming an Adult," 239.

51 Franklin-Hall, "On Becoming an Adult," 239–40.

52 Franklin-Hall, "On Becoming an Adult," 240.

53 In Re X (A Child) (Capacity to Consent to Termination), [2014] EWHC 1871 (Fam), Munby holds in respect of a minor who lacked capacity to consent to an abortion that

> It would not be right to subject X to a termination unless she was both "compliant" and "accepting." … Only the most clear and present risk to the mother's life or long-term health … could justify the use of restraint or physical force to compel compliance. … [M]ere acquiescence—helpless submission in the face of asserted State authority—is not enough. "Consent," of course, is not the appropriate word, for by definition a child of X's age who, like X, lacks Gillick capacity, cannot in law give a valid consent. (12)

If a court would generally not order an abortion unless an adolescent who lacked capacity was "accepting," *a fortiori*, it seems plausible that it would not order an abortion over an adolescent's valid refusal of treatment.

54 It is perhaps possible to argue that in the "forced, momentous" choice case, the power to give concurrent consents disappears. However, this seems inconsistent with the best interpretation of the law. In *Re W*, Lord Donaldson discusses the "hair-raising possibilities … of abortions being carried out by doctors in reliance upon the consent of parents and despite the refusal of consent by 16- and 17-year-olds." His Lordship acknowledges that "this may be possible as a matter of law," which suggests that the power to consent concurrently persists in such cases (79).

asymmetry between consent and refusal remains intact even when the adolescent faces a "forced, momentous" choice.

Having articulated how the stage-of-life account might support the concurrent consents doctrine, we now turn to two objections to relying on the former to justify the latter. First, on Franklin-Hall's view there is an important moral difference between delaying an individual in taking full control of her life and interrupting the control she has over her life. However, it is far from clear that delaying an individual in living a life in accordance with her values is less problematic than interrupting her living her life in accordance with the same. Consider a sixteen-year-old who is steadfast in her religious convictions and, though a hemophiliac, repeatedly refuses blood transfusions.[55] Would interference in this case really be less problematic because it is a case of delay rather than interruption? For the adolescent who is forced to receive treatment and perceives it as a grave insult, this may be of little or no comfort or of little moral significance. Our point is even stronger when we consider cases in which such interferences are liable to reoccur.

Even if we accept that interruption is generally worse than delay, there will still be cases in which paternalism toward adolescents is tantamount to interruption. Imagine our teen is a devout and eager member of a proselytizing religious sect. She has come sincerely to endorse various roles, projects, and so on. Preventing her from making her own choices in accordance with her values seems like an interruption, no different in kind to a similar interruption in an adult. The adolescent could very well claim that this is a case in which interference is inconsistent with being permitted to live a completely autonomous life. In this case, a different kind of justification for paternalism will be required.

Second, we doubt, in any case, that a stage-of-life justification can do the moral work required to permit paternalistic limitations on refusal of treatment. Instead, stage of life seems to be at best an indicator of the variety of considerations that do seem to matter directly to such limitations, including that adolescent values or concerns are in general provisional; that in the cases we consider, acting on these provisional values or concerns has serious consequences; that the limitations are temporary and designed to promote the development of autonomy-related skills; and that adolescent well-being possesses unique features. If we focus on these considerations directly, it may be possible to account for the asymmetry in adolescent consent to and refusal of medical treatment, without reliance on all the machinery employed in Franklin-Hall's view. In addition, it may be possible to provide a justification for paternalism in this form, even when it involves interrupting rather than delaying an adolescent living an autonomous life.

55 For a similar case, see Re E (A Minor) (Wardship: Medical Treatment), [1992] 2 FCR 219 (Fam).

5. A WELFARIST JUSTIFICATION OF CONCURRENT CONSENTS

In this section, we consider how the nature of adolescent prudential value or well-being, the provisional nature of adolescent values (in general), and the risks attached to action on such values may provide an alternative, potentially more promising, justification for the asymmetry in consent and refusal in respect of medical treatment.

There is strong reason to believe that a great measure of what makes an adult's life go well depends on what she wants or what she values.[56] That is, it seems likely that adult well-being depends in large part on what matters from the individual's own subjective point of view. It seems that much less of what makes an adult's life go well is due to the possession of so-called objective goods—things good for an individual regardless of her subjective attitudes toward them, including valuable relationships and intellectual activity—though such goods may be in part what an adult cares about or values. This is no doubt a reason why some find objective accounts of well-being for adults alienating.[57]

By contrast, it is plausible that what is good for a young child lies in part in the possession of objective goods and in part in positive experiences, including happiness and felt satisfaction. A full story about faring well for a young child plausibly involves appeal to both objective goods and positive subjective states.[58] However, much less important to what makes a young child's life go well is getting what she wants or what she values. Succinctly, the subject's point of view or schedule of concerns seems much more important to an adult's well-being than it is to a young child's well-being.

Adolescents occupy a middle position between young children, on the one hand, and typical adults, on the other hand. This is the case not only in respect of how adolescents are treated but also with regard to what might plausibly count as prudentially good for them. Indeed, the differential treatment of adolescents might result at least in some cases from the fact that what is good for them prudentially is distinctive.

We hold that the role the subject's point of view or schedule of concerns plays in an adolescent's well-being lies somewhere between children and adults. This is likely to do with the fact that as the typical human develops, their point of view matures and their schedule of concerns becomes more settled. It seems intuitive

56 For an introduction to the main theories of well-being, see Fletcher, *The Philosophy of Well-Being*; and Sumner, *Welfare, Happiness, and Ethics*.

57 Railton, "Facts and Values"; Rosati, "Internalism and the Good for a Person."

58 For discussion of children's well-being, see Skelton, "Utilitarianism, Welfare, Children," "Children's Well-Being," and "Children and Well-Being."

that adolescent well-being consists at least in part in the adolescent possessing what she subjectively cares about or values. Not just any values or concerns will do, of course. But where the values or concerns are authentic (however specified), they serve as a core feature of adolescent well-being. An adolescent is better off to the extent that her values and concerns are met. These are the subjective elements of adolescent well-being, for how well an adolescent fares depends on the adolescent's schedule of concerns—that is, what matters to her from her own perspective.

Subjective considerations likely do not exhaust what is noninstrumentally good for an adolescent. An adolescent's well-being seems to consist also in the possession of objective goods. Here, an adolescent is made at least somewhat better off to the extent to which she has or possesses these kinds of goods—for example, loving and supportive relationships, knowledge, and achievement. It might be true that more of what is good for an adolescent is determined by her schedule of concerns as she ages; that is, her well-being becomes increasingly based on subjective considerations or the passage of events meeting her expectations or aligning with her values. This is no doubt due to the maturation and development of her point of view or her subjective perspective. But it is intuitive that some constituents of her well-being will remain objective.

The above characterization of the general makeup of adolescent well-being distinguishes it from that of young children, on the one hand, and that of adults, on the other hand. What is distinctive about adolescent well-being might make a difference to our treatment of adolescents. For instance, we think that a clear articulation of the noninstrumental components of adolescent well-being may help to make philosophical sense of the asymmetry of consent and refusal in respect of medical treatment. Important for our purposes are the objective elements of adolescent well-being.

Plausibly, there is a range of objective goods that matter to adolescent well-being. Our focus here is the noninstrumental prudential good of shielding. Shielding consists in being insulated from the full brunt of, the full responsibility for, action on autonomous aims. Shielding is a variety of freedom: freedom from making certain kinds of decisions in the absence of a safety net of scrutiny and possible limitation on action. Shielding is delivered through valuable and supportive, even if not entirely personal, relationships in which adolescents are afforded the insurance of a safety net. So described, the value of shielding connects to the prudential good of valuable relationships.

Franklin-Hall suggests that one virtue of his stage-of-life account is that "it makes available to adolescents a form of freedom much scarcer in adulthood, namely, a measure of freedom from having to make certain decisions with long-

term consequences."[59] When we suggest that one can justify paternalism toward adolescents by appeal to the objective prudential good of shielding, we are expressing the idea that something like this variety of freedom is noninstrumentally prudentially good for an adolescent.

In addition, Franklin-Hall notes that at least some of an adolescent's autonomous aims are provisional. In adolescence, an individual is often attempting to determine her own values; as Franklin-Hall urges, in so doing, she is "toying with possible identities."[60] This form of play can be risky. Because of the risk and the provisional status of some of the values, there is reason for some safeguards—that is, scrutiny and possible limitation on action—even if one does not regard shielding as noninstrumentally good for adolescents: it is instrumentally prudentially good for an adolescent to be shielded from the full force of action on her autonomous, yet provisional, aims. The safety net is there to promote the prudential value for the adolescent of having the responsibility for what happens to her in part outsourced to another (sympathetic and reliable) party.

To recap, on the welfarist view that we are outlining, paternalism toward adolescents is justified in part by the fact that it is prudentially good for an adolescent to be shielded from the full brunt of the consequences of acting on her values; it is prudentially good to have the freedom from making decisions in the absence of a safety net. In addition, adolescent values are provisional in nature and action on them can be risky. It is therefore noninstrumentally and instrumentally good for an adolescent to be treated in some way paternalistically. Incorporating this value into an account of adolescent well-being, as we have done, helps to explain why the stage of life matters: in that stage lie important prudential goods.

The foregoing may justify paternalism toward adolescents in general and in the particular medical circumstances under consideration. But how might it justify the asymmetry between consent to and refusal of medical treatment? Acting on autonomous aims is developmentally important for an adolescent. Being able to exercise autonomous choice at least to some extent is useful from the point of view of preparing an adolescent for the kind of decisions she will have to make on the arrival of adulthood. When an adolescent considers treatment, she (ideally) contemplates whether to consent to or to refuse treatment (and which option to pursue in cases in which more than one intervention is offered). This involves exercising a broad range of skills, including understanding the facts of the situation, applying these to herself, and making a decision based on a sober assessment of what she most values. One might think, therefore, that the rule according to which consents always have the power to render treatment permis-

59 Franklin-Hall, "On Becoming an Adult," 246.

60 Franklin-Hall, "On Becoming an Adult," 229. See also Schapiro, "What Is a Child?" 733.

sible is justified by the fact that it involves promoting instrumentally beneficial exercises of autonomy without the threat of serious costs.[61] This account can explain why the power to consent and have that be normatively determinative is given compliance respect rather than mere consideration respect—namely, because the opportunity to consent allows the instrumental benefits of the exercise of autonomy to accrue to the adolescent to a greater relative degree.[62]

The instrumental benefits of exercising autonomy may also accrue in the case in which an adolescent refuses treatment. This may provide a reason to accord compliance respect to her refusal—that is, for it to be normatively determinative. However, greater reason seems to favor giving refusals mere consideration respect. This is supported by it being prudentially good for an adolescent to be shielded from making the decision without a safety net. In addition, the (sometimes) provisional nature of the values on which an adolescent acts and the fact that action on them may be very costly, especially in the cases we are considering here, provide a further reason not to give refusals full power. These various factors together provide strong reason to protect an adolescent from making such a decision herself. These values seem to provide us, then, with reason to treat refusals differently—that is, as not always capable of rendering treatment impermissible.

The welfarist view articulated above might, then, support an asymmetry between consent to and refusal of medical treatment—namely, the concurrent consents doctrine. The welfarist account is superior to its rivals in a number of respects. It is more comprehensive, specific, and economical. As such, we avoid the objections we have made to the welfarist account's rivals. First, our view provides a compelling reason for why (*pace* Manson) we might not accept that more autonomy is always better for an adolescent. The welfarist account situates autonomy among a wider range of values, and in turn is able to explain why refusals of treatment might not always be normatively determinative. Second, we (unlike Tucker) provide a compelling account of how to justify asymmetrical forms of transitional paternalism; we show how the values on which we draw provide specific support for the doctrine. Third, our view captures the attractive features of the stage-of-life justification of paternalism by reference to prudential values germane to that stage. It does so without reliance on the often very complex machinery found in Franklin-Hall's view. The welfarist view does not, for example, require that we put normative weight (*pace* Franklin-Hall) on the distinction between delaying and interrupting a life lived in accordance with certain values. The welfarist account is therefore more economical than the stage-of-life justifi-

61 There may also be prudential benefits if autonomy is among the prudential goods.

62 Recall that Manson was to his detriment unable to explain why more autonomy was better.

cation. In what remains, we further clarify our welfarist view and consider and reply to some additional potential objections to it.

Let us start with two clarifications. First, it might be inferred from the foregoing that refusals of treatment are alone problematic. We would stress that it need not, of course, be the case that consent is prudentially unproblematic all things considered, whereas refusal poses a prudential threat all things considered—we might envisage cases in which the converse is true. In some situations, refusal may be prudentially unproblematic because it concerns relatively insignificant medical interventions while consent is a prudential threat because it entails quite consequential risks.[63] In this case, it would be consent rather than refusal that should not always be normatively determinative. As such, an asymmetry in normative powers may track high- and low-stakes options.[64]

Second, we have assumed that an adolescent can meaningfully consent to and therefore permit treatment in circumstances in which another party has the power to override a refusal.[65] This assumption and this form of asymmetry in normative powers is a feature of English law (and that of other jurisdictions).[66] Because the concurrent consents doctrine is law, it is important to determine whether it admits of justification. Our idea is that *if* you accept the concurrent consents doctrine and its asymmetrical distribution of the normative powers of consent and refusal, then the most promising defense of this arrangement is provided by the welfarist view. We now turn to objections.

The first objection to the welfarist account focuses on the imposition that shielding involves. The idea that it is prudentially good for an adolescent to be free from making consequential decisions without a safety net has some intuitive plausibility. But for some, this intuitive plausibility may vanish when the

63 For example, we might think that consent to elective or cosmetic interventions carries higher risk of a bad outcome than refusal of the same. Thank you to David Brink for pressing us to clarify this point.

64 In *Re W*, 76 and 83–84, Lord Donaldson expresses the view that the valid consent of a minor of any age could be overridden by the court, but not parents. Interpreted in this way, Lord Scarman's dictum in *Gillick*, 188–89, may leave room for a *concurrent refusal* doctrine, but this, to our knowledge, has never been tested in litigation. If a concurrent refusal doctrine were to exist, this would support the view that the asymmetry in the normative power of consent to and refusal of medical treatment tracks high- and low-stakes options, rather than any essential feature of consent or refusal.

65 For discussion, see, for example, Manson, "Transitional Paternalism"; and Lawlor, "Ambiguities and Asymmetries in Consent and Refusal."

66 Indeed, the legal literature proceeds on this assumption. See, for example, Eekelaar, "White Coats or Flak Jackets?"; Elliston, "If You Know What's Good for You"; Harmon, "Body Blow"; Gilmore and Herring, "'No' Is the Hardest Word"; and Lowe and Juss, "Medical Treatment."

practical realities entailed in promoting or protecting the value of shielding emerge. Consider the adolescent who stubbornly and adamantly wishes to refuse treatment on the basis of her passionately expressed values. She experiences forced treatment as a deep insult, involving great pain and suffering, physical and psychological. This may intensify as the intervention becomes more invasive. These facts make it hard to maintain that it is prudentially good for her not to have the power to refuse treatment.

In reply, one option is to grant that the practical realities of shielding involve the imposition of harm but that the prudential benefits (albeit objective in nature) of being shielded, among other benefits, are worth the cost. The imposition of significant costs through forced treatment on an adolescent is, moreover, not unique to the view that we defend here. In the case of each of the views above, significant burdens will be imposed on the adolescent for her benefit. We seem to have an advantage over those accounts: we can tell the adolescent in what way denying her refusal full normative power is good for her now.

We are open to the idea, however, following Franklin-Hall, that perhaps there are cases in which it is best for an adolescent to have full power over her decisions—that is, full power to consent to or to refuse treatment. Consider two cases. The first involves a recalcitrant teen with anorexia, for whom forced feeding would be experienced by her as a form of tyranny, involving considerable confinement, violation of bodily integrity, suffering, and significant costs on those around her. In this case, we might think it is best for her to make the decision. The second involves an adolescent of First Nations descent living in a country marred by historic injustices toward her peoples, including neglect of their health needs and dismissal of their traditional forms of healing. Against such background injustice, it might all things considered be better to let the adolescent make the decision herself.

The second objection focuses on the general view that we have expressed about the differences between the nature of well-being in adults, adolescents, and young children. To justify the concurrent consents doctrine, we have relied on the idea that adolescent well-being is, in terms of its fundamental, nonderivative prudential constituents, distinct from adult well-being, on the one hand, and young children's well-being, on the other hand. More specifically, we have argued that so-called objective components of well-being are of lesser importance to adults than to adolescents and younger children and that so-called subjective constituents are of lesser significance to adolescents and younger children than to adults.[67]

67 Cf. Cormier and Rossi, "Is Children's Wellbeing Different from Adults' Wellbeing?"; Lin, "Welfare Invariabilism."

This is, of course, not the place to mount a full defense of our view. In reply to the worry, it is possible to recast our account of the prudential value of shielding in a way that makes it less dependent on the nature of well-being varying over the course of an individual's life.[68] There are two options in this regard.

The first option is to maintain that the well-being of children and of adolescents is more subjective in nature than we have suggested—more like what we maintain about adult well-being. In this case, we might hold that well-being consists in desire satisfaction or in life satisfaction or happiness for welfare subjects regardless of stage of life. In so doing, we would deny that there are radical differences in the nature of well-being across classes of welfare subject.

Taking this stance does not require denying that there are (even quite) significant differences in the instruments or causes of well-being across welfare subjects. It is likely that the breadth and depth of one's desires or expectations, not to mention the degree and sophistication of scrutiny that they are able to withstand, is going to be quite different at different stages of maturation or development. These differences are highly likely to occasion a change in the instruments of desire or life satisfaction or happiness.

It is plausible that shielding is one of the instruments of well-being for adolescents, in light of their level of maturation, the (in general) provisional nature of their values, the somewhat unstable nature of their identity, and so on. True, shielding may have some role as a cause of desire or life satisfaction or happiness even for adults. But it is likely that shielding will not have the same degree of influence given (as a class) adults' level of maturity, their stable values, their robust identity, and the value of autonomy to them.

The second option is to hold that the nature of adult well-being is more objective—more like what we maintain about child and adolescent well-being. It may be that well-being consists in the possession of some inventory of objective goods for all welfare subjects. An objective standpoint does not, however, rule out significant differences between the well-being of different welfare subjects. These differences could manifest in at least two different ways.

For one, it is possible that while the nature of well-being is objective, the items on the lists comprising the objective goods will vary across welfare subjects. This will, again, likely depend (at least in part) on the stage of life or development of the welfare subject. There is some reason to think that shielding would not feature on the list of objective prudential goods for adults. Indeed,

68 To be clear, we are not here retreating from our conception of adolescent well-being. Rather, we argue that even if one does not accept our account, shielding has an important role to play in thinking about what is prudentially good for an adolescent.

many of the objective lists for adults include autonomy, but omit shielding.[69] Of course, objective lists for adolescents do not mention shielding either, but that is because no such lists—other than our own—exist. We think, again, that adolescents' level of maturation or development, the provisional nature of (some of) their values, their need for freedom to form their own identity, and so on all make shielding a highly compelling objective good for adolescents.

For another, the objective lists for all welfare subjects might comprise the same items, but the strength of the prudential value of the goods may differ at different stages of life. For example, both autonomy and shielding may be noninstrumentally good for all welfare subjects, but autonomy may matter more (noninstrumentally) to the well-being of adults than to the well-being of adolescents and to that of young children. Likewise, shielding may matter more (noninstrumentally) to the well-being of adolescents and of young children than to that of adults. This would, again, depend in part on level of maturation, stability of values, identity formation, and so on.

It transpires then that reluctance to embrace the idea that fundamental constituents of adolescent well-being are distinct from those of adult and young child well-being, respectively, need not cast doubt on the importance of the prudential value of shielding to adolescent well-being.[70]

A third objection concerns whether our view is able to support the asymmetry between consent to and refusal of treatment. Facts about adolescent well-being may make the case for asymmetry, as suggested above. But it might be unclear whether the welfarist view indeed provides more support for the asymmetrical (constrained-power) version of transitional paternalism as opposed to the restricted-scope version. If reasons related to the prudential value of shielding are sufficient to warrant limiting refusals, these reasons may justify removing decisions about serious medical treatment from adolescents altogether.[71] For

69 Badhwar, *Well-Being*; Fletcher, *The Philosophy of Well-Being*; Griffin, *Well-Being*; Hooker, *Ideal Code, Real World*.

70 Anonymous referees for the journal suggested that we might work out a conception of well-being for children closer in nature to that of adults by reference to Rawlsian primary goods, including income, health, education, opportunity, and so on. This is a plausible suggestion. But even if primary goods play a role in well-being, it is still highly likely that at some level there will be marked differences between children and adults in the constituents or the causes of well-being. As Rawls notes, the content of primary goods depends on "various general facts about human needs and abilities, their normal phases and requirements of nurture, relations of social interdependence, and much else" (Rawls, *Justice as Fairness*, 58). We thank the referees for prompting us to clarify this point and our view in general.

71 This objection is similar to the one we leveled against Tucker's account.

even in cases of consent to treatment, an adolescent has to take responsibility for her decision.

There are two possible replies to this objection. The first involves arguing that in the cases of concern to us here, there is little reason not to grant consent compliance respect. As stipulated, consent pertains, after all, to treatment that is in an adolescent's clinical best interests and carries with it a high probability of success. There is little reason to be shielded or to have insurance against decisions that are in one's best medical interests. In any case, even if there is some reason to shield adolescents from consents, it is much weaker than the reason we might have to shield adolescents in cases in which a refusal emanates from a commitment to provisional values or in which the expected outcome runs contrary to their clinical best interests.

The second reply involves granting that the welfarist view we defend here provides only contingent support for the constrained-power view of transitional paternalism. We maintain that the welfarist view accounts persuasively for the asymmetry in consent and refusal. But it might turn out that the welfarist view provides support in some (legal and social) contexts for the restricted-scope view. We think that this is an attractive feature of the view. Whether the welfarist account in fact supports the constrained-power version over the restricted-scope version turns partly on empirical considerations and partly on facts about the institutional context, including those relating to the legal system.[72] We have told a story about how the asymmetry in consents and refusals might arise. Whether it does arise will most certainly depend on what best promotes the instrumental and noninstrumental prudential goods we discuss above and on other social and legal facts.

6. CONCLUSION

How is it that a competent minor's consent renders medical treatment lawful, yet a competent minor's refusal may not render treatment unlawful? In this article, we attempted to make philosophical sense of the concurrent consents doctrine in law, which posits an asymmetry in the normative power of adolescent consent and refusal.

We examined and rejected three possible justifications for the concurrent consents doctrine, two based on transitional paternalism and one based on stage of life. We developed a more philosophically promising, welfarist justification of the concurrent consents doctrine that takes up relevant considerations iden-

72 The kind of empirical facts that we have in mind include facts about how burdensome shielding turns out to be for individuals or classes of individuals.

tified in these rival views yet avoids their infelicities. This welfarist justification relies on the idea that there are distinct features of adolescent well-being that distinguish it from the well-being of adults, on the one hand, and young children, on the other hand. The main element of adolescent well-being of concern to us is the good of shielding. It is good for adolescents to be shielded from full responsibility for their decisions, and this explains why adolescent consent may be normatively determinative in the cases that we consider, but their refusal in such cases may not.

In this paper, our focus has been the philosophical justification of the concurrent consents doctrine in respect of serious medical treatment. However, in closing, it is important to note that the welfarist account that we defend may justify paternalism or differential treatment of adolescents more generally—that is, in other medical settings and other domains. The welfarist view is therefore a contribution to the literature on the general question of when and how paternalism toward adolescents may be justified philosophically.[73]

University of Western Ontario
askelto4@uwo.ca

University of Oxford
lisa.forsberg@law.ox.ac.uk

University College London
isra.black@ucl.ac.uk

73 We wish to thank David Brink, Simon Halliday, Matt Matravers, Kaitlin Pettit, Jenny Steele, two anonymous referees for the *Journal of Ethics and Social Philosophy*, an anonymous referee for another journal, and audiences at the Annual Congress of the Canadian Philosophical Association, Ryerson University, 2017; the Society for Applied Philosophy Annual Conference, University of Copenhagen, 2017; the Philosophy and Childhood Conference, Centre for Ethics and Poverty Research, University of Salzburg, 2017; the tenth Rocky Mountain Ethics Congress, University of Colorado, Boulder, 2017; Filosofidagarna, Uppsala University, 2017; and Gothenburg University, 2019, for helpful comments on previous versions of this paper. We would like to express our gratitude to the Fondation Brocher for hosting us as visiting researchers in 2015. Lisa Forsberg wishes to acknowledge the support of the postdoctoral fellowship program at the Rotman Institute of Philosophy and a British Academy Postdoctoral Fellowship (award PF170028). Anthony Skelton wishes to acknowledge the support of the Faculty Research Development Fund at the University of Western Ontario (award R3986A05).

REFERENCES

Badhwar, Neera Kapur. *Well-Being: Happiness in a Worthwhile Life*. Oxford: Oxford University Press, 2014.

Bainham, Andrew. "The Judge and the Competent Minor." *Law Quarterly Review* 108 (April 1992): 194–200.

Buchanan, Allen E., and Dan W. Brock. *Deciding for Others: The Ethics of Surrogate Decision Making*. Cambridge: Cambridge University Press, 1989.

Cormier, Andrée-Anne, and Mauro Rossi. "Is Children's Wellbeing Different from Adults' Wellbeing?" *Canadian Journal of Philosophy* 49, no. 8 (2019): 1146–68.

Eekelaar, John. "White Coats or Flak Jackets? Children and the Courts—Again." *Law Quarterly Review* 109 (April 1993): 182–87.

Elliston, Sarah. "If You Know What's Good for You: Refusal of Consent to Medical Treatment by Children." In *Contemporary Issues in Law, Medicine and Ethics*, edited by Sheila A. M. McLean, 29–55. Aldershot: Dartmouth, 1996.

Fletcher, Guy. *The Philosophy of Well-Being: An Introduction*. London: Routledge, 2016.

Franklin-Hall, Andrew. "On Becoming an Adult: Autonomy and the Moral Relevance of Life's Stages." *Philosophical Quarterly* 63, no. 251 (April 2013): 223–47.

General Medical Council. *0–18 Years: Guidance for all Doctors*. 2007. Updated May 25, 2018. https://www.gmc-uk.org/-/media/documents/0_18_years_english_0418pdf_48903188.pdf.

Gilmore, Stephen, and Jonathan Herring. "'No' Is the Hardest Word: Consent and Children's Autonomy." *Child and Family Law Quarterly* 23, no. 1 (2011): 3–25.

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

Harmon, Shawn E. "Body Blow: Mature Minors and the Supreme Court of Canada's Decision in *A. C. v. Manitoba*." *McGill Journal of Law and Health* 4, no. 1 (2010): 83–96.

Hooker, Brad. *Ideal Code, Real World: A Rule-Consequentialist Theory of Morality*. Oxford: Clarendon Press, 2000.

Law Reform Commission (Ireland). *Children and the Law: Medical Treatment*. LRC 103–2011. July 2011. https://www.lawreform.ie/_fileupload/Reports/r103.htm.

Lawlor, Rob. "Ambiguities and Asymmetries in Consent and Refusal: Reply to Manson." *Bioethics* 30, no. 5 (June 2016): 353–57.

Lin, Eden. "Welfare Invariabilism." *Ethics* 128, no. 2 (January 2018): 320–45.

Lowe, Nigel, and Satvinder Juss. "Medical Treatment—Pragmatism and the Search for Principle." *Modern Law Review* 56, no. 6 (November 1993): 865–72.

Manson, Neil C. "Transitional Paternalism: How Shared Normative Powers Give Rise to the Asymmetry of Adolescent Consent and Refusal." *Bioethics* 29, no. 2 (February 2015): 66–73.

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

Rawls, John. *Justice as Fairness: A Restatement*. Edited by Erin Kelly. Cambridge, MA: Harvard University Press, 2001.

Rosati, Connie S. "Internalism and the Good for a Person." *Ethics* 106, no. 2 (January 1996): 297–326.

Schapiro, Tamar. "Childhood and Personhood." *Arizona Law Review* 45, no. 3 (2003): 575–94.

———. "What Is a Child?" *Ethics* 109, no. 4 (July 1999): 715–38.

Skelton, Anthony. "Children and Well-Being." In *The Routledge Handbook of the Philosophy of Childhood and Children*, edited by Anca Gheaus, Gideon Calder, and Jurgen De Wispelaere, 90–100. London: Routledge, 2018.

———. "Children's Well-Being: A Philosophical Analysis." In *The Routledge Handbook of Philosophy of Well-Being*, edited by Guy Fletcher, 366–77. London: Routledge, 2016.

———. "Utilitarianism, Welfare, Children." In *The Nature of Children's Well-Being: Theory and Practice*, edited by Alexander Bagattini and Colin Macleod, 85–103. Dordrecht: Springer, 2015.

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

Tucker, Faye. "Developing Autonomy and Transitional Paternalism." *Bioethics* 30, no. 9 (November 2016): 759–66.

Uniacke, Suzanne. "Respect for Autonomy in Medical Ethics." In *Reading Onora O'Neill*, edited by David Archard, Monique Deveaux, Neil Manson, and Daniel Weinstock, 94–110. London: Routledge, 2013.

Wicclair, Mark R. "Patient Decision-Making Capacity and Risk." *Bioethics* 5, no. 2 (April 1991): 91–104.

Wilks, Ian. "The Debate over Risk-Related Standards of Competence." *Bioethics* 11, no. 5 (October 1997): 413–26.

Journal of Ethics and Social Philosophy
VOL. 20, NO. 3 · NOVEMBER 2021
https://doi.org/10.26556/jesp.v20i3.1127

RELIGIOUS REASONING IN THE LIBERAL PUBLIC FROM THE SECOND-PERSONAL PERSPECTIVE

A DEFENSE OF AN INCLUSIVIST MODEL OF PUBLIC REASON LIBERALISM

Patrick Zoll

ARE CITIZENS OBLIGED to refrain from using religious arguments for the public justification of political norms in a liberal democracy (e.g., a law) if these are the *only* justificatory reasons they have? Is a committed Christian, for example, who has no other means than his religious beliefs to justify his political preferences, obliged to refrain from referring to the Bible or other evaluative standards of his religious conception of a good life to justify his rejection of a law that allows abortion or the use of human embryos for research purposes?

Exclusivists like Robert Audi, Robert B. Talisse, and Jonathan Quong respond to these questions affirmatively, while inclusivists like Christopher J. Eberle, Steven Wall, and Nicholas Wolterstorff answer them negatively.[1] This

1 Cf. Audi, *Religious Commitment and Secular Reason*; Audi and Wolterstorff, *Religion in the Public Square*; Eberle, *Religious Conviction in Liberal Politics*; Quong, *Liberalism without Perfection*; Talisse, *Democracy and Moral Conflict*; Wall, *Liberalism, Perfectionism and Restraint*. One might wonder why I include Wall and Quong in this list, who are not very explicit about religious reasons. The reason is that I regard religious reasons as a kind of perfectionist reasons, and they are leading protagonists in the debate between anti-perfectionist and perfectionist liberals. I assume therefore that the debate about religious arguments in public justification is best understood as part of the more general debate about perfectionist reasons in public justification. A defense of this classificatory claim is provided by Zoll, *Perfektionistischer Liberalismus*. As will become clear in the unfolding of my argument, I am here interested in religious reasons that cannot be translated into a secular language or evaluated based on common evaluative standards because their soundness depends on the acceptance of revealed knowledge or religious authority. I am not claiming that all religious reasons are reasons that are generally inaccessible. Rather, I am arguing that the kind of religious reasons mentioned above are rejected by exclusivists because they are generally inaccessible. Although I come to a different conclusion with respect to the possible role of these reasons in public justification, I find the typology of different religious reasons offered by Andrew March quite helpful; see March, "Rethinking Religious Reasons in Public Justification" 527–30.

debate has a long trajectory in contemporary liberal political philosophy and has led not only to a constant dissent between the parties but also to the impression that a commitment to public reason liberalism necessarily implies a commitment to exclusivism.

In this essay, I will argue that the stability of this dissent is best explained as being rooted in two incompatible conceptions of public justification.[2] The exclusivist's position is grounded in a third-personal account that implies restraint, the inclusivist's in a first-personal account that rejects restraint.[3] After having criticized both accounts as insufficient, I will rely on a second-personal conception of public justification to construct and defend an inclusivist model of public reason liberalism that rejects restraint but is able to do justice to the moral intuition that motivates exclusivism.[4] Finally, I will clarify how my model of acceptable religious discourse differs from other inclusivist variants in the literature by comparing it with two proposals that have been advanced recently.

1. Why Inclusivism Fails So Far: The First-Personal Account of Public Justification

I will begin my argument with the thesis that inclusivists fail so far to convince exclusivists because inclusivism is rooted in a first-personal conception of public

2 This might be a surprise because authors like Wall and Eberle explicitly reject the term "public justification" and speak instead of "political justification." As will become clear later on, I will argue that this rejection is a reaction to the identification of the *concept* of public justification with a particular third-personal *conception* of public justification that implies restraint. Once this identification is questioned, as for example by Vallier, it becomes possible to use the term "public justification" in a broader sense than currently at use in literature; see Vallier, *Liberal Politics and Public Faith: A Philosophical Reconciliation* and *Liberal Politics and Public Faith: Beyond Separation*.

3 I assume here that the debate between political liberals and their religious critics has shown that all attempts to solve the problem by narrowing the scope of the principle of restraint (e.g., to constitutional essentials) and thereby "softening" the demand of restraint have failed. An analysis of three different models of this softening strategy and a defense of the claim that they all fail can be found in Zoll, *Perfektionistischer Liberalismus*, 93–120. For this reason, I identify "inclusivism" with "strong inclusivism" here. I will show why this strong inclusivism is preferable to a "weak inclusivism" by comparing my model of acceptable religious discourse with the models that result from two more recent versions of a weak inclusivism in section 5 of this essay. I owe thanks to an anonymous reviewer of this journal who made me aware of the need to address this issue.

4 I have learned a great deal from the works of Kevin Vallier who, to my awareness, was the first one to present a book-length defense of the possibility to construct an inclusivist public reason liberalism; see Vallier, *Liberal Politics and Public Faith: A Philosophical Reconciliation* and *Liberal Politics and Public Faith: Beyond Separation*.

justification that is not able to accommodate a valid moral intuition that motivates exclusivism.

To get a better grasp of this intuition, it is helpful to analyze an example that the prominent exclusivist Robert Audi gives to support the view that religious citizens should exercise restraint in public justification. Audi asks whether it is plausible that one would be willing to accept a law as legitimate that forbids one to mow one's lawn in one's backyard if this law is only justified on the ground that the dandelion is sacred.[5] He admits that this is not a very realistic case because no one seems to really believe that dandelions are sacred.[6] But its hypothetical and artificial nature does not weaken its force. On the contrary, even religious citizens should be able to acknowledge by this experiment of role reversal what they have to accept if they give up the idea that public justification implies an obligation to exercise restraint.[7]

On the one hand, it would be possible for them to support a coercive law just on the basis of *their* religious reasons, but on the other hand, they would also have to accept that *their* liberty could get restricted just on the basis of *other* religious reasons. But, as the absurd case of the sacred dandelion should make clear to religious citizens, if they do not embrace the possibility of getting forced on the basis of foreign religious reasons that are not accessible to themselves, they should also refrain from trying to force others solely with their own religious arguments that are not accessible to their fellow citizens who do not share their religious convictions.[8]

5 Cf. Audi, *Religious Commitment*, 93.

6 Cf. Audi, *Religious Commitment*, 93–94.

7 It could be objected that perhaps some religious citizens (e.g., fanatics) lack the necessary capacities or willingness for such an exercise, e.g., due to certain character traits, lack of training, or upbringing. Against this I would respond that the example is not a claim about what *actual* religious citizens should be able to acknowledge but a claim about what *appropriately idealized* versions of them should be able to acknowledge. This rules out cases of fanatics or other religious citizens who are either not willing or not able to play the democratic game of giving and asking for reasons. Thanks to an anonymous reviewer for pushing me here.

8 I am aware that Audi's example of the dandelion has a problematic and polemic edge because it seems to imply that all religious beliefs are somehow epistemologically flawed and therefore not apt to serve in public justifications; cf. Eberle, *Religious Conviction in Liberal Politics*, 134–40. I agree with Eberle's critique but maintain that it is possible to reinterpret the argument in the heuristic or hermeneutical way that I do above. The absurd case of the dandelion could help religious citizens precisely because of its absurd character to understand on the one hand how religious beliefs sometimes appear to *nonreligious citizens*. On the other hand, from their *own perspective* they can understand that they do not want to be coerced on the basis of absurd beliefs or beliefs that at least appear to be absurd. My point is therefore that the argument can serve for religious citizens as a heuristic device to get a

The moral intuition seeming to motivate a conception of public justification that implies restraint is that the justification of a demand that infringes the liberty of another person should take into account the epistemic perspective of this person because it is plausible to assume that one would not accept a restriction of one's own liberty by this other person if she justifies this restriction only with reasons one cannot access from one's own epistemic standpoint. I call this moral intuition "second-personal" because it makes plain that the justification of demands cannot be successful if one relies exclusively on one's own "first-personal" standpoint. It is also necessary to reason from the epistemic perspective of the addressed person.[9]

But the acceptance of this second-personal moral intuition as plausible implies the commitment to accept a "principle of moral restraint," as this argument demonstrates:

1. It is plausible to assume that nobody is willing to accept a demand that restricts one's liberty if this demand is solely justified with arguments that one cannot access from one's own epistemic perspective.
2. It is plausible to accept the general moral principle of reciprocity: "One should not treat others in ways that one would not like to be treated."
3. Therefore, if one cannot justify one's demand to another person by showing to that person that she has a weighty reason to comply with the demand, one ought to refrain from advancing and enforcing this demand.[10]

---

better grasp on the moral intuition that motivates nonreligious citizens to support restraint. This heuristic interpretation of the argument does not commit religious citizens to the truth of its problematic epistemological implications.

9 To be more precise: in addition to an individual first-personal perspective a "second first-personal perspective" has to be adopted. The epistemic duties have been fulfilled and a demand is successfully justified if it is possible to show that this demand is conclusively justified from one's own first-personal perspective as well as that there is a weighty reason to comply with the demand from the second first-personal perspective of the person who is addressed by that demand. Thus, my use of the term "second-personal" has certain similarities to the recent attempts to ground morality in the "second-personal standpoint"; cf. Darwall, *The Second-Person Standpoint*; Eilan, *The Second Person*; Pinsent, *The Second-Person Perspective in Aquinas's Ethics*. Moreover, it is important to recognize that the adoption of a second first-personal perspective that is incompatible with one's own first-personal perspective and the fact that a demand is justified from two mutually incompatible epistemic standpoints results in commitment to a kind of epistemic contextualism but not a relativism about truth. A good illustration of what it means to adopt a "second first-personal perspective" and a defense of this conception of justification against the charge of relativism can be found in MacIntyre, *Whose Justice? Which Rationality?* 349–88.

10 Alternative ways to argue for this principle can be found in Gaus, *Justificatory Liberalism*, 123–41; and Wall, *Liberalism, Perfectionism and Restraint*, 115–18.

Exclusivists therefore argue, in a further step, that a commitment to the principle of moral restraint expressed in 3 obliges everyone to adopt a "third-personal" standpoint when political norms have to be publicly justified because these norms are nothing else than liberty-restricting demands addressing all appropriately idealized citizens of a liberal democracy:[11]

4. Third-personal conception of public justification:
   a. The imposition of a political norm through a democratic procedure is only legitimate iff it is publicly justified.
   b. A political norm is publicly justified iff it is justified with reasons that are mutually accessible to the appropriately idealized members of the public.

The requirement of mutual accessibility 4b for public reasons *seems* to be a logical consequence if the premises 1 and 2 are accepted.[12] Moreover, mutual accessibility means that only those reasons can enter the process of public justification whose justificatory relevance for justifying a political norm can be recognized according to *common* evaluative standards.[13]

Therefore, exclusivists claim that the acceptance of moral restraint makes it necessary to adopt a third-personal conception of public justification and that this conception implies an obligation to exercise epistemic restraint:

5. Citizens are obliged to refrain from justifying their favored political norms with reasons that are not accessible to all appropriately idealized members of the public because the justificatory relevance of these

11 Exclusivists and inclusivists generally agree that the conception of the members of the public must be "idealized." Idealization in some form is necessary because otherwise the success of a public justification would be determined by the "actual" members of the public and arguments could be rejected on the basis of poor information, inferential mistakes, or incoherent beliefs; see Billingham, "Convergence Justifications with Political Liberalism," 137. What is controversial is how radical this idealization has to be. I believe that a "moderate idealization" is all that is demanded by a reasonable account of public justification. For a presentation and defense of this claim, see Gaus, *The Order of Public Reason*, 232–60; Vallier, *Liberal Politics and Public Faith: Beyond Separation*, 145–80. But because of the controversial character of this issue, I use the neutral expression "appropriately idealized," leaving it open what idealization exactly involves. Thanks to an anonymous reviewer for making me aware of the need to address this issue.

12 As I will later show, it constitutes a fallacy to infer from the need to justify to each citizen a need to justify using mutually accessible reasons.

13 My definition of "accessibility" is partly inspired by Vallier, *Liberal Politics and Public Faith: Beyond Separation*, 108–9. An overview of other interpretations of this concept is given by Eberle, *Religious Conviction in Liberal Politics*, 252–86.

> reasons for justifying a political norm cannot be recognized according to the common evaluative standards of a third-personal perspective.[14]

According to this line of reasoning, religious arguments cannot figure into a justification of political norms because they cannot be evaluated as justificatorily relevant from the common or "third-person" perspective that public justification demands. For example, David's biblical argument for supporting a political norm that prohibits abortion cannot play a role in public justification because atheistic Beatrice can object that she does not share David's religious evaluative standards. But without the acceptance of these evaluative standards she cannot recognize that David's biblical considerations constitute a reason *for her* that justifies abortion.

Therefore, epistemic constraint has the function to safeguard the exercise of moral restraint in politics by tying political coercion to public justification. This connection is necessary to rule out that a religious majority can impose a political norm on a religious minority solely relying on religious beliefs that the minority does not share or even rejects.[15]

We can easily see why it is problematic if public justification does not embrace moral restraint: e.g., if Adam, who is Muslim, demands from atheistic Beatrice not to wear bikinis in public swimming pools and "justifies" this demand to her only with the argument that her bikini is not appropriate according to the evaluative standards of *his* religious tradition, he is in reality not "justifying" his demand to *her*.[16] In the best case, he insists that she should believe what he wants; in the worst case, he is just browbeating her.[17] However, he misses an opportunity to address Beatrice and her epistemic standpoint in his justification. From his point of view the demand is justified but not from Beatrice's point of view. However, *public* justification needs to be bi-relational.

Thus, the lasting unwillingness of exclusivists to grant inclusivists the possibility to justify their support or rejection of political norms *solely* with religious

14 It could be objected that proposition 5 is not accepted by all exclusivists because some of them—like Cécile Laborde, Aurélia Bardon, and Will Kymlicka—content themselves to demand that just public officials, not all citizens, are obliged to exercise epistemic restraint. Thus, the strategy to avoid the problems associated with 5 consists in a limitation of the scope or application of the obligation to exercise epistemic restraint. Here I assume—as already mentioned in note 3 above—that these and similar exclusivist strategies of limitation fail for the reasons given in Zoll, *Perfektionistischer Liberalismus*, 93–120.

15 This concern is often reiterated and seems to be an important motivation for defending exclusivism; see for example Audi, *Religious Commitment*, 201; Breul, *Religion in der politischen Öffentlichkeit*, 194.

16 The example is inspired by Audi, *Religious Commitment*, 93.

17 This distinction between a moral demand and mere browbeating is taken from Gaus, *Justificatory Liberalism*, 123–29.

arguments can best be explained as rooted in the rejection of a mono-relational conception of public justification that does not do justice to moral restraint and the bi-relational character of public justification.

On this background, I will now analyze in a second step the work of Steven Wall, which is representative of a highly influential "first-personal model" of inclusivism, to give evidence for my claim that inclusivists fail so far in convincing exclusivists because their first-personal conception of public justification is not able to accommodate the second-personal moral intuition that motivates exclusivism.

Though there is a variety of first-personal models of inclusivism, common to all of them is that they try to abandon epistemic restraint by substituting the third-personal conception of public justification with a first-personal account.[18] With this different conception they want to demonstrate that there is a "gap" between public justification and epistemic restraint, i.e., that a commitment to public justification does not—in contrast to the exclusivists' claim—imply necessarily a commitment to epistemic restraint.

4*. First-personal conception of public justification:
  a. The imposition of a political norm through a democratic procedure is only legitimate iff it is publicly justified.
  b. A political norm is publicly justified iff the group of citizens *C* who want to impose the norm
    b1. give a sincere and honest justification of it, i.e., they state publicly and in a sincere and honest way the considerations that motivate them to support the imposition of this norm, and
    b2. it is intelligible for the appropriately idealized members of the public on which the norm is imposed that these considerations constitute a weighty reason for *C* that justifies the imposition.[19]

According to 4*, a commitment to public justification implies only a *weak* kind of moral restraint. Religious citizens should refrain from advancing and imposing their preferred political norms if they cannot justify them sincerely, honestly, and intelligibly. Intelligibility requires that the arguments religious citizens use for the justification of political norms be formulated in such a way that appropriately idealized nonreligious citizens should be able to track the soundness of the

18 Cf. Eberle, *Religious Conviction in Liberal Politics*, 10, 109–51, 331–33; Wall, *Liberalism, Perfectionism and Restraint*, 79–82, 115–51.

19 Cf. Wall, *Liberalism, Perfectionism and Restraint*, 108, esp. n9 and n12. Wall does not mention "intelligibility" but condition b2 is a fair interpretation of what Wall means by "subjective" and "objective" justification.

argumentation if they adopt the epistemic standpoint of their fellow religious citizens. Intelligibility does not require that they have to accept the truth of the religious presuppositions—for example, revealed truths—but only that they should be able to acknowledge that these arguments are sound for someone who does accept their truth.

Here, Wall states that this weak kind of moral restraint implied in 4* does not oblige inclusivists to exercise epistemic restraint and to refrain from enforcing their favored political norms on others if they cannot justify them with reasons that the addressed can access from their epistemic standpoint, because 4* does not commit someone to the acceptance of *strong* moral restraint in the sense of 3.

According to Wall, this is the case because the public justification of political norms and the justification of demands to others in general need not be "relational" but only "simple."[20] In contrast to a relational conception of (public) justification, the "simple" conception holds that the epistemic perspective of the persons at which a demand is directed is irrelevant for determining whether this demand is successfully justified.

The first-personal conception of public justification states instead that a political norm imposed on Beatrice is successfully justified if Adam has presented reasons that are sound and of sufficient weight to override competing reasons against the political norm from *his* epistemic first-personal perspective and that this fact is intelligible to Beatrice.[21] Therefore, the epistemic perspective of Beatrice plays no role in this first-personal account of public justification and the fact that Beatrice has no reason to comply with Adam's demand imposes no further restraint on him.

In summary, Wall abandons epistemic restraint by attacking the third-personal conception of public justification. And he attacks the principle of moral restraint by preferring a "simple" to a "relational" conception of justification that allows him to offer a mono-relational first-personal conception of public justification that unties public justification from any restraint that derives from the idea that the justification of demands to another person should take the epistemic perspective of this person into account. Consequently, Wall's argument is grounded in the claim that the exclusivist's bi-relational third-personal conception of public justification can be substituted by the mono-relational first-personal conception without thereby disconnecting the exercise of political power from public justification in any problematic way.

After reconstructing the argumentative core of this first-personal model of inclusivism, I will now argue that it has no chance of convincing exclusivists. They can correctly object that the first-personal conception of public justifica-

20 Cf. Wall, "Perfectionism in Politics," 112.

21 Cf. Wall, "Perfectionism in Politics."

tion cannot substitute their third-personal account because the tie it establishes between political coercion and public justification is too loose. Consequently, it is not public justification but a democratic decision procedure that determines ultimately whether the imposition of political coercion is legitimate or not. If the epistemic perspective of those who are addressed by a demand becomes irrelevant by definition, then the only reason left for them to accept the imposition of a political norm they reject is that the imposition of this demand is the outcome of a procedure they accept. But this means that it is not public justification anymore that legitimizes the use of political power to them, but the democratic procedure of decision-making.

Wall himself gives evidence that this objection is well grounded because he regards cases as unproblematic that exclusivists mark as highly problematic.[22] Thus, Wall's account confirms what exclusivists fear most: inclusivism leads to a legitimatization of cases where a majority can impose political norms on a minority by a democratic procedure with a "simple justification" that gives the addressed minority no reason to comply with this norm.[23] For exclusivists, these cases are highly problematic because the exact difference between "public justification from the first-personal perspective" and "political browbeating" or the arbitrary use of political power cannot be distinguished.[24]

If the first-personal model of inclusivism is adopted, minorities lack any normative resources to criticize the exercise of political power that matches democratic procedures.[25] Their epistemic standpoints and the normative resources

22 Cf. Wall, *Liberalism, Perfectionism and Restraint*, 79–82, 115–23.

23 Against this it could be objected that this is not true and represents a misdescription of Wall's position because he holds that his conception of simple justification demands that a political norm is not only subjectively justified from the perspective of the majority but also objectively justified, which means that the political norm is justified in accordance with right reason; see Wall, *Liberalism, Perfectionism and Restraint*, 102. Thus, it is wrong that the minority are given no reasons. They are given "true" or "right" reasons by the majority, which are reasons the minority ought to accept even if they cannot accept these reasons as reasons from their epistemic standpoint. I think this objection fails for two reasons. First of all, it has deeply problematic paternalistic consequences. Second, the objection presupposes the acceptance of a very strong and implausible externalism about reasons; see Wall, "Perfectionism in Politics," 109–11. But Wall's argument for such an externalism is not convincing, as shown by Zoll, *Perfektionistischer Liberalismus*, 214–25. I would like to thank an anonymous reviewer for making me aware that this objection needs to be addressed.

24 Wall partly concedes this point; see esp. Wall, "Perfectionism in Politics," 112.

25 Against this and what follows it could be objected that it applies to all democratic theories that do not include a public reason requirement. But democracies guarantee their citizens a range of constitutionally protected basic rights and incorporate their epistemic perspectives by giving all citizens equal voice and vote. Therefore, the claim that minorities are left to the complete mercy of majorities and that their epistemic perspectives are not sufficiently

that derive from them are irrelevant by definition because of the mono-relational character of public justification. Thus, the only option left for them to criticize the imposition of a political norm through a majority vote is to reason from the epistemic perspective of the powerful majority and to show that they are not justified in imposing this norm on them through democratic decision-making.

Yet, I think it is absurd—if not cynical—if one wants to sell this as a serious possibility for minorities to criticize power. Furthermore, it is quite implausible that this option for social criticism constitutes an effective mechanism to protect minorities from the abuse of political power. Thus, the substitution of the third-personal account of public justification with a first-personal account is not acceptable for exclusivists because it reduces public justification to a mono-relational enterprise with the consequence that the epistemic perspectives of minorities are systematically excluded from the process of the public justification of political norms.

For the first-personal model of inclusivism, the fact is even worse in that it does not fail just by the external standard of exclusivism but also by its own standards. This is the case because—as we have seen above—this model is only able to reject epistemic restraint if religious citizens are willing to reject moral restraint. The principle of moral restraint has to be rejected because it presupposes a relational understanding of justification that is incompatible with a simple conception of justification on which the first-personal conception of justification rests. Yet, religious citizens have a weighty reason to *accept* moral restraint.

The argument runs like this: it is constitutive for liberal and democratic societies that their appropriately idealized citizens accept a presumption in favor of liberty:

6. Citizens possess a moral status that obliges other persons to treat them as persons who are entitled and able to choose and lead a life according to the evaluative standards of their conception of a good life.[26]

taken into account without a public reason requirement is false. I would reply to this that my point is not that minorities are without any protection, etc., without a public reason requirement. Rather, I argue that without a public reason requirement they are not *sufficiently* protected from infringements of their liberties that imply a violation of the principle of moral restraint, which gives expression to a moral respect that liberal citizens owe each other. Thus, an objector is obliged to demonstrate either how he guarantees that the principle of moral restraint is not violated without a public reason requirement or that this principle need not be taken into account at all. Inclusivists like Eberle have indeed tried to formulate such an argument; see Eberle, *Religious Conviction in Liberal Politics*, 84–151, and "Basic Human Worth and Religious Restraint." But authors like Zoll have shown why these and similar argumentative strategies fail; see Zoll, *Perfektionistischer Liberalismus*, 225–36, 396–403. Thanks to an anonymous reviewer for raising this worry.

26 Cf. Wall, "On Justificatory Liberalism," 125. Wall refers here to Gaus, who in turn draws on

A commitment to the presumption in favor of liberty implies a commitment to a principle of the non-violation of moral status:

7. The justification of a political norm PN through a person *A* implying a coercive interference with the liberty of person *B* to choose and lead a life according to the evaluative standards of her own conception of a good life is solely no violation of the moral status of *B* if *A* gives *B* considerations that *B* can access as a weighty reason that justifies PN from the evaluative standards of her own conception of a good life.[27]

Additionally, a commitment to the principle of the non-violation of moral status gives citizens of liberal democratic societies an independent reason to accept moral restraint as a necessary condition for a reasonable account of public justification:

8. The principle of moral restraint expressed in claim 3 should be accepted because it excludes the possibility of public justifications of political

---

Joel Feinberg to formulate this principle; cf. Gaus, *Justificatory Liberalism*, 165; and Feinberg, *Harm to Others*, 9. This formulation of the principle is taken over from Zoll, *Perfektionistischer Liberalismus*, 232.

27 It could be objected that a commitment to a presumption in favor of liberty does not imply a commitment to a principle of the non-violation of moral status because of cases where a state overcomes the presumption in favor of liberty in order to prevent harm or promote justice or the common good without thereby appealing to citizens' own evaluative standards of a good life. I would reply that these cases constitute no counterexamples because a reference to the prevention of harm, etc., just illustrates that infringements of liberties are only legitimate if they can be justified with public reasons. In order to get the objection running, it needs to be assumed that those considerations are in principle not accessible as public reasons according to the evaluative standards of some conceptions of the good life. In other words, it must be assumed that these reasons are external reasons that have no connection at all to the evaluative standards of the conceptions of the good life of at least some citizens. First of all, I doubt that such an extreme externalism is a plausible account of reasons at all. If you tell me that you are forcing me to do something for my own good or for the good of the community in order to prevent harm or to foster the common good but neither I nor an appropriately idealized version of me is ever able to understand what the harm is or the common good consists in, what kind of "reason" are you giving me? I do not see how this does not constitute a serious violation of my moral status. Second, I would challenge the claim that the mentioned considerations are a good example of external reasons that do not appeal to citizens' own evaluative standards of a good life. Rather, I would maintain that in every reasonable liberal conception of a good life considerations of harm, justice, and the common good are playing a role in evaluating and answering the question of whether a certain *political* measure PN contributes to one's flourishing according to one's own conception of a good life. I owe my thanks to an anonymous reviewer for making me aware that this objection needs to be addressed.

norms that violate the moral status that citizens of a liberal and democratic society attribute to each other.

Steps 6–8 reveal what is ultimately wrong with the first-personal model of inclusivism. According to the presumption in favor of liberty, it is constitutive for liberal democratic societies that their citizens acknowledge as the moral status quo that people are free and entitled to live a life they judge good according to the evaluative standards of their particular conception of the good life. But this implies that it is constitutive for liberal societies to accept a principle of the non-violation of moral status and to treat liberty and the coercive interference with liberty in an asymmetrical manner.

Because of this asymmetry the burden of proof is on the side of that epistemic first-personal perspective that wants to coerce another epistemic first-personal perspective. In other words: the epistemic obligation to give priority to the first-personal perspective of the person I want to coerce and the obligation to refrain from advancing and enforcing demands on that person if I cannot justify them to her with reasons she can comply with from her epistemic perspective is rooted in the normative obligation to give priority to liberty over coercive interference. If I do not honor this epistemic obligation, I violate a moral obligation because I do not treat my fellow citizens as persons who are entitled and able to choose and lead a life according to their conception of a good life. Thus, the independent reason to accept moral restraint as a necessary condition for any reasonable conception of public justification derives from a prior commitment to liberty and equality as constitutive values for liberal democratic societies. Therefore, Wall's first-personal account of public justification as a substitute for the exclusivist's third-personal account has to be rejected because it ultimately contradicts the normative consequences that derive from a commitment to the values of liberty and equality.

This result is fatal to Wall's first-personal model of inclusivism because Wall's only possibility to demonstrate that the principle of the non-violation of moral status does not constitute an independent reason for embracing moral restraint consists in attacking the quite plausible presumption in favor of liberty. Nevertheless, he tries to undermine this presumption by claiming that it should be rejected for moral reasons because it implies an asymmetrical treatment of two cases that should be treated symmetrically from a moral point of view:

9. Two cases:
    a. Adam does something morally blameworthy if Adam interferes coercively with Beatrice's liberty to choose and lead a life according to the evaluative standards of Beatrice's conception of a good life.
    b. Adam does something morally blameworthy if Adam is able to pro-

> mote or protect something that is an important good for Beatrice but refrains from doing so.[28]

According to Wall, there is a strong moral intuition that Adam is to blame in both cases. This intuition indicates that the two cases should be treated symmetrically and not asymmetrically from a moral point of view. If this is true, it demonstrates that the presumption in favor of liberty is wrong because it implies that Adam is only in case 9a morally obliged to justify himself for his action but not in case 9b.[29] This means that liberty and the absence of coercive intervention is not the moral status quo, and that not only interference but also non-interference with liberty to promote or protect some good requires justification.[30]

But this attack on the presumption in favor of liberty is not successful because there is an easy way to show that there is a strong reason to treat cases 9a and 9b asymmetrically. Wall's rebuttal is only successful because he omits a premise that allows him to distinguish between the following cases:

> 9b1. Adam does something morally blameworthy if Beatrice is unable to realize something that is an important good for her that Adam is able to promote or protect, but Adam refrains from doing so.
> 9b2. Adam does something morally blameworthy if Adam and Beatrice are both able to realize something that is an important good for Beatrice, but Adam refrains from doing so.

I think it is quite plausible to say that in case 9b1 Adam acts morally blameworthy but not in case 9b2. This is the case because "Samaritan duties" just arise for Adam if Beatrice is not able to realize on her own what is good for her.[31] Therefore, it seems awkward to assume that Adam is required to justify that he does not help Beatrice to realize a good if Beatrice is able to realize it by her own efforts. Even worse, Wall's argument that noninterference requires in the same way a justification as interference reveals that he is not willing to accept that Adam's interference on behalf of Beatrice's good undermines her moral status if she is able to realize it on her own. Beatrice's moral status would be undermined through Adam's interference in 9b2 because Adam's interference implies that she is *not able* to choose and lead a good life by her own judgment. Yet, this is clearly an expression of a kind of paternalism no one can reasonably expect to endorse with all its annoying consequences.

28 Cf. Wall, "On Justificatory Liberalism," 130.

29 Cf. Wall, "On Justificatory Liberalism."

30 Cf. Wall, "On Justificatory Liberalism," 125, 129.

31 Cf. Wall, "On Justificatory Liberalism," 130.

2. Why Exclusivism Fails: The Third-Personal Account of Public Justification

My second thesis is now that the failure of the first-personal model of inclusivism does not count in favor of exclusivism because exclusivism is not able to accommodate a strong and plausible moral intuition that speaks in favor of inclusivism. According to this intuition, the adoption of a third-personal conception of public justification is morally problematic for religious citizens because it obliges them to untie public justification from their religious first-personal perspective at the cost of moral integrity. This argument is commonly called the "integrity objection" and can be reformulated as follows.[32]

10. A person leads a life of integrity if she acts in concert with the ideals and norms that are constitutive of her identity.
11. The ideals and norms that are constitutive of a person's identity derive from their conception of a good life.
12. The ideals and norms that derive from a religious conception of a good life require that the evaluative standards of the religious tradition someone is committed to do not have important justificatory weight just in private but in all matters, including the political ones.
13. A commitment to epistemic restraint requires that religious citizens refrain from referring to the evaluative standards of their particular religious tradition in the case of the public justification of political norms.
14. If religious citizens are obliged to refrain from referring to their religious evaluative standards in the process of the public justification of political norms, their religious evaluative standards and the reasons they generate have necessarily no justificatory weight in political matters. But this means that a commitment to epistemic restraint conflicts with a commitment to a religious conception of the good life, which is constitutive for the identity of religious citizens.
15. Therefore, a religious citizen who embraces epistemic restraint is unable to have identity integrity.

32 A classical formulation of the intuition that motivates this argument is given by Wolterstorff; cf. Audi and Wolterstorff, *Religion in the Public Square*, 105. I partly follow Vallier in the reconstruction of this argument; cf. Vallier, *Liberal Politics and Public Faith*: *Beyond Separation*, 57–66. This version of the argument has an advantage over other formulations in that it makes clearer that the integrity objection derives its force from a combination of moral and epistemological considerations. Therefore, I disagree with classificatory schemes that interpret this argument as a species of "ethical arguments"; cf. Breul, *Religion in der politischen Öffentlichkeit*; and Neal, "Is Political Liberalism Hostile to Religion?"

If this argument is sound, a third-personal conception of public justification is in a similar way as mono-relational as a first-personal conception, and exclusivists have a prudential and a principled reason for not being content with this fact. Prudentially, it seems not to be wise to violate the integrity of religious citizens in such a systematic way because it confronts religious citizens necessarily with a conflict of loyalties: either they can be fully committed to the normative ideals of liberalism and its core idea that political coercion should be tied to public justification, or they can be fully committed to the normative ideals of their religious tradition. As Paul Weitham has argued, there is a lot of empirical evidence that this does not do justice to the important contributions of religious traditions that gave rise to and that maintain democracy.[33] Even worse, as Jeffrey Stout has convincingly shown, the demand of epistemic restraint is most probably one of the main causes that led to an alienation of religious citizens from liberal democracy. This alienation is highly problematic because it has given rise to an anti-democratic radicalization of religious traditions and a dialectical backlash in the form of the so-called new traditionalism.[34]

But even if exclusivists are not convinced of this kind of prudential reasoning because they doubt the empirical evidence, they have to acknowledge that the integrity objection shifts the burden of proof in favor of inclusivism at least for the principled moral reason that the demand of epistemic restraint infringes significantly on the expressive freedom of religious citizens in the public realm. As we have seen above, a commitment to the presumption in favor of liberty obliges not only religious citizens but also exclusivists to justify their liberty-infringing demands to those addressed by these demands. Consequently, exclusivists owe inclusivists a justification for their demand of epistemic restraint. Otherwise religious citizens could rightly object that they are not treated as they should be treated because their moral status is violated.

3. Public Justification from the Second-Personal Perspective: The Construction of an Inclusivist Model of Public Reason Liberalism

So far, I have shown that there is an argumentative impasse between exclusivism and inclusivism because neither side can offer a conception of public justification that is able to accommodate the moral intuition that motivates the other side to embrace their account of public justification. Both parties can rightly claim that the opposing conception of public justification is in a problematic way mono-relational.

33 Cf. Weithman, *Religion and the Obligations of Citizenship*.

34 Cf. Stout, *Democracy and Tradition*.

Yet, I will argue in this third section that it is possible to construct an inclusivist model of public reason liberalism that is fully bi-relational because it neither obliges religious citizens to disregard the beliefs and values of their first-personal perspective when political norms have to be publicly justified nor permits a religious majority to coerce a minority without giving this minority accessible reasons to comply with. In short: I claim that this model breaks the impasse in favor of inclusivism because it can accommodate with its bi-relational character the moral intuitions of both sides.

The argumentative strategy of this inclusivist version of public reason liberalism is to demonstrate that there is a gap between the principle of moral restraint and the principle of epistemic restraint that is implied in a third-personal conception of public justification.[35] In other words: it is false to assume that the acceptance of the principle of moral restraint as a necessary condition for any reasonable conception of public justification implies a restriction of the set of possible conceptions of public justification to conceptions that demand epistemic restraint. A third-personal conception does not result necessarily from a commitment to a principle of moral restraint as an epistemic ideal for public justification because such a principle can also be respected and fulfilled by a different convergence conception of public justification that derives its normative implications from the adoption of a second-personal standpoint:[36]

4**. Convergence conception of public justification:
  a. The imposition of a political norm through a democratic procedure is only legitimate iff it is publicly justified.
  b. A political norm is publicly justified iff
    b1. the group of appropriately idealized citizens *A* who want to

35 Here I follow Vallier, who presents and defends this strategy in much more detail; see Vallier, *Liberal Politics and Public Faith: A Philosophical Reconciliation* and *Liberal Politics and Public Faith: Beyond Separation*. Although I agree with Vallier that this is the best strategy to defend inclusivism, I disagree with him about the exact outcome of this move because I defend a different convergence conception of public justification.

36 Therefore, as I have mentioned above, it constitutes a fallacy to infer from the need to justify to each citizen a need to justify using mutually accessible reasons. A concise summary of this point can also be found in Billingham, "Convergence Justifications within Political Liberalism," 136–38. The possibility of a convergence conception of public justification was developed and introduced independently into the debate by a couple of authors, but the most elaborated account can be found in the work of Vallier; cf. D'Agostino, *Free Public Reason*, 30–33; Gaus, "The Place of Religious Belief in Public Reason Liberalism"; Gaus and Vallier, "The Roles of Religious Conviction in a Publicly Justified Polity"; Stout, *Democracy and Tradition*, 65–85; Vallier, "Convergence and Consensus in Public Reason" and *Liberal Politics and Public Faith: Beyond Separation*.

impose the political norm PN on the group of appropriately idealized citizens *B* give *B* a sincere and honest justification of PN, which means that they publicly and in a sincere and honest way state what the considerations CA are that motivate them to support the imposition of PN on *B*;

b2. it is intelligible for *B* that *A* is justified according to the evaluative standards ESA of their first-personal epistemic standpoint to believe that CA justifies PN; and

b3. *A* gives *B* a consideration CB that *B* can access as a weighty reason that justifies PN according to *B*'s evaluative standards ESB.[37]

The difference between this convergence conception of public justification and the exclusivist's third-personal conception is that a commitment to the former does not demand that religious citizens exercise epistemic restraint. This means that the decisive advantage of a convergence conception over a third-personal conception is that a convergence conception is not vulnerable to the integrity objection of religious citizens because it does not disconnect public justification from their first-personal perspective.[38] Therefore, it is not mono-relational in the way that a third-personal conception is.

This is the case because a convergence conception rejects the claim that only those considerations can have justificatory weight in the process of the public justification of political norms that are *mutually* accessible. There is no need for common evaluative standards like ESAB that would enable *A* to recognize CB as a reason that justifies PN and would enable *B* to acknowledge CA as a reason that justifies PN. Common evaluative standards are not necessary because PN can be

37 In contrast to Vallier's convergence conception of public justification, I maintain that a political norm PN is publicly justified if each appropriately idealized member of the public has a "weighty"—instead of a "sufficient"—reason to endorse PN. What I call a "weighty" reason has to meet all the criteria Vallier mentions for a "sufficient" reason (epistemic justification in the form of access internalism, adequate standards of inference and evidence, etc.) minus the requirement that this reason must also override or defeat reasons that contradict it; see Vallier, *Liberal Politics and Public Faith: Beyond Separation*, 27–28, 104–6. This makes a practical difference for situations where a political norm can only be inconclusively publicly justified, which I will spell out below in more detail. Thanks to an anonymous reviewer who made we aware that I have to clarify how my view differs from Vallier's.

38 To be clear, the point of my argument is not that religious citizens do not have integrity costs or conflicts of loyalty at all or that they just have fewer costs and fewer conflicts if an inclusivist model of public reason liberalism is adopted (which I think is also true). The decisive advantage of a convergence conception over a third-personal conception is that there is no *principled* disconnection between public justification and the first-personal perspectives of religious citizens. For this reason, the integrity objection does not apply to the inclusivist position I am defending. Thanks to an anonymous reviewer for pushing me here.

publicly justified through a *convergence* of the mutually inaccessible reasons CA and CB.

The mutual inaccessibility of CA and CB constitutes no problem because citizen *A* has fulfilled their moral obligation against *B* to justify their demand to *B* with considerations that *B* can access as having justificatory weight. In other words, the political norm PN is justified through a convergence of different and mutually inaccessible first-personal standpoints and there is no additional need that the arguments that serve for the public justification of PN be evaluated from a third-personal perspective. A second-personal approach that implies that the participants adopt the second first-personal standpoints of their fellow citizens is all that is needed for the public justification of political norms.

A further advantage of this model is that it can accommodate the moral intuition that motivates exclusivism by showing that the endorsement of a third-personal perspective and of mutual accessibility through common evaluative standards is not needed in public justification to rule out the problematic cases exclusivists fear most. In contrast to a first-personal conception of public justification, a convergence conception is based on a relational conception of justification and accepts that moral restraint has to be exercised if a demand to the addressed person cannot be justified. However, it is not necessary to rely only on reasons that are mutually accessible, as a third-personal conception claims, in order to fulfill this obligation. It can also be fulfilled by reasoning from different, second first-personal perspectives.

In summary, a convergence conception provides a middle course between a third-personal and a first-personal conception because inclusivists can coherently maintain with a first-personal conception of public justification but against a third-personal conception that the religious reasons of their first-personal perspective have genuine justificatory weight in the process of public justification. Yet, with a third-personal conception and against a first-personal conception, they do not have to substitute a relational conception of public justification with the problematic simple account of public justification.

If this is right, I have demonstrated that there is a gap between moral restraint and a third-personal conception of public justification because the convergence conception fulfills with its acceptance of the principle of moral restraint the necessary condition for a reasonable conception of public justification without having to accept the principle of epistemic restraint. Therefore, inclusivists who adopt this model have an advantage over exclusivists as long as they cannot show that there are independent weighty reasons for preferring a third-personal conception to a convergence conception. If epistemic restraint in terms of accessibility is not necessary to rule out the unjustified imposition of political norms

on minorities, exclusivists have to demonstrate what exactly is problematic with religious arguments in the process of the public justification of political norms.

4. A DEFENSE OF THE INCLUSIVIST MODEL OF PUBLIC REASON LIBERALISM

A full defense of the constructed inclusivist model of public reason liberalism needs to meet two requirements. The first task is to refute objections from exclusivists (e.g., the sincerity objection) against the convergence conception of public justification on which it rests.[39] A second task consists in putting the model to work. It needs to be shown how it can rebut exclusivist arguments that try to justify the exclusion of religious arguments from the process of the public justification of political norms. In this section, I will concentrate on the second task by rebutting an important exclusivist argument that was recently presented by Jonathan Quong.

Central to Quong's defense of exclusivism is his claim that he is able to present a new argument that justifies an asymmetrical treatment of controversial perfectionist and anti-perfectionist reasons in the process of the public justification of political norms.[40] According to Quong, reasons should be excluded from the set of public reasons if their employment leads to a problematic reasonable disagreement. Unproblematic reasonable disagreements are called "justificatory" and can be defined as follows.[41] A disagreement is justificatory in nature iff

a. the participants of the debate use evaluative standards in the premises of their reasons that are incompatible but mutually accessible as having justificatory relevance, and
b. the disagreement is only about the justificatory weight of the evaluative standards and the conclusions that derive from these premises.

Such a disagreement is illustrated by Quong as a dispute between the liberals Sara and Tony over the question of whether it is just to allow the Catholic

39 This is an ongoing debate, but good defenses against a range of possible exclusivist objections can be found, for example, in Billingham, "Convergence Justifications within Political Liberalism"; Vallier, "In Defense of the Asymmetric Convergence Model of Public Justification" and *Liberal Politics and Public Faith: Beyond Separation*; and Zoll, *Perfektionistischer Liberalismus*.

40 As I said in note 1 above, I regard religious reasons as a kind of perfectionist reasons. Quong himself seems to agree with this, as his use of religious examples makes clear; cf. Quong, *Liberalism without Perfection*, 192–93.

41 Cf. Quong, *Liberalism without Perfection*, 194, 204–8. My presentation of Quong's original formulation of "justificatory disagreements" is slightly revised in order to adapt it better to the purposes of this article.

Church to discriminate on the basis of gender when employing priests.[42] Tony argues that the Catholic Church is entitled to hire exclusively male priests because it is a private institution and that a prohibition to do so would infringe on the religious liberty of Catholics.[43] Sara disagrees and responds with two arguments. First, private institutions are not exempt from laws against rape, theft, and murder and should therefore also be not exempt from laws that prohibit gender discrimination in employment. Second, if there is a compelling egalitarian reason to interfere, the right to religious liberty can be violated because it is not meant to insulate religious groups against all interference.[44]

Such a disagreement is "justificatory" because the reasons Sara and Tony give each other are derived from the fundamental normative framework they share as liberals. This means that they can reject the reasons the other gives as inconclusive, but they cannot complain that they are not addressed with reasons they can access as having justificatory relevance for them.[45]

Problematic reasonable disagreements in contrast are called "foundational" and can be defined as follows.[46] A disagreement is foundational in nature iff

a. the participants of the debate use evaluative standards in the premises of their reasons that are incompatible and not mutually accessible as having justificatory relevance, and
b. the disagreement is about the justificatory relevance of the evaluative standards themselves.

Quong's example for such a disagreement is a dispute between the liberals Mike and Sara over the question of the immorality of recreational drug use. Mike believes that the use of drugs is immoral because it constitutes an action that conflicts with what God commands.[47] Sara, in contrast, has a hedonistic conception of the good life and believes that there is nothing morally wrong with the use of drugs for recreational purposes. First of all, she rejects Mike's argument because she does not believe in the existence of God and therefore has no reason to believe that a reference to God as a moral authority is relevant to determine whether private drug consumption is morally permissible or not. Second, she herself adheres to a conception of morality according to which an action is only

42 Cf. Quong, *Liberalism without Perfection*, 205–6.

43 Cf. Quong, *Liberalism without Perfection*, 205.

44 Cf. Quong, *Liberalism without Perfection*.

45 Cf. Quong, *Liberalism without Perfection*, 204–7.

46 Cf. Quong, *Liberalism without Perfection*. Again, I have revised Quong's formulation to bring into focus some aspects I am interested in for this article.

47 Cf. Quong, *Liberalism without Perfection*, 204–5.

immoral if it does damage to another person. So, the use of drugs for one's own recreation and pleasure is simply not a matter of morality for her and consequently there is nothing morally wrong with it.[48]

This means that the conflict between Mike and Sara is somehow deeper than the conflict between Sara and Tony because they even disagree about the justificatory relevance of the evaluative standards the other party is using. Consequently, they cannot evaluate the reasons the other is giving because they disagree about the standards that evaluate reasons as good or bad. Thus, it is characteristic for foundational disagreements that there is no shared normative framework, no deeper standard of justification that could serve as the basis for adjudicating the dispute.[49]

Quong's argument for exclusivism now runs as follows:

16. Reasonable disagreements about the good life are not necessarily justificatory and will almost certainly be foundational.
17. Reasonable disagreements about justice are necessarily justificatory and not foundational.
18. The liberal principle of legitimacy is not violated when the state imposes a view that arises out of justificatory disagreement.
19. The liberal principle of legitimacy is violated when the state imposes a view that arises out of a foundational disagreement.
20. Therefore, arguments that refer in their premises to controversial evaluative standards about the good life—including religious standards—should be excluded from the set of reasons that can play a role in the process of the public justification of political norms, and there is nothing wrong in admitting arguments that refer in their premises to controversial evaluative standards about justice.[50]

I have my doubts concerning premises 16 and 17.[51] But for the sake of argument, I will grant their truth and concentrate my critique on claim 19. This claim is central to Quong's argument because its function is to explain why it is problematic when the state imposes a political norm that is justified *solely* with arguments about the good life about which a foundational disagreement exists. Quong argues that such cases are problematic because they violate the liberal principle of legitimacy:

48 Cf. Quong, *Liberalism without Perfection*, 205.

49 Cf. Quong, *Liberalism without Perfection*.

50 Cf. Quong, *Liberalism without Perfection*, 204. I have adapted Quong's argument slightly for the purposes of this article.

51 Critical remarks in this sense are offered for example by Fowler and Stemplowska, "The Asymmetry Objection Rides Again; and Zoll, *Perfektionistischer Liberalismus*, 179–83.

21. The standard of liberal legitimacy is not reasonable rejection, but it asserts that the state should not act on grounds that citizens cannot reasonably be expected to endorse.[52]

Yet, the problem with claim 21 is that it only shows that inclusivists violate the standard of liberal legitimacy if they rely on a first-personal conception of public justification. A reliance on a first-personal conception of public justification constitutes a violation of the standard of liberal legitimacy because it even permits cases when a political norm is publicly justified though it cannot be reasonably expected that all citizens have a weighty reason to endorse this norm. But inclusivism is compatible with the standard of liberal legitimacy if inclusivists rely on the convergence conception of public justification formulated by claim 4**.

According to this conception it is true that a political norm is only publicly justified if this norm is justified to each citizen with a weighty reason he can reasonably be expected to endorse.[53] But this conception allows for cases when a political norm is publicly justified through a convergence of mutually inaccessible reasons. This gives room for the employment of arguments that rely on controversial evaluative standards about the good life. This means that Quong's attempt to justify an asymmetrical treatment of controversial perfectionist and anti-perfectionist reasons in the process of public justification fails. Yet, if there is no justification of an asymmetrical treatment of religious reasons, the mono-relational character of a third-personal conception of public justification remains problematic and a bi-relational inclusivism is preferable to exclusivism.

I will return to Quong's example of the conflict between Mike and Sara to illustrate *how* my inclusivist model of public reason liberalism can counter Quong's argument by accommodating the moral intuition that motivates exclusivism without thereby having to accept epistemic constraint.[54] According to Quong, we have a scenario here that can be characterized in the following way:

22. Mike has a religious and non-hedonistic conception of the good life that is controversial, because it is rejected by Sara who has a nonreligious and hedonistic conception of the good life. According to the evaluative standards of Mike's conception of the good life (ESM),

52 Cf. Quong, *Liberalism without Perfection*, 209.

53 In contrast, on Vallier's account it would be necessary that each citizen has a *sufficient* reason to endorse the political norm. This means that each citizen would need to have a reason that overrides all the reasons he might have for rejecting the political norm; see Vallier, *Liberal Politics and Public Faith: Beyond Separation*, 27–28.

54 Due to his emphasis on sufficient reasons, it is not possible for Vallier to rebut Quong's argument as I will do in the following. I regard this as an advantage of my model of inclusivist public reason liberalism over Vallier's.

Mike's consideration (α) that drug abuse is against God's will because it is a vice according to biblical teaching (e.g., Gal. 5:19–21 or Eph. 5:18) constitutes a weighty reason RM that justifies the political norm NOPE that prohibits recreational drug use.

23. Sara has a nonreligious and hedonistic conception of the good life that is controversial, because it is rejected by Mike who has a religious and non-hedonistic conception of the good life. According to the evaluative standards of Sara's conception of the good life (ESS), Sara's consideration (β) that recreational drug use is permissible because it does no harm to others and produces a considerable amount of pleasure constitutes a weighty reason RS that justifies the political norm DOPE that allows recreational drug use.

From this follows a foundational disagreement between Mike and Sara:

24. Because of the mutual rejection of their conceptions of the good life, they also reject the evaluative standards of each other as justificatorily relevant with the following consequences:
    24a. Mike cannot reasonably expect from Sara that consideration (α) constitutes a reason for her that justifies NOPE; and
    24b. Sara cannot reasonably expect from Mike that consideration (β) constitutes a reason for him that justifies DOPE.

An imposition of NOPE or DOPE through the state via a democratic decision procedure would be problematic, because either Sara or Mike could object that they have not been addressed with a consideration that they can reasonably be expected to endorse as a reason that justifies NOPE or DOPE. Quong and other exclusivists follow from scenarios like this that this supports their view that only reasons that are mutually accessible can play a role in the process of public justification.

But this inference is wrong, as I will show with the following case:

25. Mike addresses Sara with the consideration (δ) that there are empirical studies that show that frequent and continuous drug abuse leads to a decrease in personal well-being and pleasure in the long run, which means that even by her own evaluative standards ESS she has a weighty reason RS′ that justifies NOPE.
26. Sara addresses Mike with the consideration (ϕ) that his own religious tradition acknowledges that it is not the purpose of the state to eliminate all vices and to make its citizens holy, which means that even by

> his own evaluative standards ESM he has a weighty reason RM′ that justifies DOPE.[55]

In this scenario, NOPE can for example be publicly justified through a convergence of the reasons RM (religious reason for NOPE) and RS′ (hedonistic reason for NOPE), which are not mutually accessible for Mike and Sara. RS′ (hedonistic reason for NOPE) is no reason Mike accepts because he rejects the hedonistic evaluative standards of Sara as having justificatory relevance, but he fulfills his obligation to address Sara with a reason he can reasonably expect Sara to endorse. And his own reason RM (religious reason for NOPE) can enter the process of the public justification of NOPE because a convergence conception does not demand from him to refrain from arguing for his preferred political options with controversial reasons that derive from his particular first-personal epistemic standpoint and their evaluative standards.

In contrast to the first-personal model of inclusivism proposed by Wall, the problematic cases exclusivists fear most are ruled out by this inclusivist model of public reason liberalism because religious reasons can never justify a political norm *alone* in a plural society.[56] It is not the democratic procedure that legitimizes NOPE in this case. In this example NOPE is legitimately imposed through a democratic procedure because beforehand it was publicly justified to each citizen with reasons these citizens can be reasonably expected to endorse. Once citizens have fulfilled their duty to address each other with considerations they can reasonably expect the other to endorse as reasons, a democratic decision procedure is no more problematic than in the case of the justificatory disagreement between Tony and Sara.

What remains between Sara and Mike is a disagreement about the weighting of RM (religious reason for NOPE) against RM′ (religious reason for DOPE) and RS (hedonistic reason for DOPE) against RS′ (hedonistic reason for NOPE). Therefore, the public justification of NOPE or DOPE is necessarily inconclusive. But this inconclusiveness, as Quong himself explicitly admits, does not make it

55 If Mike is a Catholic, Sara could, for example, refer to Augustine, De Lib. Arb. I, 5, 6 or Thomas Aquinas, ST I-II, q. 91, a. 4., corp.

56 This is not in contradiction to the inclusivist's claim that religious citizens do not have to exercise restraint if they have only religious reasons that can justify a political norm. They have to exercise restraint if they cannot give their fellow citizens weighty reasons for the acceptance of the norm these can access from their epistemic standpoints. But religious citizens themselves are not obliged to accept these reasons as reasons for themselves. Therefore, inclusivists can maintain that there are cases where they do not have to exercise restraint even if they have only religious reasons to justify their preferred political norms. This is the decisive difference to the Rawlsian *proviso* model of religious reasoning in a liberal public or other variants of "weak inclusivism."

illegitimate for the state to act on the basis of either the combination of RM and RS′ or RS and RM′ because the standard of liberal legitimacy is not reasonable rejectability but that the state should not act on grounds that citizens cannot reasonably be expected to endorse.[57] In either case, Sara and Mike are addressed with a weighty reason they can access from their epistemic standpoint so that neither Mike nor Sara can make a reasonable complaint against the outcome of the democratic procedure that decides on the prohibition or permissiveness of recreational drug use.

Against this outcome Quong could object that the scenario I described with the claims 25 and 26 still constitutes a violation of the standard of liberal legitimacy expressed in claim 21 for the following reason: if it is a combination of RM and RS′ or RS and RM′ that publicly justifies NOPE or DOPE, either Mike could object that Sara justifies DOPE to him with a reason RS (hedonistic reason for DOPE) he cannot reasonably be expected to endorse or Sara could object that Mike justifies NOPE to her with a reason RM (religious reason for NOPE) she cannot reasonably be expected to endorse.

But this objection fails because it does not acknowledge that reasons have a different function in the process of public justification according to a convergence conception. In the case that a combination of RM and RS′ justifies NOPE, it is RM (religious reason for NOPE) that justifies Mike's support for NOPE. A convergence conception of public justification just demands from Sara with regard to RM that it should be intelligible for her that consideration (*a*) constitutes a weighty reason for Mike according to his evaluative standards that justify NOPE. It is reason RS′ (hedonistic reason for NOPE) that has the function to justify NOPE to Sara. Likewise a convergence conception just demands from Mike with re-

57 Cf. Quong, *Liberalism without Perfection*, 209. Vallier is developing his "principle of convergent restraint" in the same direction when he substitutes the criteria of reasonable rejectability with reasonable expectability; cf. Vallier, *Liberal Politics and Public Faith: Beyond Separation*, 185–88. But this case illustrates nicely to what extent our different convergence conceptions of public justification lead to different practical results in situations where a political norm is not conclusively publicly justified. For example, according to Vallier, DOPE would not be publicly justified through the fact that both Sara and Mike have a weighty reason to endorse DOPE. This is the case because RM′ (religious reason for DOPE) is not necessarily a sufficient reason for Mike to endorse DOPE. According to the evaluative standards he is committed to, he has to acknowledge that RM′ is a weighty reason for DOPE. But he may be justified in regarding RM (religious reason for NOPE) or other reasons as overriding RM′ with the consequence that RM′ is not a sufficient reason for him to endorse DOPE. Thus, according to Vallier, it would be illegitimate for the state to act on a combination of RS (hedonistic reason for DOPE) and RM′ (religious reason for DOPE). This is not the case on my account of convergence justification. Therefore, I think that my approach is better suited to rebut the challenge formulated by Quong. I owe my thanks to an anonymous reviewer for making me aware that I should exemplify how Vallier's and my convergence conception of public justification differ.

gard to RS′ that it should be intelligible for him that consideration (δ) constitutes a weighty reason for Sara according to her evaluative standards that justify NOPE.

To reiterate that RM (religious reason for NOPE) cannot play a role in the public justification because RM is not accessible to Sara is to beg the question because the application of the convergence conception of public justification to the case of Sara and Mike has demonstrated that the rejection of epistemic restraint in terms of accessibility does not necessarily imply a violation of the liberal principle of legitimacy expressed by claim 21. The lack of common evaluative standards and the resulting foundational disagreements can make the process of the public justification of political norms more complicated, but this fact does not justify the exclusion of a significant number of reasons as justificatorily irrelevant with the rationale that they do not fulfill the equally demanding formal norm of being mutually accessible.

Therefore, the upshot of this fourth section is that my inclusivist version of public reason liberalism is able to rebut Quong's attempt to justify the claim that the employment of religious reasons is problematic and that religious citizens are therefore obliged to accept epistemic restraint. Thus, if the argument of this paper is sound, the burden of proof is on the exclusivist's side. This is the case for two reasons. First of all, I have demonstrated that inclusivists can accommodate the valid moral intuition that motivates exclusivism without thereby accepting epistemic restraint by relying on a convergence conception of public justification. Second, I have shown in a case study how my inclusivist model of public reason liberalism can be defended against an exclusivist attempt to justify the exclusion of religious reasons from the set of public reasons. Therefore, the answer to the question "Are citizens obliged to refrain from using religious arguments for the public justification of political norms in a liberal democracy (e.g., a law) if these are the *only* justificatory reasons *they* have to embrace this norm?" is no unless exclusivists present new arguments that suggest otherwise.

5. HOW MY INCLUSIVIST PROPOSAL DIFFERS FROM OTHER MODERATE INCLUSIVIST ACCOUNTS OF RELIGIOUS REASONING IN THE LIBERAL PUBLIC

This leaves me with the task to clarify how my proposal for acceptable religious discourse in the liberal public differs from other "moderate" inclusivist accounts advanced in the literature, i.e., accounts that have in common that they try to reconcile an inclusivism with the normative demands that derive from a commitment to liberal core values like freedom and equality.[58] I will address this task in two steps. First of all, I will explain in a summary fashion how my position differs

58 I am thankful to an anonymous reviewer for making me aware of the need to address this issue.

from the exclusivist and inclusivist positions I have discussed in detail up to this point. Second, I will compare two weak inclusivist accounts recently advanced in the academic literature with my strong inclusivist account, according to which there is no principled reason to exclude religious reasons from the process of public justification, i.e., religious reasons that cannot be translated into a secular language or evaluated based on common evaluative standards because their soundness depends on the acceptance of revealed knowledge or religious authority. Different from the strong inclusivism I am advocating for, the relevant authors defend a weak inclusivism according to which not all religious reasons have to be excluded from public justification but certainly those religious arguments that refer to revelation and are therefore in principle inaccessible for nonreligious citizens. I will limit myself to the discussion of two weak inclusivist accounts not only for reasons of space but also because they exemplify two of the most promising strategies to justify the exclusion of this specific type of religious argument. And the comparison with these inclusivist variants is sufficient to sharpen the specific contours and conditions of acceptable religious discourse I am proposing. According to the first weak inclusivist strategy, religious arguments referring to revelation have to be excluded in virtue of a moral consideration, namely, that they violate certain moral obligations such as respect that we owe to each other in a liberal society. The second strategy justifies the exclusion of these reasons on the basis of an epistemic criterion, namely, their lack of accessibility.

5.1. The Difference My Inclusivist Public Reason Liberalism Makes to the Exclusivist and Inclusivist Positions Discussed So Far

My inclusivist version of public reason liberalism makes a practical difference to the alternative positions discussed so far in three respects. First of all, in agreement with an exclusivist position and Vallier but against Wall's first-personal model of inclusivism I accept that the epistemic perspectives of those addressed by demands are relevant for a successful public justification of political norms. In consequence, I agree that religious citizens should exercise *moral restraint* if they cannot justify their demand to their fellow citizens with considerations that these citizens can access as a weighty reason to comply with the demand.

Second, I part company with exclusivists like Quong and side with the inclusivist models presented by Wall and Vallier in their rejection of *epistemic restraint*. Religious citizens are obliged to exercise moral restraint but not epistemic restraint. There is no reason to exclude reasons that derive from particular first-personal epistemic perspectives and are not mutually accessible, like religious reasons that refer to revelation from the process of the public justification of political norms. Citizens are just demanded to fulfill their duty to address

their fellow citizens with different considerations they can access from their different epistemic standpoints as having justificatory relevance.

Finally, I part company with Vallier's inclusivist model of public reason liberalism in proposing that public justification just demands that each appropriately idealized citizen has a weighty—instead of a "sufficient"—reason to endorse a political norm. As a consequence, religious reasons lose some force as *defeaters* against the imposition of political norms that conflict with the respective religious conceptions of a good life. However, religious reasons do not only enter the process of public justification as justificatory reasons to determine whether each person has sufficient reason to endorse a proposed political norm.[59] Rather, the state may permissibly act on religious reasons in combination with other, nonreligious reasons in situations of inconclusiveness. In these cases, the inconclusiveness can be resolved by a democratic procedure like voting as is possible in Quong's model of exclusivism.

5.2. The Difference from Andrew March's Weak Inclusivist Proposal for Acceptable Religious Discourse in the Liberal Public

Central to March's weak inclusivist proposal for acceptable religious discourse is the presentation and defense of a typology of different kinds of religious reasons and a typology of different areas of political and social life that coercive laws regulate or about which political communities deliberate.[60] On the basis of these typologies he argues that religious reasons should be excluded from public reasoning if two criteria are met. First, a religious argument refers to a scriptural, revealed, or clerical command, i.e., a command that is extracted from a revealed text, religious authority, or personal mystical or revelatory experience.[61] Second, such an argument is given to justify a law that restricts the personal freedoms of others to make decisions about their bodies and property.[62] Put simply: religious arguments that do not appeal to revelation are welcome in political areas like social justice but not in areas that deal with issues like sexuality or marriage.[63] Thus, he defends a weak inclusivist model for acceptable religious discourse.

In comparison to March, my strong inclusivist model for acceptable religious discourse is more liberal in two respects. First, it does not exclude religious arguments that refer to revelation from public deliberations. Second, religious reasoning is not restricted to specific political areas like social justice. For this

59 Cf. Vallier, *Liberal Politics and Public Faith: Beyond Separation*, 106.

60 Cf. March, "Rethinking Religious Reasons in Public Justification," 523–24, 527, 530.

61 Cf. March, "Rethinking Religious Reasons in Public Justification," 529, 527.

62 Cf. March, "Rethinking Religious Reasons in Public Justification," 532.

63 Cf. March, "Rethinking Religious Reasons in Public Justification," 532–37.

reason, it exemplifies a strong inclusivist position. This more liberal stance of my strong inclusivism has two advantages. First, it imposes fewer integrity costs upon religious citizens (see section 3). Therefore, it is more probable that it will convince inclusivist religious critics of liberalism and thereby help to overcome the impasse between inclusivists and exclusivists. Second, March's weak inclusivist model depends on the plausibility of a couple of classificatory assumptions, for example, that marriage is a political issue that belongs to the political area that deals with the personal freedom to make decisions about one's body and that abortion is a political issue that belongs to the political area that deals with social justice. But why regard abortion as a matter of concern with the basic, uncontroversial interests of persons (including future persons) and not as a matter that has to do with the personal freedom to make decisions about one's body? And why regard the question of same-sex marriage as a matter of personal freedom and not as a political matter that has to do with marriage as a basic social institution?[64] My point is not that March does not justify his classificatory decisions or that his arguments for doing so are bad. My point is simply that his weak inclusivist model of acceptable religious discourse is dependent on his arguments for his classificatory claims, according to which a certain political issue belongs to a certain political category and not another. And since my strong inclusivist model is not dependent on such classificatory issues it is not vulnerable to counterexamples or objections that suggest otherwise. I simply do not have to distinguish between political areas where the use of religious arguments is permitted and those where it is not.

I think March's introduction of the two above-mentioned restrictions for religious reasoning in the liberal public is motivated by the fear that otherwise we end up with the possibility of cases where nonreligious citizens have to endure an objectionable paternalism and are therefore not treated with the respect that we owe each other in a liberal society because it is expected of them that they accept a coercive law as publicly justified and legitimate—especially in very sensitive political areas that involve issues of sexuality—which is justified to them with religious arguments referring to revelation.[65] So, March's weak inclusivist proposal exemplifies the first of the two most promising strategies to justify the exclusion of a subset of religious reasons, namely, those reasons that refer to revelation.

I agree with March that situations like these have to be prevented. It cannot be reasonably expected from nonreligious citizens that they accept a coercive law on the basis of religious reasons that refer to revelation. But I reject March's assumption that such cases can only be prevented if his two principled restrictions

64 Cf. March, "Rethinking Religious Reasons in Public Justification," 533–35.

65 Cf. March, "Rethinking Religious Reasons in Public Justification," 525–30, 532–33, 536–37.

for religious discourse are adopted. First, I do not find plausible March's interpretation of the function of religious arguments that refer to revelation. How could an *appropriately idealized* religious citizen reasonably expect that his arguments, whose soundness depends on the acceptance of revealed knowledge or religious authority and which are therefore by definition not accessible to citizens outside his own religious tradition, provide those citizens with reasons for accepting a coercive law? This simply does not work. I think I am able to propose a more plausible and charitable interpretation of the function of those arguments that connects with what in fact is considered by March himself, namely, that they have the more expressive and prophetic function of "not in my name," i.e., that religious citizens communicate to their fellow nonreligious citizens the reasons for their stance on a political issue, knowing that these reasons are not reasons that have justificatory weight for their fellow nonreligious citizens.[66] If one understands the function of religious arguments in this way—as I do—there is nothing inherently authoritarian, theocratic, paternalistic, disrespectful, demeaning, or humiliating about it if religious citizens use those arguments to justify their preference for a political decision from their first-personal epistemic perspective and their particular evaluative standards.[67] Second, I have shown in this article that something like the two principled restrictions for religious discourse that March proposes are not necessary in order to prevent cases where a nonreligious minority is forced to accept a coercive law imposed by a religious majority without justifying it publicly to the minority with reasons that are accessible to them as having justificatory weight. According to my strong inclusivist position and the second-personal conception of public justification on which it is based, public justification does not require that a law is justified with reasons that are accessible to all citizens from a third-personal standpoint, i.e., with reasons that have justificatory weight for all citizens. According to my second-personal account of public justification, it is just required that a law is justified to each citizen with considerations he or she can access as being justificatorily relevant and constituting a weighty reason according to his or her first-personal epistemic standpoint and particular evaluative standards. In short: what rules out the problematic cases is the requirement to exercise moral restraint that is built into the second-personal conception of public justification and that does not require the exercise of epistemic restraint. And since the exercise of epistemic restraint is not required, even religious reasons that refer to revelation can play a role in the public justification of a law because a law can be publicly justified through a convergence of mutually inaccessible reasons. The only case that is ruled out by my strong inclusivist

66 Cf. March, "Rethinking Religious Reasons in Public Justification," 528.

67 Cf. March, "Rethinking Religious Reasons in Public Justification," 529–30, 534–35.

model—given its acceptance of the fact of a reasonable pluralism and its requirement to exercise moral restraint—is that a law in a liberal society can be justified *solely* with religious reasons that refer to revelation. What my model demands from religious citizens is not that they cease to justify their political preferences with arguments that in principle are inaccessible to their fellow citizens but that they are willing to make this kind of reasoning *intelligible* to those who do not share their religious worldview and that they are willing to show that although their reasoning is derived from resources that are inaccessible without faith it is not against reason to reason in this way. But if religious citizens exercise moral restraint and address their fellow citizens with considerations they can accept as weighty reasons from their different epistemic standpoints, there is nothing disrespectful or paternalistic about a law that is publicly justified through a convergence of mutually inaccessible reasons—religious reasons that refer to revelation included—and about which a democratic procedure of decision-making decides whether it gets implemented or not.

5.3. The Difference from Aurélia Bardon's Weak Inclusivist Proposal for Acceptable Religious Discourse in the Liberal Public

Like Andrew March, Aurélia Bardon distinguishes between different kinds of religious arguments in order to exclude some from the process of the public justification of political decisions. And similar to March's proposal, the religious reasons that should be excluded are arguments that refer to revelation, i.e., arguments that operate with premises whose truth is in principle inaccessible to nonbelievers.[68] But Bardon's weak inclusivism is different from March's because she offers a different rationale for why at least agents in the political sphere like politicians should refrain from using such arguments to justify their preferred political options.[69] Her paradigmatic example for the kind of religious argument that should be excluded from the process of public justification is from John Locke, who justifies his support for the redistribution of wealth in favor of poor citizens with pressing needs with the religious argument that God has created the world and all its goods for the sustenance of all persons.[70] Thus, a policy tool like a wealth tax with which the state redistributes money from rich citizens to poor citizens in dire need is justified with the religious consideration that God is the creator of all goods and that it is His will that these goods are used in order to sustain all human beings. Consequently, the right to an accumulation of goods and their private use is not absolute but is relativized by the right of people to

68 Cf. Bardon, "Religious Argument and Public Justification," 274, 283–84.

69 Cf. Bardon, "Religious Argument and Public Justification," 288.

70 Cf. Bardon, "Religious Argument and Public Justification," 285–86.

make use of those goods in accordance with God's will in order to sustain their lives in situations of dire need. From Locke's religious, first-personal, epistemic standpoint, such a right derives from the fact that God as the creator of these goods has created them with a certain purpose and that human beings just exercise a kind of stewardship over these goods that they practice with some liberty as long as it accords with the purpose with which God has created them.

Now, what is wrong with such an argument? Why should it be excluded from the process of the public justification of a wealth tax? According to Bardon, Locke's religious argument should be excluded from such a political discussion because it is only a good argument for religious citizens but not for nonreligious citizens.[71] Put another way: it should be excluded because it is not accessible to nonreligious citizens as an argument that has justificatory weight for them.[72] Without the acceptance of Locke's religious evaluative standards, a nonreligious citizen cannot recognize that Locke's religious considerations constitute a reason for her that justifies the imposition of a wealth tax. But what cannot be reasonably expected of a nonreligious citizen is that she accepts the religious evaluative standards of her fellow religious citizen as justificatorily relevant for the issue at hand.

So, unlike March's proposal for acceptable religious discourse that exemplifies the strategy to justify the exclusion of certain religious arguments on the grounds that they violate certain moral obligations such as respect that we owe to each other in a liberal society, Bardon's proposal exemplifies the strategy to justify the exclusion of certain religious reasons on the basis of an epistemic criterion, namely, their lack of accessibility. Her argument runs like this: since a nonreligious citizen does not share the belief that God is the creator of all goods and that He has created these goods with the purpose to sustain all human beings, a reference to God or sacred texts like the Bible that are supposed to reveal His will are justificatorily irrelevant for her in order to evaluate and determine whether wealth should be redistributed by means of a wealth tax or whether poor citizens have the right to make use of accumulated goods of rich citizens in order to sustain their lives in situations of dire need. A nonreligious citizen cannot evaluate Locke's reason as good or bad because as a nonreligious citizen she does not share the relevant religious evaluative standards. Thus, Bardon's rationale for the exclusion of religious arguments referring to evaluative standards that are in principle inaccessible to nonreligious citizens is very similar to Quong's: the use of such arguments in a political discussion is problematic and dangerous because it leads to foundational disagreements, i.e., disagreements

71 Cf. Bardon, "Religious Argument and Public Justification," 286.

72 Cf. Bardon, "Religious Argument and Public Justification," 284.

where there is no shared normative framework, no deeper standard of justification that could serve as the basis for adjudicating the dispute.[73] According to Bardon, such foundational disagreements are dangerous because they can result in the imposition of evaluative standards that are not part of a shared liberal normative framework because they are too fundamental to become the object of any compromise, negotiation, or argumentation, with the consequence that the political discussion depending on the possibility to question, review, and criticize arguments breaks down.[74] In short: Bardon argues that religious arguments that refer to revelation should be excluded from public justification because their inclusion would lead to situations where it is expected of nonreligious citizens to accept religious evaluative standards as justificatorily relevant for them.

According to my proposal for acceptable religious discourse, this is not expected of nonreligious citizens. On the contrary, I fully acknowledge that religious arguments that refer to revelation are by their very nature inaccessible to nonreligious citizens and that they therefore have no justificatory weight for them. Nevertheless, I disagree with Bardon as well as with Quong that this fact justifies their exclusion from the process of the public justification of a political decision. I have shown in detail with the example of the foundational disagreement between Mike and Sara (see section 4) how one can deal with such situations in a way that accords with the requirements of moral restraint.

Foundational disagreements, i.e., disagreements that are not only about the justificatory weight of evaluative standards and the conclusions that derive from these premises but also about the justificatory relevance of the evaluative standards itself are only problematic if one is committed to a third-personal conception of public justification. According to such a conception, the public justification of a political decision with reasons not justificatorily relevant for all parties is problematic because public justification demands that a political decision is justified with reasons that are mutually accessible to the appropriately idealized members of the public from a third-personal standpoint. But such foundational disagreements are not problematic according to the second-personal conception of public justification I have advanced in this article because a political decision can be justified through a *convergence* of mutually inaccessible reasons. So, according to my model of an acceptable religious discourse, something like Locke's religious argument for a wealth tax does not have to be excluded from the political discussion. The imposition of a wealth tax through a democratic procedure is publicly justified, iff the following three conditions are fulfilled. First, an appropriately idealized version of Locke who wants to

73 Cf. Bardon, "Religious Argument and Public Justification," 284–85.

74 Cf. Bardon, "Religious Argument and Public Justification," 284, 285, 287.

impose a wealth tax WT on the group of appropriately idealized rich citizens *B*, gives *B* a sincere and honest justification of WT, which means he in a sincere and honest way states publicly the considerations that motivate him to the support the imposition of WT on *B*. In the case at hand, these considerations are his religious considerations RC that he believes that God has created all goods with the purpose to sustain all human beings. Second, my model requires that it is intelligible for *B* that Locke is justified from his own first-personal epistemic standpoint that includes certain religious evaluative standards RES (e.g., living in accordance with God's will as it is revealed in the Bible) to believe that RC justifies WT. Contrary to Bardon, I think there is no principled problem for *B* to understand that there is a logical relation between RES, RC, and WT from Locke's first-personal epistemic standpoint.[75] *B* can for example have a look at the Bible, a catechism, or a book on theology in order find out whether there is a logical relation between Locke's religious beliefs and his support for a wealth tax, i.e., if Locke's religious reasoning is sound from the epistemic perspective and the evaluative standards Locke himself is committed to. Intelligibility requires that the arguments religious citizens use for the justification of political norms have to be formulated in such a way that appropriately idealized nonreligious citizens should be able to track the soundness of the argument if they adopt the epistemic standpoint of their fellow religious citizens. Intelligibility does not require that they have to accept the truth of the religious presuppositions—for example, revealed truths—but only that they should be able to acknowledge that these arguments are sound for someone who does accept their truth. For this reason, it is not expected of nonreligious citizens to accept religious evaluative standards as justificatorily relevant for them.

The third and last condition that needs to be fulfilled to justify publicly the imposition of a wealth tax through a democratic procedure is that Locke gives *B* a consideration CB that *B* can access as a weighty reason that justifies WT according to *B*'s evaluative standards ESB that could be, for example, "One should do what promotes and secures one's wealth." Now, Bardon herself acknowledges that this last condition should also not be very difficult to fulfill.[76] For example, Locke could argue that, without a certain redistribution of wealth from the rich to the poor, economic inequality and mass poverty reaches a point where a democratic society becomes instable and where a great mass of impoverished citizens is inclined to follow radical populist parties that promise to take wealth from the rich by force. So, Locke could argue that even from *B*'s nonreligious epistemic perspective and by its own evaluative standards ESB, there is a consideration that

75 Cf. Bardon, "Religious Argument and Public Justification," 286.

76 Bardon, "Religious Argument and Public Justification."

constitutes a weighty reason for *B* to support WT. This shows that what is dangerous is not so much the reality of foundational disagreements in a democratic society but rather a lack of willingness or capacity to adopt a different epistemic first-personal perspective than one's own. Again, all that is required to exclude what exclusivists most fear about a very liberal stance on religious reasoning in the public is to demand from religious citizens what is demanded of all citizens, namely, that they are committed to liberal core values like freedom and equality and the principle of moral restraint that derives from these values, i.e., that they refrain from advancing and enforcing a demand if they cannot justify this demand to their fellow nonreligious citizens by showing them that they have a weighty reason to comply with this demand. Thus, my model of acceptable religious discourse makes a difference to proposals of weak inclusivists like Bardon and March insofar as it shows that even religious arguments that refer to revelation can figure into the public justification of political decisions without making those cases possible that exclusivists most fear, namely, that a political measure counts as publicly justified without it being the case that each citizen is given a weighty reason to comply with this measure.[77]

Munich School of Philosophy
patrick.zoll@hfph.de

REFERENCES

Audi, Robert. *Religious Commitment and Secular Reason*. Cambridge: Cambridge University Press, 2000.

Audi, Robert, and Nicholas Wolterstorff. *Religion in the Public Square: The Place of Religious Convictions in Political Debate*. Lanham, MD: Rowman and Littlefield, 1997.

Bardon, Aurélia. "Religious Argument and Public Justification." In *Religion, Secularism and Constitutional Democracy*, edited by Jean L. Cohen and Cécile Laborde, 273–92. New York: Columbia University Press, 2016.

Billingham, Paul. "Convergence Justifications within Political Liberalism: A Defense." *Res Publica* 22, no. 2 (May 2016): 135–53.

Breul, Martin. *Religion in der politischen Öffentlichkeit: Zum Verhältnis von re-*

77 I would like to thank three anonymous reviewers, Martin Breul, Godehard Brüntrup, Chad Flanders, Ralf Klein, Felix Körner, James Dominic Rooney, and Eleonore Stump for many helpful and valuable comments on previous versions of this article.

ligiösen Überzeugungen und öffentlicher Rechtfertigung. Paderborn: Ferdinand Schöningh, 2015.

D'Agostino, Fred. *Free Public Reason: Making It Up as We Go*. New York: Oxford University Press, 1996.

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

Eberle, Christopher J. "Basic Human Worth and Religious Restraint." *Philosophy and Social Criticism* 35, nos. 1–2 (2009): 151–81.

———. *Religious Conviction in Liberal Politics*. Cambridge: Cambridge University Press, 2002.

Eilan, Naomi, ed. *The Second Person: Philosophical and Psychological Perspectives*. New York: Routledge, 2015.

Feinberg, Joel. *Harm to Others*. New York: Oxford University Press, 1984.

Fowler, Timothy, and Zofia Stemplowska. "The Asymmetry Objection Rides Again: On the Nature and Significance of Justificatory Disagreement." *Journal of Applied Philosophy* 32, no. 2 (2015): 133–46.

Gaus, Gerald F. *Justificatory Liberalism: An Essay on Epistemology and Political Theory*. Oxford: Oxford University Press, 1996.

———. *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*. Cambridge: Cambridge University Press, 2011.

———. "The Place of Religious Belief in Public Reason Liberalism." In *Multiculturalism and Moral Conflict*, edited by Dimova-Cookson and Peter M. R. Stirk, 19–37. London: Routledge, 2010.

Gaus, Gerald F., and Kevin Vallier. "The Roles of Religious Conviction in a Publicly Justified Polity: The Implications of Convergence, Asymmetry and Political Institutions." *Philosophy and Social Criticism* 35, nos. 1–2 (January/February 2009): 51–76.

MacIntyre, Alasdair. *Whose Justice? Which Rationality?* 3rd ed. Notre Dame, IN: Notre Dame Press, 2003.

March, Andrew F. "Rethinking Religious Reasons in Public Justification." *American Political Science Review* 107, no. 3 (August 2013): 523–39.

Neal, Patrick. "Is Political Liberalism Hostile to Religion?" In *Reflections on Rawls*, edited by Shaun P. Young, 153–56. Burlington, VT: Ashgate, 2009.

Pinsent, Andrew. *The Second-Person Perspective in Aquinas's Ethics: Virtues and Gifts*. New York: Routledge, 2012.

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

Stout, Jeffrey. *Democracy and Tradition*. Princeton, NJ: Princeton University Press, 2004.

Talisse, Robert B. *Democracy and Moral Conflict*. Cambridge: Cambridge University Press, 2009.

Vallier, Kevin. "Convergence and Consensus in Public Reason." *Public Affairs Quarterly* 25, no. 4 (October 2011): 261–79.

———. "In Defense of the Asymmetric Convergence Model of Public Justification: A Reply to Boettcher." *Ethical Theory and Moral Practice* 19, no. 1 (February 2016): 255–66.

———. *Liberal Politics and Public Faith: Beyond Separation*. New York: Routledge, 2014.

———. *Liberal Politics and Public Faith: A Philosophical Reconciliation*. Ann Arbor, MI: ProQuest, UMI Dissertation Publishing, 2012.

Wall, Steven. *Liberalism, Perfectionism and Restraint*. Cambridge: Cambridge University Press, 1998.

———. "On Justificatory Liberalism." *Politics, Philosophy and Economics* 9, no. 2 (May 2010): 123–49.

———. "Perfectionism in Politics: A Defense." In *Contemporary Debates in Political Philosophy*, edited by Thomas Christiano and John Philip Christman, 99–117. Oxford: Wiley-Blackwell, 2009.

Weithman, Paul J. *Religion and the Obligations of Citizenship*. Cambridge: Cambridge University Press, 2002.

Zoll, Patrick. *Perfektionistischer Liberalismus: Warum Neutralität ein falsches Ideal in der Politikbegründung ist*. Freiburg im Breisgau: Alber Verlag, 2016.

Journal of Ethics and Social Philosophy
VOL. 20, NO. 3 · NOVEMBER 2021
https://doi.org/10.26556/jesp.v20i3.948

DISAGREEMENT, UNILATERAL JUDGMENT, AND KANT'S ARGUMENT FOR RULE BY LAW

Daniel Koltonski

IT IS A COMMON THOUGHT that authoritative law is necessary because we disagree about justice. This idea often rests on law's purported instrumental value, on its ability to get us, imperfect and biased agents, closest to a just society: we do best, from the perspective of justice independently defined, by having clear legal rules to follow and rights to respect. In *The Doctrine of Right*, Kant rejects such an instrumental conception of law and instead defends the more controversial claim that, absent authoritative law, there will often be no answer to be had about what justice (or, for Kant, right) requires of us in our interactions with one another. On this view, in a situation without authoritative law—in a state of nature—a person is unable coherently to pursue the aim of acting rightly. Authoritative law is required for Kant, then, not because a person, in obeying the law, is thereby more likely to do what right demands; rather, it is required because without it, there will often be no sense to be made of this question of what right demands.

The problem with the state of nature, according to Kant, is that it is "a state *devoid of justice* . . . , in which when rights are *in dispute* . . . there would be no judge competent to render a verdict having rightful force."[1] Kant argues:

> However [good and right-loving] human beings might be, it still lies a priori in the rational idea of such a condition (one that is not rightful) that before a public lawful condition is established, individual human beings . . . can never be secure against violence from one another, since each has its own right to do *what seems right and good to it* and not to be dependent upon another's opinion about this.[2]

1 Kant, *The Metaphysics of Morals*, 6:312. All citations to Kant are to the Akademie numbers listed in the margins of most editions; unless otherwise stated, all English translations of Kant's *Doctrine of Right* are from *Practical Philosophy*, translated and edited by Mary Gregor.

2 Kant, *The Metaphysics of Morals*, 6:312. Translation in brackets from Ripstein, *Force and Freedom*, 146.

He continues:

> So, unless it wants to renounce any concepts of right, the first thing it has to resolve upon is the principle that it must leave the state of nature, in which each follows its own judgment, unite itself with all others (with which it cannot avoid interacting) … and so enter into a condition in which what is to be recognized as belonging to it is determined *by law* and is allotted to it by adequate *power* … that is, it ought above all else to enter a civil condition.[3]

As the literature on this argument makes clear, the basic difficulty with the state of nature is not that persons, even entirely "good and right-loving" ones, cannot be secure from violence—though this is an important consequence—but rather that they cannot respect one another's equal freedom when they act on their own judgments of right in circumstances of disagreement among them. The idea here is that when someone interacts with another according to her own judgment of right, she implicitly claims that her judgment governs this interaction—it binds them both—while his conflicting judgment does not. And he implicitly claims something similar when he interacts with her according to his judgment of right. They each implicitly claim a power to bind the other that they deny the other has to bind them, a claim inconsistent with "innate equality," itself an aspect of our freedom: "independence from being bound by others to more than one can in turn bind them."[4] Kantians call this "the problem of unilateralism": in the state of nature, acting on your own judgment of right contains within it the claim that your will is unilaterally lawgiving for those with whom you interact.[5]

As one quite natural understanding of this problem of unilateralism has it, you judge that *x*, she judges that not-*x*, and for you to take your judgment to be what governs this interaction is for you to treat as special the fact that *you* judge that *x*. You treat your judgment of right as having authority over the interaction, and so also over her, simply because the judgment is yours; in so doing, you deny that same authority to her differing judgment in violation of innate equality. This understanding of the problem, however, faces the objection that it misunderstands the reason you act on when you act on your own judgment in

3 Kant, *The Metaphysics of Morals*, 6:312.

4 Kant, *The Metaphysics of Morals*, 6:237.

5 Persons' lack of security from violence in the state of nature is thus the result of the absence of a mechanism for resolving for them their inevitable disagreements about right. When such disagreements arise, each will act on their own judgment of the matter, standing up to others in defense of what they take right to be. Everyone will thus be subject to coercive threat—as doing only what you judge consistent with right will be no protection from others who may disagree—and, in that way, no one will be secure from violence.

circumstances of disagreement. Your judgment governs this interaction, on your view, not because it is yours but rather because it is correct. Or, as David Enoch puts it, "your reason for action is not that you believe so-and-so (where others believe otherwise), but rather that so-and-so."[6] Here, then, "it's just not about you at all. So, there is no sense in which you're giving extra weight to your beliefs over others—you're giving no weight to your beliefs here."[7] And so, in acting on your judgment in the state of nature, imposing it on others who disagree, you are not thereby claiming a special power to bind them in violation of innate equality. Call this *Enoch's Objection*.[8]

My aim here is to defend the Kantian account of the problem of unilateralism against Enoch's Objection and, in so doing, to illuminate the feature of the Kantian conception of right that accounts for why, however "good and right-loving" they might be, persons in a state of nature about right are unable coherently to pursue the aim of acting rightly. Notably, this problem of unilateralism is not unique to the domain of right but rather arises more generally, and less controversially, in state-of-nature versions of other rule-governed interactions. The case against Enoch's Objection thus begins by arguing that when it comes to state-of-nature versions of these other interactions, the objection fails. The focus here will be on certain multiplayer games and, in particular, on those games whose rules give players the freedom to decide for themselves what moves to play on a field of play that they share, with chess and baseball as paradigm cases. In these games, players have a right of self-governance: the right to choose for themselves which move, among the legitimate moves open to them, to play. The difficulty is that some of these games would be unplayable were the game's rules to give players a second right of self-governance: the right to act on their own judgments of which moves, according to the game's conduct rules, are legitimate in the first place. Giving players this second right would be to put them in a state of nature with one another about the game's conduct rules, and so the claim is that some of these games are unplayable as state-of-nature games.

6 Enoch, "Not Just a Truthometer," 982. The notion of a reason here is "the agent's reason," or "the consideration in light of which the agent acted, the feature of the situation that made the relevant action one the agent thought worth performing." See Enoch, *Taking Morality Seriously,* 221–23.

7 Enoch, "Against Public Reason," 131. As Enoch notes, Joseph Raz makes this same general point in his "Disagreement in Politics."

8 Enoch offers this objection specifically against what he calls "public-reason accounts." His specific targets are John Rawls's *Political Liberalism* (Enoch, "Against Public Reason," 130–34) and Gerald Gaus's *The Order of Public Reason* (Enoch, "The Disorder of Public Reason," 156–60). But, as Thomas Sinclair notes, this objection can also be directed at Kant's problem of unilateralism. See Sinclair, "The Power of Public Positions," 31.

The basic issue with state-of-nature versions of these games is that players can responsibly disagree about what the conduct rules say about the legitimacy of some move. When such disagreement is possible, there will not be one status for the move—either legitimate or not—that they all are accountable for recognizing as the status it has within their game. The result is that the status of this move in their game, and so the status that governs (or has authority over) their individual game play, is indeterminate. (While what the conduct rules say about the move may be determinate, it may not be *accessible* to these players.) When a player nevertheless exercises her rights of self-governance in this context of responsible disagreement, playing a move she responsibly judges that the conduct rules say is legitimate, she cannot help but implicitly claim that her judgment has decided for their game the move's status within it. She thus claims for herself and her judgment unilateral authority over the game and so over her fellow players—authority she denies to them and their responsible judgments. Her judgment, she claims, resolves the indeterminacy by *making* the move legitimate in their game. Directed at state-of-nature versions of these games, Enoch's Objection thus fails, for this player is claiming a special power to bind the others in violation of innate equality: in exercising her second right to self-governance in this state-of-nature game, she implicitly denies that her fellows have that same second right. The result is that she is unable to exercise both rights responsibly in a state-of-nature game while understanding it as such—that is, as a game in which everyone, and not just she, has both rights.

After showing that Enoch's Objection fails when applied to state-of-nature versions of these games, I argue that on a Kantian conception of right, the rules of right are relevantly analogous to the rules of these games. It is as if the rules of right place us, along with those with whom we "cannot avoid interacting," into a multiplayer "game" of equal freedom, one whose players, even when idealized as entirely good and right-loving, can responsibly disagree about which "moves" the conduct rules of right declare to be legitimate or not. As a result, Enoch's Objection similarly fails when it comes to a state of nature about right (or when it comes to a state-of-nature "game" of equal freedom). The Kantian account of the problem of unilateralism in a state of nature about right does not make the mistake that the objection claims.

1. THE INDETERMINACY-OF-RIGHT RESPONSE

One of Kant's central claims is that right is, in some important sense, indeterminate.[9] We might try to avoid Enoch's Objection, then, by understanding the

9 See, for instance, Kant, *The Metaphysics of Morals*, 6:266.

disagreements about right at issue as instead disputes about how to resolve some indeterminacy of right itself. This indeterminacy claim would give us a version of the problem of unilateralism: by acting on my own judgment when right itself is indeterminate, I am thereby claiming the power to make my answer about right *the* answer for all of us, a claim of unilateral authority incompatible with innate equality. Recent defenders of the Kantian view have either offered this indeterminacy response or presented an account of the indeterminacy of right that makes such a response possible. Thomas Sinclair, for instance, does the former and Arthur Ripstein, the latter.

For Ripstein, the problem lies in the application of concepts of right: "The application of concepts to particulars is always potentially indeterminate, and so requires judgment, as a result of which the classification of particulars is always, at least in principle, indeterminate."[10] This indeterminacy is a source of disagreement about right:

> There are some cases in which concepts of right completely determine the outcome of a dispute.... In other cases, however, even if it is agreed that concepts of right apply, there can be a dispute about how they apply to particular cases.... Although their internal structure requires a single answer, neither the normative concepts nor the relevant facts nor any combination of them guarantees agreement.[11]

These disagreements are a problem, according to Ripstein, because no one need accept another's answer:

> If I believe in good faith that the boundary between our property is in one place, and you, equally in good faith, believe that it is somewhere else, neither of us has any obligation of right to yield to the other.... More generally, neither of us needs to give in to the unilateral judgment of the other as to how to classify particulars.[12]

Thomas Sinclair understands disagreement about right similarly, as having its source in the "inevitable indeterminacy in the application of *any* general principle [including principles of right] to concrete particulars."[13] For example:

> You and I might agree on the authoritativeness of a law that says clamshells on the beach are mine and clamshells in the sea are yours, and yet

10 Ripstein, *Force and Freedom*, 170.

11 Ripstein, *Force and Freedom*, 170.

12 Ripstein, *Force and Freedom*, 172.

13 Sinclair, "The Power of Public Positions," 33.

> disagree about this clamshell, which is being moved back and forth by waves on the beach. . . . If my claim were authoritative, then you (and everyone else) would be subject to constraints privileging my judgment, but I would not be subject to constraints privileging yours.[14]

Both accounts thus hold that were I to regard my judgment as resolving the indeterminacy of right at issue, I would be treating the fact that it is *my* judgment as special, as what gives it the requisite authority.

One might reasonably wonder whether this indeterminacy response succeeds. In Ripstein's example, the judgments we each make concern where the boundary lies: I believe it is here; you believe it is there. It would seem, then, that I believe that I am correct and you are mistaken, while you believe the opposite. But if so, we are both thus presuming about this question of the boundary that there is a determinate answer, one that we each think we have gotten right. Ripstein's account thus seems to face something like a dilemma. If there is no such answer—if, as Ripstein argues, right itself is indeterminate here—then it would be a mistake to exercise judgment such that one arrives at, as Ripstein's account has it, a good-faith belief about where the boundary lies. Or if this is not a mistake—if it is somehow appropriate to form good-faith beliefs about where the boundary lies in a case like this—then this response does not avoid Enoch's Objection, for it will still be that when I act on my judgment in a case of disagreement, thereby imposing it on you, my reason is of the form "that *x*" ("that the boundary is here" or "that the boundary is there") not "that I judge that *x*" or "that I believe that *x*." Granted, if right is indeterminate in these cases, then I will in fact be imposing my answer by acting on my judgment. But I will not thereby be implicitly claiming the unilateral power to make right determinate, for the way I approach the question of right at issue—as one I aim to make the correct judgment about—rules out this implicit claim.

The indeterminacy-of-right response may be able to counter this objection. It is not clear, for instance, whether Sinclair's version of the response is in fact vulnerable to the objection. That will depend, I think, on whether there is available a plausible account of judgment such that my judgment that I own the clamshell does not bring with it a claim of correctness. If there is not, then the authority of my judgment will not, on my view, come from me but from the rules themselves, and Enoch's Objection will still apply. Regardless, a response to Enoch's Objection that depends on the claim that right itself is often indeterminate seems of limited use dialectically, for many of those who reject the Kantian account of the problem of unilateralism for the reasons articulated by Enoch's Objection

14 Sinclair, "The Power of Public Positions," 32.

will likely also reject this claim that right itself is often indeterminate.[15] What we require, it seems to me, is a different response to Enoch's Objection, one that neither relies on nor denies this claim that right itself is often indeterminate, but instead remains neutral with regard to it. Such a response is what I develop in what follows. (This response still holds that there is an important sense in which right is indeterminate, just not in the way that Ripstein and Sinclair claim.)

2. DISAGREEMENT AND GAMES

The defense of this different response to Enoch's Objection is somewhat indirect. The argument is first that the Kantian problem of unilateralism is not specific to a state of nature about right but instead arises more generally, and less controversially, in state-of-nature versions of other rule-governed interactions. After showing, in this section, that when it comes to state-of-nature versions of these other interactions, Enoch's Objection fails, I then proceed, in the next section, to show that on a Kantian conception of right, the rules of right are relevantly analogous to the rules of these other interactions. And because Enoch's Objection fails when it comes to state-of-nature versions of these other interactions, it similarly fails when it comes to a state of nature about right. The Kantian account of the problem of unilateralism in a state of nature about right thus does not make the mistake that Enoch's Objection claims.

The focus in this section will be on certain multiplayer games, with chess and baseball as paradigm cases. In the games at issue, players share a field of play, and when it is a player's turn, there are normally multiple moves among which she may choose according to her own view of her ends within the game and how to pursue them. (Winning the game may not be a player's only or even primary end. Indeed, as it is not an obligatory end, it may not be one of her ends at all.) The rules of these games thus give players a right to self-governance:

> *Right to Self-Governance 1* (RSG1): The right to act within the game according to one's own judgments (rather than deferring to some other's judgments) of which moves, among the legitimate ones, to play.

This section argues that these games must also be structured such that players stand in relationships of accountability with one another about the rules.

For a subset of these games, however, players will not stand in such relation-

15 And this includes Enoch: "Indeterminacy can perhaps play some role in accounting for moral disagreement, but not the key role some thinkers attribute to it" (*Taking Morality Seriously*, 192n20).

ships of accountability if the game also gives them a second right to self-governance:

> *Right to Self-Governance 2* (RSG2): The right to act within the game according to one's own judgments (rather than deferring to some other's judgments) of which moves, according to the game's conduct rules, are legitimate in the first place.

Why is this? The possibility of a certain kind of disagreement about how the conduct rules apply and so about which moves are legitimate—responsible disagreement between players—makes these relationships of accountability impossible when players have not only the first but also the second right to self-governance. Baseball is one such game.

What these games require, then, is some authoritative mechanism that resolves these responsible disagreements such that within the game, moves have one status (i.e., either legitimate or not) that players are to recognize as governing all players, as this will make relationships of accountability between them possible. That the game requires such a mechanism, and so cannot be played as a state-of-nature game, is what would give rise to the problem of unilateralism were players nevertheless to attempt it as a state-of-nature game: when one player exercises her RSG2 and so acts in the game according to her own responsible judgment of the legitimacy of some move, a judgment with which other players might responsibly disagree, she cannot help but implicitly claim, in this instance, to be the authoritative mechanism, required by the game, that decides the legitimacy of moves. This is a claim of unilateral authority over the game and so over her fellow players.

2.1. Freedom and Accountability within Certain Multiplayer Games

A player's exercise of the first right of self-governance (RSG1) is, of course, still governed by the game's rules. When pursuing her ends within the game, she may do so only within the bounds set by those rules: for any turn *t* in the game, she may choose only among those moves that are legitimate at *t*. In this way, her RSG1 is both the right within the game to pursue her ends in her own way and the responsibility for doing so only within whatever bounds the rules, at any *t*, give to that right. We can understand these bounds as *the state of play* within the game. The state of play at some *t* is, roughly, where things stand in the game at *t*—which moves at *t* are legitimate (or not) and for whom—and it is the product not only of the rules themselves but also of what has happened in the game up to *t*, the legitimate moves that have already been played and, as a result, shape the field of play at *t*. In this way, the state of play at any *t* in the game is normative, for it tells

players what at *t* they are and are not permitted to do next. To say that a player's game play is governed by the rules, and so that she is responsible for abiding by those rules, is thus to say that when pursuing her ends in her own way within the game, she is responsible for doing so using only those moves the current state of play has as legitimate for her.

Because the state of play at any *t* is, in part, the product of whatever has happened in the game up to *t*, a player's choice of moves, in a game where players share a field of play, is not merely a choice of how to pursue her ends but also a choice of the state of play that will result in that because it will partly determine which moves are subsequently legitimate (or not) for other players on the field of play. There is thus a sort of interdependence between players' freedom within a game in which they share a field of play—a player's pursuit of her ends within the game shapes the bounds the rules subsequently give to her fellows' space for their pursuit of their ends, and vice versa—and so the players are in this way governed *together* by the rules. One result is that a player's RSG1 contains, as it were, an additional right:

> *Right to a Legitimate Field of Play*: The right that, at any turn *t*, the moves that are legitimate for one to play are indeed playable on the field of play.

Put another way, a player has the right that her fellow players, in their turns shaping the field of play for her, play only moves that the state of play has as legitimate for them. This right to a legitimate field of play, as part of a player's RSG1, thus correlates with her fellow players' responsibility (or duty), as part of their RSG1, to choose only among legitimate moves. The result is not only that players are responsible for choosing only among moves that are legitimate for them but also that they are accountable to one another for doing so: other players have the standing to demand, as something owed to them, that one fulfill this responsibility. And so, because they are governed together by the rules in this way, players stand in relationships of accountability with one another about their exercise of their RSG1. Or at least this is what a game structured in this way, with players each having (and exercising) the RSG1 on a field of play that they share, commits itself to.

But if players are to stand in these relationships of accountability, the game must be structured so as to make such relationships possible. And they will be possible between players, it seems to me, only so long as the state of play at any *t* is accessible to them. By "accessible" I mean that, barring special (and unfavorable) circumstances, it is the case that players exercising judgment responsibly are able to identify the state of play as such. The basic idea is that a game that gives players the RSG1, if it is to be playable, must also see to it that at least when

players exercise judgment responsibly in whatever the game presupposes as "ordinary" (or "normal" or "standard") circumstances of game play, these players can fulfill the responsibility contained within their RSG1, one they owe to each other, of playing only legitimate moves. (Special, and unfavorable, circumstances can thus excuse a player's failure to fulfill this responsibility.)[16]

Why is it that players can be accountable to one another in the way these games require only if they themselves can identify the state of play within the game, whatever it is, as such? It is because the claims the game licenses players to make of each other within the game must be compatible with the others' exercise of their RSG1. Suppose one player makes a demand of another about some attempted move: "But you can't do that!" If her demand of him is to be compatible with his RSG1, it cannot be a demand that he defer to her judgment that his move is illegitimate; it must instead be a demand that he recognize its illegitimacy himself, and as such, it presupposes that he can do so, at least in ordinary circumstances, by exercising judgment responsibly himself.[17] Thus, if she is to be able to make such demands of him, ones her RSG1 entitles her to make, it must be the case that the illegitimacy of his move is accessible not only to her but also to him. The result is thus that players will stand in relationships of accountability with one another only if, for any move played, its status as legitimate or not in the state of play is something that, provided they exercise judgment responsibly, players are able to identify themselves. All of this follows from the game giving players each an RSG1 that they are to exercise on a field of play they share.

2.2. The Case of Chess

We can see this at work in a game of chess. In chess, each player has this RSG1: during her turn, she may choose for herself which move, among the legitimate moves, to play in pursuit of her ends. (A player's primary end, of course, may be to win. Even so, she may have several different strategies to choose among, or she may wish to win but, for whatever reason, not too quickly. Or, as winning is not an obligatory end, her end may be something else: to let her opponent win, to

16 In what follows, I do not explicitly include this proviso about valid excuses, but solely for convenience; its presence should be understood as implied throughout.

17 This claim about accountability in these games is related to a claim Stephen Darwall makes about moral obligation and accountability:

> If . . . you address a putatively authoritative demand to someone to get off your foot and hold him answerable for doing so, you do assume, do you not, that this is something he should be able to see for himself, or at least to appreciate when it is pointed out to him? After all, how can you hold him responsible for doing something for reasons he cannot himself appreciate even when they are pointed out to him? ("Law and the Second-Person Standpoint," 174)

practice a new and complicated strategy of play, to help her opponent practice it, etc.) Once one player exercises her RSG1, her choice of moves is incorporated into the resulting state of play, and straightforwardly so: her move changes the field of play—the arrangement, and perhaps number, of pieces on the board—and in so doing, it partly determines the moves that are now open (or not) to her opponent for his pursuit of his ends.[18] Her exercise of her RSG1 thus shapes the bounds the rules impose on his exercise of his RSG1 (and vice versa)—the rules govern them together—and so the game has it that they are accountable to one another for choosing only among the moves the state of play has as legitimate for them. ("You can't take my queen like that! This pawn is in the way.")

In fact, the rules of chess give players not just the first but also the second right of self-governance (RSG2): they may act within the game according to their own judgments of which moves, according to the game's conduct rules, are legitimate in the first place. Chess is thus structured so that players are *entirely* self-governing. Or, to put it another way, chess puts its players in a Kantian state of nature with one another about its conduct rules: "each has [their] own right to do what seems right and good" to them within the game rather than defer to another.

That chess puts players in a state of nature about its conduct rules does not, however, lead to the problem of unilateralism Kant claims to find in a state of nature about right. It is important to be precise about why. When players exercise their RSG2, the judgments they act on are their judgments of what the conduct rules say the state of play is. This means that by giving players this RSG2, the game defines the state of play at any *t* as simply what the conduct rules say that it is at that *t*. But what the rules say the state of play is can fulfill the role of the state of play in the game only if what they say is accessible to players—that is, only if, barring special (and unfavorable) circumstances, responsibly exercising this RSG2 will lead them to the correct answer as to what the conduct rules say the state of play is. Otherwise, players will not be accountable to one another for choosing only among the moves the state of play has as legitimate. And in chess the conduct rules are indeed such that what they say about the legitimacy of any move, whether actual or possible, is accessible to players. What is presupposed by a demand addressed by one player to another (that the other player is able, via her own judgments of what the conduct rules say, to recognize the authority of the one's claim about the legitimacy of some move) obtains in a game of chess. The one player is able to make this demand of the other, as it were, on behalf of

18 That her move will partly determine the moves that are subsequently open (or not) to her opponent might be precisely why she chooses the move she does.

the rules governing them both, and so its authority comes not from the one but from the rules.

In chess, then, players can have (and exercise) both rights of self-governance, and yet when they contemplate their next move or evaluate the legitimacy of their opponent's move, they are able to do so from a view of what the rules say the state of play is that they both are accountable for identifying as correct and so as the state of play in their game. That chess has its players in a state of nature about the conduct rules—their individual game play entirely unregulated by any external authority (such as a referee or an umpire)—does not undermine the maintenance throughout the game of an accessible state of play. The result is that chess players can be entirely self-governing and yet stand in a relationship of accountability with one another.

2.3. The Case of Baseball, and Enoch's Objection

But some such games are impossible to play when players have (and exercise) both rights of self-governance. Baseball is one such game. Why might this be? Recall that a game structured so that players have both the RSG1 and the RSG2 is one that has the state of play at any *t* as simply what the conduct rules say it is at that *t*. For such a game to be playable, then, it must be that what those rules say the state of play is at any *t* is accessible to players. And this is not the case in baseball, for, unlike in chess, two players each deliberating entirely responsibly in normal circumstances can disagree about how the conduct rules apply to some move played and so about what those rules say the resulting state of play is. Because what the conduct rules say the state of play is at that *t* is not accessible to the players, there is not one view of the state of play that they all are accountable for recognizing as the state of play in their game. We can thus understand the problem a state of nature poses here for a game of baseball as one of *indeterminacy*: when disagreement of this sort is possible, what the conduct rules say the state of play is cannot be the state of play in their game—it is unable to fulfill that role—and, as a result, the state of play is indeterminate. (This will be the case even if what the conduct rules say is not itself indeterminate, for what they say can be determinate and yet not accessible to players.)

Alternatively, we can understand the problem the state of nature poses as one of *unilateralism*: possible disagreement of this sort makes it the case that when one player plays a move that she responsibly judges the conduct rules say is legitimate, she thereby implicitly claims that her judgment of what those rules say about the state of play resolves the indeterminacy and, in that way, *decides* the state of play in their game. She thus claims for herself and her judgment authority over the game and so over her fellow players, authority she necessarily denies

to them and their conflicting responsible judgments of what those rules say. As Enoch's Objection is directed at the problem of unilateralism, we will consider it in detail.

Consider a pitch in a state-of-nature baseball game in which Astrid and Bashir are opponents. Exercising judgment responsibly in normal circumstances, Astrid judges the pitch to be a ball while Bashir judges it a strike, and Bashir proceeds to act on his judgment. (Suppose that as a strike, but not as a ball, the pitch ends the inning.) A Kantian might say that in acting here, Bashir is claiming unilateral authority over the status of the pitch in the state of play and so over Astrid as a fellow player. And Enoch might respond that Bashir is not claiming unilateral authority over the pitch's status, for he is acting on his judgment that the pitch is a strike not because it is his but because it is, as he thinks, correct.

Implicit in Bashir's action is not merely "The conduct rules say that the pitch is a strike." Bashir is making a claim *within* the game *of* a fellow player—we can imagine Bashir saying to Astrid, "It's a strike!"—and so it is an implicit demand he addresses to Astrid as someone accountable to her fellow players for acting from the state of play. Now, Enoch might argue that it is no problem for his view that Bashir, in acting on his judgment in their game, thereby claims authority for the demand of Astrid implicit in his action, for it has this authority within the game and so over Astrid, on Bashir's view, not because it is his but because it is the correct application of the conduct rules to the pitch. Bashir can thus regard himself as making this demand on behalf of the conduct rules, and so the authority he is claiming for it, on his view, will come not from him but from the rules themselves. As we do not yet have a claim of unilateral authority by Bashir over Astrid, Enoch's Objection seems to stand.

Bashir's view that he is making this demand on behalf of the conduct rules, with its authority over Astrid thus coming not from him but from the rules, presupposes not merely that the rules do, in fact, count the pitch as a strike but also that Astrid is accountable for recognizing that fact and so for identifying "The pitch is a strike!" as part of the state of play. But, as we have established, Astrid is not accountable for identifying this as part of the state of play unless it is accessible to her; and it may not be accessible to her, for it may be that from where she stands in the game, entirely responsible deliberation about what the conduct rules say about the pitch cannot but lead her to the judgment that the pitch is a ball, not a strike. When this is the case and so responsible disagreement about the pitch is possible, the presupposition is false—Astrid is not accountable for recognizing "The pitch is a strike!" as part of the state of play. Thus, Bashir's view that he is making this demand on behalf of the conduct rules, with its authority

over Astrid thus coming from them, is undermined. (And, again, this will be the case even if the conduct rules do in fact say that the pitch is a strike.)

Now, as this dispute about the pitch's status is one within their game, Astrid and Bashir cannot simply agree to disagree about its status while continuing their game, for the game requires that the pitch have one status in the state of play—*either* a strike *or* a ball—that is authoritative for them both. This is why any act of playing the game contains an implicit demand addressed to other players: it makes a claim about the state of play governing them all. By playing their game according to his own judgment that the conduct rules declare the pitch a strike—and thus implicitly claiming that it is a strike in the state of play—Bashir cannot help but address such a demand to Astrid, and so he addresses it whether or not this presupposition obtains (that is, whether or not Astrid is accountable for recognizing "The pitch is a strike!" as what the conduct rules say the state of play is). In acting here, then, Bashir commits himself to the view that there is an available basis for his demand's authority over Astrid that is not the authority of those rules themselves.

What might this basis be? Notice that because this presupposition is false, Bashir cannot say that Astrid would be accountable for recognizing, as part of the state of play, that the pitch is a strike even if he had not yet acted on the relevant judgment and, in so doing, addressed it to her as a demand. But in acting on that judgment and so addressing it to her as a demand, Bashir is thereby claiming authority for it. It seems, then, that Bashir commits himself to the view that it was by his acting on the judgment—and, in doing so, addressing it to her as a demand—that Astrid came to be accountable for abiding by it. Thus, by playing their game according to his own responsible judgments of the conduct rules, even when those judgments admit of responsible disagreement, Bashir commits himself to the view that his judgments decide the state of play—and so that they have authority over Astrid—not because these judgments are correct (although they might be) but because he has acted on them in their game. Bashir thus cannot help but claim unilateral authority over those parts of the state of play at issue.

Where Enoch's Objection goes awry is that Bashir's demands cannot have the authority he claims for them simply because the judgments are, as he thinks, correct applications of the conduct rules, and this is because in baseball (unlike in chess) it may not be that other players are accountable for recognizing the judgments as correct. In playing the game according to his own responsible judgments, what Bashir cannot help but claim is that the demands he implicitly addresses to other players have authority over them because *he thinks* the judgments his demands contain about how the conduct rules apply are correct. Thus,

what Bashir demands of Astrid is that she regard his judgment as deciding for their game the status of the pitch in the state of play simply because it is his. And were Astrid to act on her different judgment that the pitch is a ball, she would thereby be making a similar demand of Bashir.

2.4. Indeterminacy and Unilateralism in the State of Nature

The problem of unilateralism reveals that there is something incoherent about the very idea of a state-of-nature baseball game. By exercising his right to act on his own responsible judgments of how the conduct rules apply (his RSG2), even when those judgments admit of responsible disagreement, Bashir thereby denies that other players like Astrid have a similar RSG2, for he cannot help but claim unilateral authority within the game for those judgments. In this way, Bashir cannot play in a state-of-nature baseball game while understanding it as a state-of-nature game. And he cannot because, unlike a state-of-nature game of chess (or, simply, a game of chess), a state-of-nature game of baseball is unplayable. The problem of indeterminacy explains why. In normal circumstances, two players can responsibly disagree about what the conduct rules say about a pitch—is it a strike or a ball?—and, as a result, they will be unable to play from a view of what those rules say that they all are accountable for identifying as correct and so as part of the state of play. Consequently, the status of the pitch in their game will be indeterminate, and because the pitch's status is indeterminate and the pitch ends the inning only if it is a strike, whether the inning has ended will also be indeterminate. Unless the game is restructured so that the pitch is given one accessible status in the state of play, thereby resolving the indeterminacy, they will be unable to continue the game.

What Astrid and Bashir's baseball game requires, then, is that it replace a situation where the players all have this RSG2 with some mechanism for the authoritative resolution of disputes arising from players' disagreement about what the conduct rules say. Astrid and Bashir need not think that, as a judgment of what those rules say, this mechanism's resolution of their dispute about the pitch is correct; they need only recognize it as authoritatively settling what the pitch *counts as* in the resulting state of play. What their game requires, then, is a mechanism that they are to recognize as authoritative for their game even as their disagreements about the correct application of the conduct rules in their game might remain. This mechanism may be formal (e.g., an umpire or one player possessing unilateral authority) or informal (e.g., case-by-case negotiations, as in a casual game). But there must be some such mechanism if they are to be able to play a game of baseball together.

In a baseball game with an umpire, anyone—a spectator, a coach, a player—

might judge that a pitch satisfies the criteria for a strike, but only the umpire can give that pitch the status of a strike within the resulting state of play. Thus, if a player declares "It's a strike!" she is merely doing the former, judging the pitch to be what we might call a strike$_1$, while if the umpire declares "Strike!" he is doing the latter, making the pitch a strike$_2$. In making the pitch a strike$_2$, the umpire's judgment settles the state of play for the players such that the pitch counts as a strike in their game. The umpire's judgment is authoritative—it governs their subsequent game play—while a player's judgment that the pitch is a ball$_1$ is practically inert. (This is the case even if, as an application of the conduct rules to the pitch, the player's judgment is correct.) Of course, these two—strike$_1$ and strike$_2$—are not unrelated: when a spectator declares "It's a strike," she is saying not only that the pitch is a strike$_1$ but also that it ought to be a strike$_2$; and when the umpire calls "Strike!" he is not merely making the pitch a strike$_2$ but also doing so, it is implied, because he has judged it a strike$_1$. But they are nevertheless importantly distinct, and what matters for playing the game is not, it turns out, what the conduct rules themselves might actually say (i.e., whether the pitch is a strike$_1$) but what the umpire says that they say (i.e., whether it is a strike$_2$), for the latter is what settles the state of play and makes possible relationships of accountability between the players.

Suppose that our baseball game has an umpire and that he has called the pitch a strike, a call Bashir acts on. The claim implicit in Bashir's action is still addressed to Astrid as a demand, but now we see that it is "The pitch is a strike$_2$!" While the authority Bashir is claiming for this demand does not, on his view, come from him, it also does not come from the rules themselves; it comes from the umpire and, in particular, from the fact that the umpire has issued the judgment making the pitch a strike$_2$. What Bashir's demand of Astrid presupposes, then, is not that she is accountable for recognizing "The pitch is a strike$_2$!" as what the conduct rules themselves say but that she is accountable for recognizing it as what *the umpire says* that those rules say. Of course, what the umpire says must be accessible to her, but under normal conditions this requirement is easily met. Provided that the umpire has indeed called the pitch a strike, thereby making it a strike$_2$ in the state of play, Bashir's demand of Astrid is vindicated as an authoritative demand within the game, one from within the relationship of accountability they stand in as fellow players. In this way, the mechanism of the umpire makes it possible for players in a baseball game to hold one another accountable for choosing only among legitimate moves in their game play. And, unlike other possible mechanisms, it makes it possible for them to do this *as equals*: because the umpire is not a player, no player is subject to the authority of any other player.

We will return now to our state-of-nature game, for we can use this distinction between what the conduct rules say the pitch is (a strike$_1$) and what its status is in the state of play (a strike$_2$) to restate the problem of unilateralism. In acting within the game, those judgments a player acts on are judgments of the state of play. In the case of Bashir and Astrid, then, they are judgments that the pitch is a strike$_2$ or a ball$_2$. But because responsible disagreement about what the conduct rules say about the pitch is possible, it cannot be what the conduct rules say about the pitch—that it is a strike$_1$—that makes the pitch a strike$_2$, for then this status will not be accessible to players. And so when Bashir plays their game according to his judgment that the conduct rules say the pitch is a strike, a judgment others can responsibly disagree with, he cannot help but claim that *he* is what gives the pitch the accessible status their game requires—that is, that what he says the conduct rules say about the pitch *makes* the pitch a strike$_2$ in the state of play while what Astrid says they say is merely the judgment that the pitch is a ball$_1$. In this way, in acting within the game, the judgment Bashir acts on is not the (correct) judgment that the pitch is a strike$_1$ but rather the judgment that it has the status of a strike$_2$. And so while Enoch's Objection is right that Bashir acts on his judgment here because it is, as he thinks, correct and not because it is his, it is precisely in thinking that this judgment—that the pitch is a strike$_2$—is correct that he claims unilateral authority over the status of the pitch in their game and so over Astrid as a fellow player.

3. DISAGREEMENT ABOUT RIGHT

The task now is to argue that this defense of the problem of unilateralism against Enoch's Objection is available not just for state-of-nature versions of certain multiplayer games but also for a state of nature about right. We must show two things. First, we must show that right puts persons in a situation relevantly analogous to that in which players are put by these games. We must show, in other words, that it is as if right places persons in a "game" of equal freedom with those with whom they "cannot avoid interacting," a game whose rules—the rules of right—give them the first right of self-governance (RSG1). By showing this, we will establish that as a game of equal freedom, right must be structured so that persons stand in relationships of accountability with one another about the conduct rules of right. Second, we must show that when it comes to applying these rules of right in particular situations, responsible disagreement between "players" is possible, for this will mean that they will not stand in relationships of accountability in a state-of-nature game of equal freedom in which they also have the second right to self-governance (RSG2). By showing this, we will establish

that what is required for persons to make valid claims of right of one another is some authoritative mechanism for resolving disputes arising from responsible disagreement such that the various "moves" have one status (i.e., they count as right or not) that those persons are accountable, absent a valid excuse, for recognizing as governing them all.

The result is that a state of nature about right will face the problem of unilateralism defended in the previous section. When a person in a state of nature acts on her own responsible judgment of right, one that admits of responsible disagreement, she cannot help but claim that *she* is the mechanism for the authoritative resolution of any disputes arising from such disagreement and, consequently, that her judgment of what the conduct rules of right say decides what counts as right. This assertion of unilateral authority means that she cannot help but deny the equal freedom of others, for she is claiming for herself the RSG2—the right to act within the "game" according to her own judgments (rather than deferring to some other's judgments) of which "moves," according to the conduct rules of right, are legitimate in the first place—while denying this same RSG2 to them. What right requires is therefore not just any authoritative mechanism for resolving disputes arising from responsible disagreement about right but rather one that is itself compatible with the equal freedom of persons—it must be that no "player" is subject to the authority of another. This mechanism, on Kant's view, will be law.

3.1. Right as a "Game" of Equal Freedom

Kant's account of right begins, plausibly enough, with the claim that persons have an innate right to freedom: "*Freedom* (independence from being constrained by another's choice), insofar as it can coexist with the freedom of every other in accordance with a universal law, is the only original right belonging to every man by virtue of his humanity."[19] As Ripstein explains it, "you are independent if you are the one who decides what ends you will use your means to pursue, as opposed to having someone else decide for you."[20] The freedom at issue here, then, is that of being one's own master, and so whatever else it may require, it at least requires that you have the RSG1: the right to choose for yourself (rather than deferring to some other's judgments), within the bounds set by the rules of right, what ends to pursue and how to pursue them. In this way, we can understand the rules of right as, in part, giving each of us an equal such RSG1, and so as securing for each of us, as it were, equal space in which we may pursue our own ends in our own way.

19 Kant, *The Metaphysics of Morals*, 6:237.

20 Ripstein, *Force and Freedom*, 33.

My exercise of my RSG1 will change the circumstances on the ground, and in doing so, it may shape the bounds the rules of right give to your exercise of your RSG1. If I acquire some piece of unowned land, you now can do so only by acquiring it from me; if I sell something to another, whether you might acquire it is determined now not by my price but by that other's; if I instead destroy that thing, you simply cannot acquire it. There is thus a sort of interdependence here—my exercise of freedom in our shared circumstances can expand or contract the set of possibilities the rules of right give you to exercise yours—and so, as in a game, we are governed *together* by those rules of right. It is as if we are players together in a multiplayer "game" of equal freedom, one structured by the rules of right and played on a field of play we share.[21] We can thus understand the bounds that right gives to the space we each have to pursue our ends as *the state of play* within our game of equal freedom: these bounds define the choices (or "moves") that in the circumstances are or are not legitimate for you and for me, and these definitions are the product not only of the rules of right themselves but also of what has happened in the game up to this point, the legitimate moves that have already been played and whose results now shape the field of play for us.

Because persons are governed together by the rules of right as if in a game of equal freedom, the rules not only make persons responsible for choosing only legitimate moves but also make it such that they are accountable to their fellows for doing so, for the moves they choose will change the field of play and their fellows are entitled to those changes being legitimate ones. If what I destroyed was not in fact my property but rather yours, I have violated your right and so am accountable to you for that violation—you have standing to demand to be made whole—for you are entitled to have available to you, when you pursue your ends, the moves that access to this property would have made possible. The claim here, then, is that on a Kantian account, what holds for the games considered above holds also for right: the rules of right constitute a system in which persons each have the RSG1. As a result, this system of right—or this game of equal freedom—must be structured such that these persons are accountable to one another for staying within the bounds the rules give that right. And, as we saw with those games, persons will be in these relationships of accountability with one another about the rules of right only if the state of play at any *t* in their game of equal freedom is accessible to them. If I am to be accountable for destroying your property, it must be the case that the fact that it was yours—and so that destroying it required your consent—was accessible not only to you but also to me.

21 Thomas Pogge notes that the system of right can be conceived of as a game ("Is Kant's *Rechtslehre* Comprehensive?" 170–71). He calls it "Kant's *Rechtslehre* game."

There is, however, one important disanalogy. Unlike with these other games, playing this game of equal freedom with those with whom we "cannot avoid interacting" is not optional. Whenever we act in a world we share with others, we thereby act on a judgment of right: at a minimum, the judgment that our action is permitted by right (or, in other words, that it is a legitimate "move" for us). In acting in such a world, then, we implicitly address a claim of right as a demand to those others, a demand that presupposes that they are accountable for recognizing the claim as authoritative: at a minimum, the claim that they not interfere with our action (or, in other words, that they recognize our "move" as legitimate for us). Whenever we act in such a world, it is as if we simply find ourselves already as players in a game of equal freedom with those others, for the claims of right that we, in acting, cannot help but make of others are demands one addresses *within* such a game *to* fellow players *about* the state of play.

3.2. *Responsible Disagreement about Right*

If we cannot help but play in a game of equal freedom with those with whom we cannot avoid interacting, then we must accept whatever is necessary to make such a game playable, for otherwise we will be unable to make valid claims of right of those others. What Kant's state-of-nature argument reveals is that this game of equal freedom, unlike chess, is not playable if the rules of right give players not only the first right of self-governance (RSG1) but also the second (RSG2). We might attempt a close analogy here with our state-of-nature baseball game. When it comes to many questions of right, two people can each find that, even in normal circumstances, from where they stand, responsible deliberation about what the conduct rules of right say about some situation cannot but lead them to different judgments. Persons inevitably face what John Rawls has called "the burdens of judgment," and so such disagreement about many questions of right is inevitable.[22] And, just as in baseball, the possibility of this sort of disagreement will undermine persons' ability to make claims of right of each other in a state of nature, for they will not be accountable for recognizing the demands others thereby address to them not merely as what those others demand but as what the rules of right themselves demand of them.

But things are not quite so straightforward, as Kant idealizes those in the state of nature about right: they are "however good and right-loving human beings might be." And, as he explains, he does so in order to show that "it is not experience from which we learn of the maxim of violence in human beings and of their malevolent tendency to attack one another before external legislation endowed with power appears, thus it is not some deed that makes coercion through public

22 For discussion of the burdens of judgment, see Rawls, *Political Liberalism*, 56–57.

law necessary."[23] On Ripstein's reading, this idealization signals that "Kant does not follow Hobbes and Locke in focusing on the empirical defects of the state of nature, such as self-preference and limited knowledge." The defects in the state of nature, he argues, "do not reflect human limitations."[24] It is not clear, however, that we need read this idealization quite so strongly, as removing any and all human limitations. Being a maximally good and right-loving person does not imply, for instance, that one is also maximally informed or perfectly situated, or that one reasons perfectly. Some human limitations would seem to survive this idealization.

I suggest instead that we should imagine persons who are *fully* committed to pursuing their ends only in ways permitted by right. Such a thoroughgoing commitment will manifest itself not just in the maxims a person acts on in their interactions with others. For instance, a person thus committed will cultivate within themselves all those "rules of moral salience" that, as Barbara Herman has emphasized, are required for one to deliberate and judge well (here, about questions of right).[25] In this way, it is not simply that these idealized persons mean well—all and only good intentions are not sufficient for such a thoroughgoing commitment to acting rightly—but rather that they have done *everything* a person might do to see to it that in their interactions with others, they actually succeed in acting rightly. The point is thus that even these idealized persons in normal, or even reasonably favorable, circumstances will disagree with one another at times about how the rules of right apply in their interactions. Or, to put it another way, the point is that there are hard cases of right—cases for which, if there are determinate answers, those answers are not accessible even to these idealized persons. In these cases, a person may entirely succeed in holding herself responsible for judging correctly—she may do everything that could possibly be asked of her as a human agent exercising judgment—and yet, even if she is suitably placed to judge, fail to judge correctly. On this reading, then, the problems of a state of nature about right (indeterminacy and unilateralism) do arise from certain human limitations—namely, the limitations of our faculty of judgment. That we are subject to these limitations means that for many questions of right in a game of equal freedom—the hard cases—responsible disagreement is possible even between our idealized players.

23 Kant, *The Metaphysics of Morals*, 6:312.

24 Ripstein, *Force and Freedom*, 146.

25 As Herman explains, rules of moral salience "structure an agent's perception of his situation so that what he perceives is a world with moral features. They enable him to pick out those elements of his circumstances or of his proposed actions that require moral attention" ("The Practice of Moral Judgment," 77).

Of course, hard cases in this sense might simply be cases where, as Ripstein and Sinclair argue, right itself is indeterminate. On this view, the correct answers are indeed not accessible even to these idealized persons, but that is simply because, in these cases, there are not any such answers in the first place. Though I incline toward understanding many of these hard cases as having determinate answers (because if they do, it will not be a mistake to aim at the correct answer when exercising judgment about them), the core claim of the larger argument remains even if we accept this alternate view, for what is responsible for the Kantian problems of indeterminacy and unilateralism in a state of nature about right will still be that for many questions of right, there is not one accessible correct answer. The correct answer may not be accessible because, as this alternate view has it, there is not one in the first place, or there may be one, but because of the limitations of human judgment, it is not accessible to us. Either way, what playing a game of equal freedom together requires is that there be one accessible answer, and so the indeterminacy at issue in the state of nature—the one that gives rise to the problem of unilateralism—is the lack of this one accessible answer to these questions of right. (That right itself is indeterminate would thus be sufficient to generate this indeterminacy, but it is not necessary.)

3.3. Indeterminacy about Right and Unilateralism

Consider now a responsible disagreement in this idealized state of nature about right. Carlos claims ownership over some plot of land, while Dana claims ownership over a neighboring one. Unfortunately, their claims overlap: there is one field over which they both claim exclusive ownership. While they agree that Carlos's attempted acquisition of that previously unowned field came first, they disagree about whether it was legitimate (and so successful). Carlos thinks the answer is yes (he laid down clear boundary markers and did some preliminary preparation of the field for planting), while Dana thinks the answer is no (while Carlos worked other portions of his plot intensively, there were no clear indications that this field had been worked on). Their disagreement thus concerns the status of the field in the state of play: Carlos judges that he owns it and so both that Dana has an obligation not to trespass on it and that he has the right to use coercion to keep her off of it; Dana makes the opposite judgment. Suppose that the principles of acquisition declare Carlos's acquisition legitimate and so a success. But also suppose that this question—"According to the principles of acquisition, was it legitimate?"—is a hard case and that both Carlos and Dana have deliberated responsibly. What we have, then, is a case of responsible disagreement about a question of right.

Carlos and Dana cannot simply agree to disagree here about who owns this

field, at least not if their claims of ownership are to be valid claims of right marking the boundary between their respective spheres of freedom. The problem, however, is that they are unable to act from a view of what the rules of right say about this field's ownership that they are both accountable for recognizing as the status of the field in the state of play. In the state-of-nature game in which they find themselves, then, the status of the field—who owns it—is indeterminate. So, what this game requires is some mechanism for resolving their dispute about the status of the field, about who *counts as* the owner in the state of play, one that will be authoritative for them even if they continue to disagree about what the principles of acquisition say. There is thus a distinction in this case between ownership_1 and ownership_2, and it is ownership_2 of the field—who counts as the owner—that governs their interactions. More broadly, it is what counts as right (their rights_2 and obligations_2) that governs them, while what according to the conduct rules is right (their rights_1 and obligations_1) is practically inert. Within a game of equal freedom, then, the claims of right that persons address to one another as demands are claims of rights_2 and obligations_2, not of rights_1 and obligations_1, for they are claims about what the other is accountable for recognizing as authoritative limits on their pursuit of their ends. In this way, the problem for a state of nature about right, even our idealized one, is that it may be indeterminate what these rights_2 and obligations_2 are, even if it is not indeterminate what the rights_1 and obligations_1 are.

Were Carlos to act on his judgment here, moving to force Dana off of the field, the judgment at issue would not be the (correct) judgment that he owns_1 the field but rather the judgment that he owns_2 it. And so, it is true that Carlos's reason for action here would not be "that I believe that Dana is obligated_2 not to trespass on my land" but rather simply "that Dana is obligated_2 not to trespass on my land." Enoch's Objection is thus correct that on Carlos's view, what justifies his acting on this judgment is that the judgment is, as he believes, correct and not that it is his. The problem cannot be that Carlos treats as special the fact that *he* judges that Dana has this obligation_2. The problem lies instead within Carlos's judgment itself. For *this* judgment to be, as he believes, correct—for the state of play to have it that he owns_2 the field and so that Dana has this obligation_2 not to trespass—it will need to be the case that his responsible judgment of what the principles of acquisition say here *makes* it the case that he owns_2 the field while Dana's responsible judgment of what they say here is merely that she owns_1 the field. It is precisely in taking his judgment that he owns_2 the field to be correct, and so as what governs their interactions concerning this field, that Carlos implicitly claims unilateral authority over the state of play in this part of their

game of equal freedom and, in that way, over Dana as a fellow player. Enoch's Objection thus fails.

4. CONCLUSION

If our idealized Carlos and Dana are to be able to address valid claims of right to one another, their "game" of equal freedom must have some authoritative mechanism that makes determinate in a way accessible to them what counts as right in it. And since innate equality requires that no particular player's judgment of right decides what counts as right in their game—otherwise we run into the problem of unilateralism—what is required is some other agent whose judgment decides what counts as right for them and who is not a player in the game. This agent is what Kant calls "the omnilateral will": it is empowered to make determinate in a way accessible to them all everyone's rights and obligations within the game—their rights_2 and obligations_2—such that players can hold one another accountable for recognizing them as authoritative over their exercise of their RSG1 in their game play; and, crucially, it is not empowered to do anything else.[26] The basis of the law's claim of authority over citizens is that it is this omnilateral will—a will that by giving laws to everyone solves the problems of indeterminacy and unilateralism that citizens would face were they entirely self-governing, and, in so doing, makes it possible for citizens to stand as equals in relationships of accountability with one another about right.[27]

University of Delaware
dkoltons@udel.edu

REFERENCES

Darwall, Stephen. "Law and the Second-Person Standpoint." In *Morality, Authority, and Law: Essays in Second-Personal Ethics I*, 168–78. Oxford: Oxford University Press, 2013.

Enoch, David. "Against Public Reason." In *Oxford Studies in Political Philosophy*,

26 For a more complete account, see Ripstein, *Force and Freedom*, 190–98.

27 For very helpful comments, objections, and discussion, I am grateful to Seth Shabo, Kaila Draper, Rochelle Duford, and two anonymous reviewers. I also thank audiences at Amherst College, the Eighth Multilateral Kant Colloquium (Catania, Italy), and the University of Delaware to whom I presented earlier versions of some of the ideas contained in this paper.

vol. 1, edited by David Sobel, Peter Vallentyne, and Steven Wall, 112–42. Oxford: Oxford University Press, 2015.

———. "The Disorder of Public Reason." *Ethics* 124, no. 1 (October 2013): 141–76.

———. "Not Just a Truthometer: Taking Oneself Seriously (but Not Too Seriously) in Cases of Peer Disagreement." *Mind* 119, no. 476 (October 2010): 953–97.

———. *Taking Morality Seriously: A Defense of Robust Realism*. Oxford: Oxford University Press, 2011.

Gaus, Gerald. *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*. Cambridge: Cambridge University Press, 2011.

Herman, Barbara. "The Practice of Moral Judgment." In *The Practice of Moral Judgment*, 73–93. Cambridge, MA: Harvard University Press, 1993.

Kant, Immanuel. *The Metaphysics of Morals*. In *Practical Philosophy*, translated and edited by Mary J. Gregor, 363–603. Cambridge: Cambridge University Press, 1996.

Pogge, Thomas W. "Is Kant's *Rechtslehre* Comprehensive?" *Southern Journal of Philosophy* 36, Spindel Supplement (Spring 1998): 161–87.

Rawls, John. *Political Liberalism*. New York: Columbia University Press, 1993.

Raz, Joseph. "Disagreement in Politics." *American Journal of Jurisprudence* 43, no. 1 (1998): 25–52.

Ripstein, Arthur. *Force and Freedom: Kant's Legal and Political Philosophy*. Cambridge, MA: Harvard University Press, 2009.

Sinclair, Thomas. "The Power of Public Positions: Official Roles in Kantian Legitimacy." In *Oxford Studies in Political Philosophy*, vol. 4, edited by David Sobel, Peter Vallentyne, and Steven Wall, 28–52. Oxford: Oxford University Press, 2018.

Journal of Ethics and Social Philosophy
VOL. 20, NO. 3 · NOVEMBER 2021
https://doi.org/10.26556/jesp.v20i3.1630

NONIDEAL JUSTICE, FAIRNESS, AND AFFIRMATIVE ACTION

Matthew Adams

Affirmative-action policies aim to increase the representation of a target group. If such policies try to realize this aim by giving preference to members of a target group, then public controversy is often aroused on the basis of a perceived unfairness.

This controversy is current again. In 2014, Students for Fair Admissions filed a lawsuit against Harvard University, alleging that the admissions preference given to African and Latinx Americans results in unfair discrimination against Asian Americans.[1] A US District Court ruled, in 2019, that Harvard does not discriminate in this way. But Students for Fair Admissions plan to appeal and it is expected that the case will eventually be heard by the Supreme Court.[2] Significantly, the changing ideological profile of the Supreme Court has led to speculation that "affirmative action could be dead not only at public schools but also at private ones whose practices have largely escaped legal scrutiny until now."[3]

In addition to being politically pressing, affirmative action raises a parallel set of theoretical issues. The appeal of affirmative-action policies is that they can be an effective means of, at least partially, overcoming legacies of injustice. But they can also be contested because they require *prima facie* unfair treatment, at least in some instances. Affirmative action is, therefore, a good test case for the adequacy of a theoretical conception of justice. An adequate conception must specify the conditions (if any) under which affirmative action is just; a successful philosophical defense must explain how the unfairness objection can be overcome. The topic of affirmative action thus invites careful reflection on the nature of justice in unjust conditions. In particular, at least given the core commitments of nonconsequentialist liberalism that I presuppose, a compelling explanation as to why it is permissible for affirmative-action policies to treat

1 See Moses, "After *Fisher*."

2 Following Anderson, "Federal Judge Rules Harvard Does Not Discriminate against Asian Americans in Admission."

3 Gerstein and Haberkorn, "It's Not Just Abortion."

certain individuals in a *prima facie* unfair way is required. The mere fact that such treatment would be an effective means of realizing a more just society in the future is not a sufficient explanation.[4]

In this paper I argue for two related claims, one substantive and the other methodological. First, I forge a new justification for affirmative action, via a basic-liberties argument. And second, in doing so, I illustrate the value of a new conceptual innovation that I term "nonideal principles of justice." More precisely, I defend affirmative action on the ground that it increases certain comparatively disadvantaged people's ability to exercise their basic liberties. I argue that, given the particular empirical conditions that obtain in the contemporary US, this basic-liberties justification supports attaching special weight to being African or Latinx American in admissions procedures. Furthermore, it can provide a compelling response to the unfairness objection.

My approach has a certain affinity with so-called integrationist justifications, insofar as it construes the appropriate function of affirmative action as overcoming legacies of injustice rather than promoting diversity.[5] But my argument does not presuppose that achieving racially integrated educational institutions is necessary for democratic legitimacy, or that racial integration is an imperative of social justice.[6] Racial integration plays a function in my argument to the contingent extent that it is an effective and fair means of promoting certain people's ability to exercise their basic liberties.

As I noted above, my argument has a second, methodological upshot. In order to tackle racial injustice in particular, a number of philosophers argue that the Rawlsian paradigm of justice should be abandoned. Elizabeth Anderson claims that Rawls's theory of justice is inadequate because it focuses on ideal principles of justice, which specify how a perfectly just society should be arranged. She argues instead that philosophers should take an empirically informed "bottom-up"

4 In contrast, consequentialists might acknowledge that affirmative-action policies are unfair to certain individuals while arguing that, within certain empirical parameters, the benefits of affirmative-action policies outweigh this cost of unfairness. See Beauchamp, "In Defense of Affirmative Action."

5 The Supreme Court has ruled that colleges can consider race to enhance student diversity. Regents of the Univ. of Cal. v. Bakke, 338 US 265 (1978) at 300. However, the diversity justification does not support anything like the scope of actual affirmative-action policies. Such policies standardly focus on promoting a racially diverse student body. But if the justification for affirmative-action policies is that they promote a diverse student body, then there is no reason why racial diversity should be the only type of diversity that is given significant weight. A commitment to diversity also supports attaching significant weight to other sources of diversity within the student body, such as creationism, Scientology, and climate-change denial. Following Anderson, *The Imperative of Integration*, 142.

6 See Grutter v. Bollinger, 539 US 306 (2003) at 336; and Anderson, *The Imperative of Integration*.

approach that addresses the pressing problems that they actually face, such as racial injustice in the US.[7]

I agree with Anderson that Rawlsian ideal principles of justice are not (at least in and of themselves) the best way of tackling pressing problems, such as whether affirmative action should be used to ameliorate racial injustice. But I argue that her bottom-up approach is also inadequate: it does not provide a sufficiently determinate conception of justice to overcome the unfairness objection. I show how a sufficiently determinate conception can be developed by using a Rawlsian contractualist framework to forge what I term a "nonideal principle of justice."

My paper has the following structure. I begin by presenting the unfairness objection and clarifying its scope (section 1). I then examine an unsatisfactory response to the objection (section 2), and I survey the fertile—but limited—implications of Rawls's theory of justice for affirmative action (section 3). After that, I forge a nonideal principle of justice (section 4) that supports affirmative-action policies like those in the contemporary US (section 5) and blocks the unfairness objection (section 6). I close by showing how my account can be used to refine some features of contemporary affirmative-action policies, and I reflect more generally on the value of nonideal principles of justice for tackling exigent topics (section 7).

A few preliminary clarifications are in order: I focus on affirmative action in a contemporary US educational context. In the US, the term "affirmative action" has been used to label a disparate set of policies. Such policies range from measures that simply outlaw group-based discrimination, to soft (i.e., non-explicit) quotas and hard (i.e., explicit) quotas for members of target groups.[8] I put affirmative-action policies that simply outlaw group-based discrimination to one side because they can be given a straightforward defense: they block group-based discrimination and safeguard equality.

I use the term "ideal theory/justice" to refer to a conception of how a perfectly just society should be structured and "nonideal theory/justice" to refer to a conception of what justice requires in conditions that fail to realize ideal justice.[9]

My argument is exclusively forward looking; I am neutral about whether a backward-looking justification of affirmative action can also be provided.[10]

7 Anderson, *The Imperative of Integration*, 3–7. Relatedly, Charles Mills argues that a nonideal contract to end racial domination should supplant Rawlsian ideal theory (see *The Racial Contract*).

8 Following Nagel, "Equal Treatment and Compensatory Discrimination," 349–51.

9 My explication of the distinction between ideal and nonideal justice follows Simmons, "Ideal and Nonideal Theory," 7.

10 See Thomson, "Preferential Hiring."

1. THE UNFAIRNESS OBJECTION

Many critics of affirmative action argue that such policies are *unfair*, at least in some instances of their application.[11] The following hypothetical example illustrates this charge of unfairness: a white man called Simon applies to Texas Law School and is rejected; Simon, however, would have been admitted but for a soft-quota affirmative-action policy that "added points" to favor the admission of a target African American group. A proponent of the unfairness objection can grant that African Americans are underrepresented in Texas Law School but urge that people like Simon are not directly responsible for perpetrating the historic conditions that led to such underrepresentation. It is not, therefore, fair for the admissions procedure to employ such a quota—even if this would be an effective means of realizing a more just society. This imposes the unfair burden of non-admission on Simon.

Ronald Dworkin argues that the unfairness objection presupposes a commitment to meritocracy: an affirmative-action policy is only unfair to Simon if he deserves to be admitted because he is an intellectually superior candidate. Such a presupposition is false, according to Dworkin, because no one deserves to be admitted to an academic institution because they possess some particular combination of talents.[12]

Note, however, that my intuitive presentation of the unfairness objection does not presuppose any particular independent standard of merit. Someone defending the unfairness objection can remain neutral about what standard or procedure (e.g., academic merit, or a type of lottery) should be used to determine admission. They are merely committed to the claim that it is wrong for something like membership in a particular race to have significant weight.

Simon's predicament can be generalized into the following formulation of the conditions under which affirmative action can *prima facie* plausibly be contested as unfair, in any particular admissions procedure:

> Membership in a target group—that is not directly relevant to an ability to complete/excel in the program of study—is given preferential weight. This results in the non-admission of a subset of people who are not members of the target group—who would have been admitted but for the affirmative-action policy.
>
> Caveat: The preferential weight does not merely block/partially block the

11 See Cohen, "Why Race Preference Is Wrong and Bad," 33–37; Lynch, *Invisible Victims*; Pojman, "The Case Against Affirmative Action," 98–105; and Scalia, "The Disease as Cure," 153–57.

12 See Dworkin, *A Matter of Principle*, 299–300.

discrimination that members of the target group standardly experience in the admissions procedure.

The caveat is necessary because people's biases against members of the target group might be so strong that the only way to cancel out (or reduce the effect of) such biases is to give preference to members of the target group. Essentially, an affirmative-action policy that cancels out, or reduces, an unfair advantage that Simon enjoys *qua* white male in the selection process is not unfair to Simon. (Moreover, *not* to cancel out that advantage is unfair to everybody who does not belong to his group.)[13]

Early social-scientific research focused on the burdens that affirmative action imposes on white men.[14] More recent empirical research concludes that—at least in the context of admission to elite educational institutions—the burdens fall heaviest on Asian Americans. Thomas Espenshade and Alexandria Radford calculate that, *ceteris paribus*, an Asian American needs an SAT score 140 points higher than a white American and 450 points higher than an African American to have the same chance of admission.[15]

Yet, this empirical evidence is contested. In response to the lawsuit filed by

13 Some argue that the caveat needs to be expanded as follows: the preferential weight does not merely block the advantage that members of the non-target group standardly benefit from because of discrimination against members of the target group in the past, rather than in the present admissions procedure. See Boxill, "The Morality of Preferential Hiring"; and Thomson, "Preferential Hiring." This extension of the caveat faces the non-identity problem. See Morris, "Existential Limits to the Rectification of Past Wrongs." Furthermore, Kasper Lippert-Rasmussen argues that it is hard to cash out the relevant counterfactuals for this extension of the caveat to provide additional support for race-based affirmative action in the contemporary US. He notes that if there had been no past racial injustice, the US would now be a much richer society. This is because, for instance, slavery resulted in a suboptimal use of the large pool of talents among African Americans. Accordingly, if there was no past racial injustice there would have been a greater number of university places because the US would have been richer. Consequently, there would have been more African Americans in universities but *also* more white Americans in universities. "Hence, if beneficiaries are those individuals who are better off given the relevant past injustice than without it … then there *might* be no beneficiaries of past injustice, even if some contemporary people have been harmed less than others" ("Affirmative Action, Historical Injustice, and the Concept of Beneficiaries," 82). I will show that the unfairness objection can be defeated, under the range of conditions that I specify, without expanding the caveat in this way.

14 See Lynch, *Invisible Victims*.

15 Espenshade and Radford, *No Longer Separate, Not Yet Equal*, 92. A simulation that determines the effect of race-based preferences at private institutions predicts that in 1997 Asian Americans would have comprised nearly 40 percent of all accepted students compared to less than 25 percent under current policies (Espenshade and Radford, *No Longer Separate, Not Yet Equal*, 344–46).

Students for Fair Admissions, Harvard University vigorously denied that their admissions policy made it harder for Asian Americans to be admitted. One way in which the evidence has been disputed is by arguing that, despite best efforts, empirical studies standardly fail to control sufficiently for variables such as legacy and athletic status.[16] Essentially, the primary reason that it is harder for Asian Americans to be admitted is not the preference given to African American and Latinx American students but the preference given to predominantly white legacy applicants and recruited athletes.[17] More generally, of course, it is difficult for even rigorous empirical studies to measure the variable of bias.

It is difficult to maintain, however, that no actual affirmative-action policies can *prima facie* plausibly be contested as unfair. After all, many admissions procedures are primarily dependent on standardized discrete data—the scrutiny of which leaves relatively little room for bias; for example, admission to law school is primarily dependent on an applicant's LSAT score and undergraduate GPA. It is also important to highlight that a number of affirmative-action policies take a particularly strong form. In the University of Texas Law School's affirmative-action policy under challenge in the case of *Hopwood v. State of Texas* the presumptive admit score for African Americans (a combination of undergraduate GPA and LSAT score) was lower than the presumptive deny score for white Americans.[18] It seems very unlikely that this preference given to African Americans is merely blocking the unfair discrimination that they standardly face in the University of Texas Law School's admissions procedure. Furthermore, perhaps some empirical studies of admissions procedures fail to control sufficiently for variables such as legacy status. Even still, legacy status has a comparatively minor impact in some educational admissions procedures, such as those in law schools.[19]

Accordingly, my argument presupposes the relatively uncontroversial claim that some actual affirmative-action policies can *prima facie* plausibly be contested as unfair under the conditions that I have specified.

16 See Espenshade, Chung, and Walling, "Admission Preferences for Minority Students, Athletes, and Legacies at Elite Universities." In this study the authors tried to control for variables such as legacy preference.

17 See the review of Harvard's admissions policy by the US Department of Education's Office for Civil Rights: United States Commission on Civil Rights, "Civil Rights Issues Facing Asian Americans in the 1990s," 104.

18 Hopwood v. Texas, 78 F.3d 932 (5th Cir. 1996).

19 See Schmidt, "A History of Legacy Preferences and Privilege," 57–59.

2. AN UNSATISFACTORY RESPONSE TO THE UNFAIRNESS OBJECTION

Anderson writes that the unfairness objection

> neglects the fact that as long as discrimination or its effects persist, there will be innocent victims suffering unjust burdens. The only question is whether these burdens should be borne exclusively by disadvantaged racial groups or more widely shared. There is no injustice in sharing the costs of widespread injustice.[20]

Anderson's analysis highlights that there is nothing problematic with the government legislating to share the burdensome effects of injustice in an appropriate way, and that just policies can impose certain costs on private individuals.

But these general claims about what is permissible are not sufficient to establish that the particular burdens imposed by affirmative-action policies are fair. Indeed, a proponent of the unfairness objection can grant these general claims and simultaneously argue that the particular burdens imposed by affirmative-action policies are not fair to private individuals such as Simon; essentially, by granting that something should be done to distribute the burdens of injustice more evenly while denying that affirmative action is a permissible means of achieving such an end.[21]

Kwame Anthony Appiah tries to block this move by arguing that the burdens that are in fact imposed by affirmative-action policies are analogous to the burdens imposed by other clearly permissible policies. He writes: "If justice requires restitution to Japanese Americans for the wrongs they suffered in internment in World War II, I cannot complain, when my taxes are raised to pay this restitution, that I did not do the interring."[22]

The problem with Appiah's argument is that the burdens imposed by affirmative-action policies do not seem analogous to such clearly permissible policies in the relevant sense. In order to see why, it is instructive to consider why critics argue that affirmative action is particularly objectionable: it imposes heavy burdens on a small subset of innocent individuals (such as Simon) in admissions procedures.[23] This is disanalogous to—and comparatively more controversial than—the US paying out reparations to the victims of state injustice, and the cost of these reparations being evenly distributed among all innocent taxpayers. Indeed, the case of affirmative action seems more analogous to the following

20 Anderson, *The Imperative of Integration*, 139–40.

21 See Pojman, "The Case against Affirmative Action," 108.

22 Appiah, "'Group Right' and Racial Affirmative Action," 273.

23 See Cohen, "Why Race Preference Is Wrong and Bad," 33–34.

modified version of Appiah's example: imagine that in order to compensate the interned Japanese Americans, a heavy tax was exclusively levied on a group of randomly selected non-Japanese people who comprised 5 percent of the population. This would distribute, at least in one sense, the costs of injustice more evenly. But it is an arrangement that the 5 percent group could plausibly contest as unfair—not because it imposes some costs on them, but because it imposes particularly heavy costs exclusively on them.[24]

Affirmative action and just taxation are disanalogous in a further sense. In a just scheme of progressive taxation relative privilege and relative burden are correlated, in the sense that the rich pay more and the poor pay less. But this correlation does not hold with respect to affirmative action, at least in a contemporary US context. Indeed, many argue that affirmative action is particularly unfair because this correlation is inversed: the least privileged members of the non-target group, such as poor white men from Appalachia, are more likely to lose out on admission than comparative privileged members of the non-target group.[25]

Anderson's and Appiah's combined attempt to overcome the unfairness objection, therefore, fails because it faces the problem of "under-theorization." It does not provide sufficient theoretical resources to determine whether affirmative action is a just policy that imposes a fair set of burdens: the claim that the demands of justice can impose certain costs on private individuals is not sufficient to establish that a set of actual costs is fair, and affirmative action is not relevantly analogous to other clearly permissible policies.

3. Rawlsian Justice and Affirmative Action

The problem of under-theorization can be overcome by developing Rawls's nonideal theory in a novel way. In this section, I pave the way for this endeavor by surveying the limitations of Rawls's theory of justice as it stands.

Rawls's large corpus of work contains little explicit discussion of group-based

24 Relatedly, James Sterba tries to diffuse the unfairness objection by arguing that, from the perspective of fairness, the preference given to legacy students is at least as bad as affirmative action ("Defending Affirmative Action, Defending Preference," 266–67). Sterba's argument can be used to present an *ad hominem* objection against some conservatives: it is inconsistent to object to race-based affirmative action but to approve of legacy preferences—for predominantly upper-middle-class, white Americans. But this is not sufficient to show that the unfairness objection to affirmative action does not have real moral force. Even if affirmative-action policies are no worse (or even better) than legacy preferences this does not establish that they are defensible. After all, legacy preferences can plausibly be contested as a deeply unfair feature of a society structured by economic class.

25 See Hurst, Fitz Gibbon, and Nurse, *Social Inequality*.

injustices, such as racism, that affirmative-action policies are designed to ameliorate. That said, Tommie Shelby has shown that Rawls's ideal theory of justice has substantive implications for such injustices. Rawls's liberty principle condemns race-based slavery and apartheid: under either institutional arrangement, *not* everyone would have access to a fully adequate scheme of basic liberties.[26] Similarly, the fair equality of opportunity (FEO) principle condemns any educational procedure that discriminates against a racial group.[27] Essentially, from the perspective of Rawlsian ideal theory, these group-based injustices are objectionable because they are deviations from ideal justice.

In order to determine whether affirmative-action policies are an acceptable way of ameliorating group-based injustice, we must turn to Rawls's nonideal theory of justice. Drawing on *The Law of Peoples*, A. John Simmons argues that Rawls's nonideal theory has the following content and structure:

> The specific "policies and courses of action" it mandates must be (i) "morally permissible," (ii) "politically possible," (iii) "likely to be effective" in moving society toward the ideal of perfect justice.[28]

Rawlsian nonideal theory is transitional. From the perspective of Rawlsian nonideal theory the goals of affirmative-action policies are good, at least insofar as they are an effective means of transitioning toward ideal justice. The crucial question is whether such policies satisfy the "moral permissibility" condition. This condition entails that not all paths that would be an effective means of transitioning toward ideal justice are necessarily permissible.

But the limitation of this permissibility condition is that neither Rawls nor Simmons offer any real guidance for determining which transitional paths are permissible. Perhaps they intend for permissibility to be judged intuitively. This clearly can be done in some cases: for instance, it seems obvious that despotic rule by a dictator should be judged impermissible, even if (surprisingly) this would be an effective means of bringing about ideal justice in the very long term.

It is not, however, always so easy to determine the permissibility of certain possible transitional paths. As I noted above, affirmative action is a difficult test case for nonideal theorists. It can *prima facie* plausibly be contested as unfair. But—in contrast to the example of dictator rule—it is not intuitively clear that

26 See Rawls, *Political Liberalism*, 292. Rawls's conception of the liberty principle evolved between *A Theory of Justice* and *Political Liberalism*. Most important, given my present purposes, "a fully adequate scheme" replaced "the most extensive total liberty."

27 Following Shelby, "Race and Ethnicity, Race and Social Justice. See also Rawls, *A Theory of Justice*, 266.

28 Simmons, "Ideal and Non-ideal Theory," 18. Following Rawls, *The Law of Peoples*, 89.

this unfairness objection is sufficient to rule out affirmative action. Consequently, in order to determine whether affirmative action is permissible a more normatively determinate conception of permissibility is required. On Simmons's reconstruction, at least, Rawlsian nonideal theory cannot determine whether affirmative action is in fact just, for (like Anderson's and Appiah's approach) it faces the problem of under-theorization.

4. A NONIDEAL PRINCIPLE OF JUSTICE

Given the limitations of Rawls's theory of justice, philosophers who wish to offer a broadly Rawlsian treatment of affirmative action need to be creative. My way of developing Rawls's theory has two stages. First, I use a contractualist framework to derive a nonideal principle of justice that applies in all empirical conditions. Second, in section 5, I argue that, given the particular empirical conditions that obtain in the contemporary US, this nonideal principle of justice supports affirmative action.[29]

Nonideal principles of justice are "idealized" in the sense that they abstract away from certain feasibility constraints and specify what justice *simpliciter* requires.[30] But they are "nonideal" in the sense that they specify how an unjust society should transition toward becoming a perfectly just society, rather than how a perfectly just society should, itself, be structured. The innovation of nonideal principles of justice is, I suggest, the key to giving substantive normative content to the under-theorized "moral permissibility" condition in Rawls's nonideal theory. I will not derive a complete set of nonideal principles; rather I will derive a single principle that justifies affirmative action under a broad range of conditions.

Although my approach is Rawlsian, it (arguably) abandons one Rawlsian orthodoxy. Rawls argues that justice exclusively regulates the basic structure: the main institutions of society. The precise scope of the basic structure is disputed. But universities are (standardly) not construed as part of it.[31] Consequently, some Rawlsians would argue that justice is silent about whether affirmative action should be used in universities.[32] In contrast, I assume—at least in nonideal

29 For different ways of developing a nonideal theory within a contractualist framework, see Arvan, "First Steps toward a Nonideal Theory of Justice"; and Mills, *The Racial Contract*.

30 I am neutral about whether justice *simpliciter* depends on some feasibility constraints. See Wiens, "Motivational Limitations on the Demands of Justice."

31 Following Hodgson, "Why the Basic Structure?" 303–32.

32 Even for such Rawlsians my argument has some value: it shows that if universities choose to implement affirmative-action policies that are supported by my nonideal principle of justice, then such policies cannot plausibly be contested as unfair.

conditions—that justice has direct implications for university policies, like affirmative action.[33] This assumption is motivated by the claim that if institutions—even those outside the basic structure—are capable of ameliorating injustice, then they ought to do so.[34]

4.1. The Basic Liberties and Two Distinctions

The requisite nonideal principle of justice hinges on the concept of "basic liberties" that are protected by Rawls's ideal liberty principle. Rawls argues that a liberty should be classified as basic if and only if it is essential for the adequate development and full exercise of the two moral powers: the capacity for a sense of justice and the capacity for a conception of the good.[35] Such liberties fit into five categories: freedom of thought and liberty of conscience, freedom of association, equal political liberty, rights and liberties protecting the integrity and freedom of the person, and rights and liberties covered by the rule of law.[36]

Rawls notes that the various basic liberties are bound to conflict with one another; consequently, particular liberties can be restricted so that a complete and coherent scheme of liberties is generated.[37] I acknowledge the need for such holistic specification. But, given my present purposes, the rights and liberties protecting the integrity and freedom of the person are particularly important. Such liberties are valuable in themselves and also, as Samuel Freeman notes, because they are instrumental to the exercise of the other basic liberties.[38] To explain Freeman's point, suppose that someone's rights protecting the integrity and freedom of their person are infringed. Then, plausibly, they will also lack the ability to engage effectively in politics and hence fail to have equal political liberty.

It is necessary to make two related distinctions concerning the basic liberties that are salient in nonideal conditions but not in ideal conditions. Both distinctions can be illustrated using the same example. Between 2005 and 2012, the New York Police Department increasingly implemented a "stop-and-frisk" practice. As the name of this practice suggests, it allowed police officers to stop,

33 This assumption is shared by a number of philosophers who extend Rawlsian justice to the regulations of corporate entities outside the basic structure. See Donaldson, *Corporations and Morality*; Donaldson and Dunfee, *Ties That Bind*; and Taylor, "Rawlsian Affirmative Action."

34 For related discussion, see Berkey, "Rawlsian Institutionalism and Business Ethics."

35 Rawls, *Political Liberalism*, 293, 332–33.

36 Rawls, *Political Liberalism*, 291.

37 Rawls, *Political Liberalism*, 295.

38 Freeman, *Rawls*, 56.

question, and frisk pedestrians under a standard of reasonable suspicion.[39] This practice—particularly given that the standard of "reasonable suspicion" was so vague that it could be interpreted to apply to almost any case—clearly violated the basic liberty of freedom and integrity of the stopped person. This is because, as Rawls notes, this basic liberty includes freedom from psychological oppression.[40] And, quite understandably given its nature, the practice induced a great deal of fear, which prevented innocent citizens from exercising this basic liberty in a public space.

The first distinction to note is between the state's official recognition of basic liberties in documents such as a written constitution and the degree to which people are in fact able to exercise their basic liberties because of their reasonable reaction to practices such as stop-and-frisk.[41] A consequence of this distinction is that a state can be in nonideal conditions even if its official recognition of basic liberties conforms to the ideal liberty principle. This is because people may not actually be able to exercise their basic liberties to a fully adequate degree.

Being able to exercise one's basic liberties to a fully adequate degree is a threshold sufficientarian concept. But—as a second, related distinction illustrates—if it is not reached then this can also give rise to certain egalitarian concerns: in nonideal conditions an inability to exercise one's basic liberties is a burden that could fall disproportionately on certain types of people. For example, 90 percent of the people who were stopped and frisked in New York City were Black or Latinx and had committed no crime.[42] Due to the stop-and-frisk practice, African American and Latinx American New Yorkers were *ceteris paribus* less able to exercise their basic liberties than other citizens because of either the direct effects of this practice or the fear that it induced. Such an inequality seems problematic in itself. And, from a Rawlsian perspective, it is also problematic in a deeper sense. For a central Rawlsian commitment is that political liberty must be (at least approximately) equal.[43] But as I noted above, the rights and liberties protecting the integrity and freedom of the person are instrumental to the realization of equal political liberty. Consequently, if, for example, African Ameri-

39 NY Criminal Procedure Law §140.50.

40 See Rawls, *A Theory of Justice*, 53.

41 The caveat of "a reasonable reaction" rules out cases in which actual people feel psychological oppression; however, this oppression should be judged as either psychologically eccentric or stemming from an unjustifiable set of beliefs. Hosein makes a similar move in "Racial Profiling and a Reasonable Sense of Interior Political Status," e6.

42 Center for Constitutional Rights, "Racial Disparity in NYPD Stops-and-Frisks."

43 Rawls, *Justice as Fairness*, 148–49.

can and Latinx American New Yorkers are disproportionately unable to exercise such rights and liberties, then this will also undermine equal political liberty.

Stop-and-frisk practices provide an especially clear illustration of the two distinctions. The distinctions, however, also apply to practices that take place on a more diffuse social level. For example, middle-class African Americans often report that they are avoided like criminals even when they dress in respectable clothing.[44] As Anderson notes, "to be subject as a matter of public reputation to the default presumption of criminal suspicion is … to be publicly dishonored and degraded.… Even those with thick skins and high self-esteem suffer harm to their public standing due to racial stigmatization."[45] Essentially, diffuse racial stigmatization constitutes a type of psychological oppression. African Americans are, consequently, *ceteris paribus* less able to exercise their basic liberties to a fully adequate degree than other citizens because of their reasonable reaction to such psychological oppression.

Some philosophers, such as Iris Marion Young, argue that Rawls's theory of justice is insensitive to many modes of social oppression.[46] The two distinctions, concerning the basic liberties, are not explicitly articulated by Rawls. But, I suggest, they are in keeping with the spirit of his theory; furthermore, once added they help to illuminate how Rawlsian theory can be sensitive to an important type of oppression in nonideal conditions.

4.2. Deriving and Defending a Nonideal Principle of Justice

In order to derive the nonideal principle, I begin by clarifying how the contractualist framework and parties are modeled. The nonideal contracting parties—exactly like Rawls's ideal contracting parties—are rational, in the sense that they want to advance their ends as effectively as possible. The parties are placed behind a veil of ignorance. This veil precludes knowledge of the particular social position that they will actually occupy when the veil is lifted; it thereby prevents the parties from tailoring principles of justice to advance the particular social position they will occupy, such as a particular race or social class.[47]

The nonideal original position has an intergenerational component: the parties are ignorant of when they will be born prior to the realization of ideal justice. This stipulation is introduced to ensure that the path to ideal justice is intergenerationally fair—as opposed to merely intragenerationally fair.[48]

44 See Feagin, "The Continuing Significance of Race."

45 Anderson, *The Imperative of Integration*, 55.

46 See Young, *Justice and the Politics of Difference*.

47 Rawls, *A Theory of Justice*, 17.

48 I remain neutral about how, precisely, this intergenerational component is modeled. I favor

The nonideal contracting parties are presented with the following "nonideal" scenario, which accounts for the two previously drawn distinctions: not all actual people will be able to exercise their basic liberties to a fully adequate degree because of significant noncompliance with justice and/or structural injustice. Furthermore, there is a chance that different people will be unable to exercise their basic liberties to different degrees.[49]

The contracting parties must select a nonideal principle of justice for this nonideal scenario: a principle that it is rational for them to adopt, given that they do not know which social position they will occupy or when they will be born.

Furthermore, they select the principle against the backdrop of three presuppositions. First, they assume that Rawls's ideal principles of justice are correct. Second, they assume that there are sufficient economic resources for it to be possible to increase people's ability to exercise their basic liberties to a significant degree. Third, they assume that all actual people will strictly comply with the nonideal principle of justice that they select.

To clarify this third presupposition, one of the causal reasons that a society can be in nonideal conditions is that actual people have failed to comply fully with the demands of justice. (Stop-and-frisk illustrates this point.) But this causal genesis is compatible with the claim that the contracting parties should assume that actual people will strictly comply with the nonideal principle of justice that they select. This assumption is not realistic. But it is adopted because it allows the contracting parties to select a principle that specifies what justice *simpliciter* requires, without that selection being tainted by actual people's expected noncompliance.[50] Although I make this assumption, I grant that the derived

---

the "narrowing choice" model: the parties know that they all belong to the same generation but they do not know to which generation they belong. They are mutually disinterested and their selection of principles is constrained by certain formal features, such as universality. For a good defense of this model, and a survey of competing models, see Attas, "A Transgenerational Difference Principle."

49 In the context of ideal theory, Rawls describes a four-stage sequence in which the veil is gradually lifted in order to determine principles of justice, then a constitution, then laws, and then the application of laws to particular cases. In this sequence, each stage is guided and constrained by the results of the previous stages (*A Theory of Justice*, 171–74). My nonideal scenario is comparable to the later stages of this sequence, insofar as the veil is partially lifted because more information is introduced. But it is different from the four-stage sequence because more information is introduced in order to determine a *sui generis* nonideal principle of justice rather than to guide the application of ideal principles of justice to things such as the constitution and law.

50 It might be objected that this idealizing assumption of strict compliance is inappropriate in the context of Rawlsian nonideal theory. After all, Rawls sometimes defines nonideal theory as partial compliance theory. See Rawls, *A Theory of Justice*, 215. But Rawls defines nonideal

principle is not directly action guiding. For it is necessary to consider how actual people will react to policies that are supported by the principle in order to assess the efficacy and feasibility of such policies.[51] Accordingly, in section 5 I consider the possibility that affirmative-action policies may have negative stigmatizing effects even if they are just.

I am now in a position to present the (irreducibly baroque) nonideal principle of justice that, I contend, the contracting parties would settle on. After presenting the principle, I will outline the reasoning that leads to its selection. In the principle, "comparative disadvantage/advantage" is with respect to an ability to exercise one's basic liberties to a fully adequate degree because of one's reasonable reaction to other people's noncompliance with justice and/or structural injustice.

1. Measures should be instituted to increase the standing of comparatively disadvantaged people. The required measures are specified by the following clauses.
2. If comparatively disadvantaged people are disadvantaged to different degrees, priority should be given to the most disadvantaged.
3. Comparatively disadvantaged people's standing should be increased in such a way that it imposes as few demands on comparatively advantaged people as possible.
4. If it is not possible to increase the standing of disadvantaged people without imposing costs on comparatively advantaged people, such costs should be distributed according to the following two principles: (i) costs should be imposed evenly on comparatively advantaged people at the same level of advantage; (ii) the relative significance of these costs should be determined by the priority ordering of Rawls's ideal principles of justice (e.g., the basic liberties have priority over equality of opportunity).[52]
5. In determining what measures should be implemented in a particular set of nonideal conditions it *is not* sufficient to consider what measures

---

theory in two different ways: as a conception of what we ought to do in conditions that fail to realize ideal justice, and as partial compliance theory. These two definitions potentially cut against one another as there is no reason to think that the best theoretical account of what we ought to do in conditions that fail to realize ideal justice could not assume strict compliance, at least at some levels of theorizing. I expand on this point in Adams, "The Value of Ideal Theory."

51 For related discussion, see Carroll, "In Defense of Strict Compliance as a Modelling Assumption."

52 Assuming, of course, that such costs are not *so* great that bearing these costs would make people who were antecedently comparatively advantaged more disadvantaged than people who were antecedently comparatively disadvantaged.

> would be the most effective means of increasing the standing of comparatively disadvantaged people at that time. Rather, measures should be introduced that are the most effective means of increasing comparatively disadvantaged people's standing prior to the realization of ideal justice—with priority given to the most disadvantaged, regardless of which generation they are born into. This clause also applies *mutatis mutandis* to the distribution of costs: such costs should be distributed according to clause 4, without discriminating between comparatively advantaged people because they are born into different generations.

Clause 1 is selected because—at the very minimum—the parties want, *ceteris paribus*, to increase the standing of comparatively disadvantaged people in case they end up occupying this unfortunate position. They are, however, also concerned with the costs and opportunity costs of achieving this end. Consequently, they select a set of clauses that specify the parameters under which this end should be achieved.

Clause 2 is chosen because, given the choice problem posed, it is rational for the nonideal contracting parties—like the ideal contracting parties—to be guided by maximin: to instigate measures to guarantee that their social standing is as good as possible in case they end up occupying the position of the most disadvantaged.[53]

Clause 3 is adopted because, although the parties are prepared to impose costs on comparatively advantaged people for the sake of improving the social standing of the disadvantaged, they are not indifferent to the nature of these costs. After all, they could end up occupying the social position of comparatively advantaged people. Therefore, *ceteris paribus*, they prefer for the costs that fall on comparatively advantaged people to be as small as possible.

The first part of clause 4, (i), is adopted because the parties prefer to impose costs on the comparatively advantaged so long as this increases the standing of the disadvantaged and does not bring the overall new standing of the comparatively advantaged down to a level below the new standing of the comparatively disadvantaged. As I noted above, given the choice problem posed, it is rational for the parties to improve the standing of the most disadvantaged—given that they could end up occupying this most disadvantaged position—rather than to produce the best aggregated outcome. They decide that costs should be imposed evenly on comparatively advantaged people at the same level of advan-

53 There is a vast literature discussing both why and whether it is rational for the parties to favor the interests of the comparatively disadvantaged. For a good overview see Gaus and Thrasher, "Rational Choice in the Original Position." I will not, here, attempt to defend Rawls's position further.

tage because they are as likely to become any one of these particular advantaged people as any other; consequently, they want the costs on each comparatively advantaged person to be as small as possible.

The mere fact that the priority ordering structures the ideal principles of justice does not straightforwardly entail that the nonideal contracting parties would also choose for it to structure the nonideal principle, as clause 4 (ii) states. But the salient point concerns the relative ordering of value that the priority ordering reflects: the priority ordering reflects the fact that, for example, the ideal contracting parties attach greater value to the basic liberties than other considerations of justice. This point about value also applies in nonideal conditions in the sense that, for instance, the nonideal contracting parties—like the ideal contracting parties—would attach more value to the basic liberties than other considerations of justice. Consequently, they would prioritize measures to increase their ability to exercise their basic liberties—in case they end up in a position in which their exercise of their basic liberties is compromised—over other possible considerations. Therefore, clause 4 (ii) is endorsed by the parties because the priority ordering of the ideal principles determines the relative importance, or urgency, of different types of injustice in nonideal conditions.[54]

Finally, clause 5 is selected because the parties do not know when they will be born; consequently, they reject measures that privilege the interests of a particular generation prior to the realization of ideal justice.

5. How the Nonideal Principle of Justice Supports Affirmative Action

The nonideal principle of justice supports affirmative action if and only if the following conditions are satisfied:

a. People who are unable to exercise their basic liberties to a fully adequate degree because of their reasonable reaction to other people's noncompliance with justice and/or structural injustice are in that position (at least partly) because they possess the characteristic(s) that affirmative-action policies are designed to target (from clause 1).

54 Following Rawls, *A Theory of Justice*, 216; and Korsgaard, *Creating the Kingdom of Ends*, 148. As noted above, I assume that there are sufficient resources for it to be possible to increase comparatively disadvantaged people's ability to exercise their basic liberties to a significant degree. Robert Taylor helpfully elaborates the threshold of resources that is necessary for the priority ordering of liberty to apply: "a society must have achieved a level of wealth sufficient for it to allow its citizens to engage in meaningful formation of life plans" ("Rawls's Defense of the Priority of Liberty, 263).

b. Affirmative-action policies are a generally effective means of increasing comparatively disadvantaged people's ability to exercise their basic liberties in a way that gives priority to the most disadvantaged and *does not* discriminate between equally advantaged/disadvantaged people who are born into different generations (from clauses 1, 2, and 5).
c. There is not another candidate policy that would be at least as effective a means of increasing comparatively disadvantaged people's ability to exercise their basic liberties, but would impose less significant costs on comparatively advantaged people (regardless of which generation they are born into) as specified by the priority ordering of Rawls's ideal principles of justice (from clauses 3, 4, and 5).

Given the particular empirical conditions that obtain in the contemporary US, I argue that a strong case can be made that the nonideal principle of justice supports affirmative action. For the sake of simplicity, and because there is the most relevant social scientific data on the topic, I will focus on affirmative action for African Americans. (I suggest that a similar conclusion applies for Latinx Americans.)

It is uncontroversial that condition a is satisfied. There is overwhelming evidence that racial discrimination harms African Americans.[55] One of the ways in which it does so, as my previous example of stop-and-frisk illustrates, is to interfere with African Americans' ability to exercise their basic liberties to a fully adequate degree.

Condition b rests on the following presupposition: using affirmative-action policies to ensure that a sufficient threshold of African Americans is placed in certain educational institutions can be a causally effective means of reducing racism of various sorts. And, consequently, given that racism is a causal mechanism that undermines African Americans' ability to exercise their basic liberties—affirmative action can be a causally effective means of increasing African Americans' ability to exercise their basic liberties.

This presupposition can be defended in two main ways. First, contact theory postulates that intergroup contact can reduce prejudice and discrimination if the following four conditions are satisfied: the members of the different groups have equal status (at least in certain relevant respects), they work toward common goals, they engage in cooperation, and the contact is supported and regulated by institutional authority.[56] These conditions are satisfied in educational

55 For a detailed but nontechnical overview of the relevant statistical data, see Anderson, *The Imperative of Integration*, chs. 1–3; and Sterba, *Affirmative Action for the Future*, "Introduction" and ch. 1.

56 Allport, *The Nature of Prejudice*; Dhont, Van Hiel, and Hewstone, "Changing the Ideological Roots of Prejudice"; and Pettigrew, "Intergroup Contact Theory."

institutions, in which students cooperate on equal terms (at least in certain relevant respects) to pursue educational goals that are regulated by institutional norms. Affirmative-action policies ensure that there is a greater representation of African Americans in educational institutions and, thereby, facilitate greater intergroup contact. This helps break down prejudice and discrimination within educational institutions. Clearly, education is a particularly formative time in many people's lives; consequently, the reduction of racial prejudice within educational institutions can have an impact not just within such institutions but also on graduates' subsequent professional and personal lives.[57]

The presupposition underpinning b can be defended in a second way. Affirmative action can reduce discrimination in a far broader sense and, thereby, benefit African Americans who neither attend the particular institutions that practice affirmative action nor directly encounter the graduates of such institutions. Especially on a large intergenerational scale, it can do so by breaking down negative race-based stereotypes and thereby reduce so-called statistical discrimination against all African Americans. To explain, in the US being African American is correlated with variables such as having relatively low educational attainment and social class.[58] People's knowledge about these variables, with respect to particular individuals, is standardly imperfect and increasing this knowledge is costly. Race, however, is a visible feature that can be instantly assessed with relative reliability at almost no cost. Consequently, rational economic actors who do not have any racial prejudices may use race as a proxy for this knowledge. This discrimination is statistical because individuals are judged in terms of the group averages of all African Americans rather than in terms of their individual merits and level of achievements.[59]

Affirmative action can reduce statistical discrimination by helping under-

57 Elizabeth Anderson also uses contact theory to argue that achieving racial integration—in all walks of American life—is an imperative of justice (*The Imperative of Integration*, 123–27). In contrast to Anderson, I defend the more restricted claim that contact theory supports affirmative action in an educational context. There are two major advantages to my approach. First, as Anderson herself acknowledges, contact theory most obviously supports racial integration in formal settings such as educational institutions and the workplace. For in such settings—in contrast to residential neighborhoods—the integration is backed by institutional authority. See Anderson, *The Imperative of Integration*, 123. Second, in contrast to Anderson, I can acknowledge that it is permissible for African Americans to resist integration for a variety of reasons, for example, in a residential context, due to the short-term threat of greater interracial conflict, or out of solidarity with other African Americans. Following Shelby, "Integration, Inequality, and Imperatives of Justice," 267–82.

58 See Jones, Schmitt, and Wilson, "50 Years after the Kerner Commission," 1–8.

59 Following Phelps, "The Statistical Theory of Racism and Sexism"; and Arrow, "Models of Job Discrimination."

mine the rational basis of such discrimination: the general correlation between being African American and having relatively low educational, occupational, and social status. It does this by increasing the number of African Americans studying and, consequently, also ultimately working in institutions of power and prestige. As Dworkin argues, in doing so it thereby decreases the degree of racial identification and by extension statistical discrimination in the US, by reducing the existing correlation between being African American and having an assumed social standing.[60]

It might be objected that when all the effects of affirmative-action policies are taken into consideration they will not satisfy condition b. In particular, many argue that affirmative action has a stigmatizing effect because granting preferential treatment to African Americans implies acknowledging the inferiority of their average strength as applicants.[61] This effect could be broad because it is usually impossible to identify the subset of African Americans who would not have been admitted without affirmative action. Consequently, affirmative action may have the ironic effect of making people view all African American students as inferior.[62] This stigma could prevent affirmative-action policies from satisfying b for two different reasons. First, this stigmatizing effect could undermine my argument that contact theory supports affirmative action. This is because contact theory requires that different groups have *equal* status and affirmative action makes people view African Americans as inferior rather than as equal. Second, the psychological oppression caused by stigmatization could undermine African Americans' ability to exercise their basic liberties.

Much of the evidence in support of this alleged stigmatizing effect is anecdotal.[63] The most comprehensive statistical study on the effects of affirmative action by William Bowen and Derek Bok surveyed over eighty-thousand students at twenty-eight top-tier institutions. It concludes that the effects of stigmatization were comparatively low and that most alumni thought that affirmative action helped reduce stereotypes and mutual animosity.[64] Similarly, Deirdre Bowen's study finds that African Americans experience greater stigma in educational institutions located in states that have banned affirmative action.[65] Thus, it seems

60 Dworkin, *A Matter of Principle*, 294. See also Goffman, *Stigma*.

61 Following Sabbagh, *Equality and Transparency*, 109.

62 See Eastman, *Ending Affirmative Action*; and Scalia, "The Disease as Cure," 219.

63 See, in particular, Clarence Thomas's opinion in Adarand Constructors Inc. v. Peña, 515 US 200 (1995).

64 Bown and Bok, *The Shape of the River*.

65 Bowen, "Brilliant Disguise."

plausible to conclude that although there may be some stigmatizing effect it is not sufficient to prevent many affirmative-action policies from satisfying b.

Note that my nonideal principle specifies the threshold at which a stigmatizing effect would be intolerable. In order for condition b to be satisfied, priority must be given to the most disadvantaged without intergenerational discrimination. It would, therefore, be intolerable for an affirmative-action policy to impose a stigmatizing effect that has the net effect of making people the most disadvantaged. This would be the case even if the affirmative-action policy was a causally effective means of increasing certain less disadvantaged people's ability to exercise their basic liberties, who were members of a different future generation.

Finally, consider condition c. As I will explain in section 6, affirmative action suspends certain features of fair equality of opportunity. This is preferable to an alternative policy that restricts comparatively advantaged people's ability to exercise their basic liberties. This is because c requires that the relative significance of the costs must be determined by the priority ordering of Rawls's ideal principles of justice, and the ideal liberty principle is lexically prior to the FEO principle.[66]

The nonideal principle of justice would, however, support abolishing affirmative action in favor of alternative policies that merely redistribute wealth, if these alternative policies were an equally effective means of increasing African Americans' ability to exercise their basic liberties. This is because such alternative policies would impose a less significant cost on comparatively advantaged people, given that the FEO principle has priority over the distribution of economic goods according to the difference principle.[67]

It might be argued that the nonideal principle supports abolishing affirmative action because any benefit that could be produced by affirmative action could also be produced exclusively by a redistribution of economic goods: in the US there is a significant correlation between being African American and relative poverty. In 2018 African Americans were about 2.5 times as likely to be in poverty compared to white Americans, and in 2016 the median African American family had only 10.2 percent of the median white American family's wealth.[68] This relative poverty, it might be claimed, is the primary cause of African American underrepresentation in academic institutions. But a sufficient redistribution of economic goods would remove this relative poverty and, consequently, the primary cause of African American underrepresentation. Therefore, after sufficient

66 See Rawls, *A Theory of Justice*, 266.

67 Rawls, *A Theory of Justice*, 266.

68 See Jones, Schmitt, and Wilson, "50 Years after the Kerner Commission," 3–4.

economic redistribution, no possible affirmative-action policy could satisfy condition b. This is because a consequence of sufficient economic redistribution is that there would be enough African Americans in academic institutions such that no affirmative-action policy could increase African Americans' ability to exercise their basic liberties.

This argument can be challenged empirically. It is not clear that relative poverty is the primary cause of African American underrepresentation. Susan Mayer, for instance, argues that the degree to which children's educational achievement is dependent on the financial resources of their parents is far weaker than standardly supposed. Indeed, she argues that it has relatively little effect as long as the parents are not in extreme poverty and the basic material needs of their children are met.[69] If this is correct, then mere economic redistribution would (almost certainly) be insufficient for any affirmative-action policy to be *un*able to satisfy b.

More important, suppose for the sake of argument that a sufficient redistribution of economic goods would make it impossible for any affirmative-action policy to satisfy b. Even so, this would only be something that could be achieved in the relatively long term—plausibly, at the very minimum, after there is no significant correlation between being African American and relative poverty for at least one generation. Consequently—at least in the short term, before this is achieved—there is reason to retain affirmative-action policies.

In summary, the nonideal principle of justice supports affirmative action because, at least in the short term, it is an effective and fair means of increasing some African Americans' ability to exercise their basic liberties to a significant degree.

Two clarifications about the scope of this claim are required. First, the claim that the nonideal principle supports affirmative action in the short term is compatible with the claim that actions should be undertaken to make affirmative action unnecessary in the long term. Indeed, suppose that some possible set of policies that merely redistribute wealth would make affirmative-action unnecessary in the long term (for instance, by increasing the funding of predominantly African American high schools or alleviating African American poverty). Then, it would be obligatory to implement this set of policies. For they would impose less significant costs on comparatively advantaged people, as specified by condition c.

Second, note that I use the phrase to a "significant degree" rather than to the "fully requisite degree." Indeed, in my view, affirmative action is a small com-

69 Mayer, *What Money Can't Buy*. See also Satz, "Equality, Adequacy, and Education for Citizenship," 633.

ponent of what is required to enable African Americans to exercise their basic liberties to the fully requisite degree even in the short term. Other important courses of action will include ending practices like stop-and-frisk and the disproportionate mass incarceration of African Americans, which (among other things) interferes with African Americans' ability to exercise their basic liberties because of the stigmatizing effect that it induces.[70]

Still, even if much more than affirmative action is required, this should not distract from the fact that affirmative action can perform the *sui generis* function of reducing racial discrimination within certain educational institutions and a type of statistical discrimination in society as a whole.

6. TWO GOOD-MAKING FEATURES OF MY ARGUMENT

6.1. Blocks the Unfairness Objection

The unfairness objection gains critical traction because when affirmative-action policies are viewed in isolation they appear unfair because individuals such as Simon experience burdens in virtue of their membership in a particular racial group. But this is not sufficient to ground a charge of unfairness. For under the conditions that I have specified, affirmative-action policies reflect a fair distribution of the burdens that are required to transition to a more just society. The explanation for why they are fair can be presented using the nonideal contractualist framework: it would be rational for parties who do not know what social position they will occupy to assent to a principle that condones affirmative action under the specified conditions. Therefore, the policy is impartial—hence fair—in the appropriate sense.

The topic of affirmative action invites reflection on the relationship between "fairness" and "justice." Many use these two concepts interchangeably in everyday speech. However, some philosophical theories of justice render these concepts completely distinct: for instance, a theory in which justice is explicated as "nondomination."[71] Given such a conception of justice, demonstrating that affirmative action is just would not be sufficient to obviate the unfairness objection. Some additional account would have to be supplied in order to explain why considerations of justice trump considerations of fairness.

In contrast, a Rawlsian contractualist framework is uniquely suited to over-

70 In 2000, Human Rights Watch reported that in at least fifteen states African Americans constituted 80 percent to 90 percent of all drug offenders sent to prison, despite the fact that African Americans were no more likely to be guilty of drug crimes than white people (*Punishment and Prejudice*). Following Alexander, *The New Jim Crow*, 98–99.

71 See Pettit, *Republicanism*.

coming the unfairness objection because it presupposes such a tight connection between the concepts of "justice" and "fairness." Rawls describes his view as "justice as fairness" because "the principles of justice are the result of a *fair* agreement or bargain."[72] It would be an exaggeration to say that, in a Rawlsian framework, fairness and justice are synonymous concepts. Even so, if conditions a–c are satisfied, it is difficult to see how a plausible charge of unfairness could be presented against such policies. For they are justified by a nonideal principle of justice that is determined by a *fair* agreement.

6.2. Not Narrowly Focused on Achieving Equality of Opportunity

Rawls's ideal FEO principle states that "those who are at the same level of talent and ability, and have the same willingness to use them, should have the same prospects of success regardless of their initial place in the social system."[73] In the contemporary US this principle is not satisfied because of factors such as extreme poverty and the effects of past racial discrimination. Many Rawlsian and non-Rawlsian philosophers argue that affirmative action can be justified because it helps to counteract the unequal opportunities that are rooted in such past discrimination and, thereby, results in admissions procedures that are closer to the ideal of fair equality of opportunity.[74] Thus, affirmative action suspends features of fair equality of opportunity in the sense that race—which would be arbitrary if the FEO principle were realized—is given preference.[75] This is done, however, in order to realize fairer equality of opportunity in the long term.

Robert Taylor mounts a powerful challenge to this approach. He argues that "we simply cannot know what the counterfactual result of a "clean" competition would look like unless we run one."[76] For under genuinely fair equality of opportunity there could be disproportional group outcomes that are (at least in part) due to cultural reasons, such as Jewish overrepresentation among academics.[77] Given the epistemic opacity of counterfactuals about what the outcomes of genuinely fair equality of opportunity would look like, Taylor argues that we ought to

72 Rawls, *A Theory of Justice*, 11, emphasis added.

73 Rawls, *A Theory of Justice*, 63.

74 See Mason, *Levelling the Playing Field*, 38n29; Miller, *Principles of Social Justice*, 175–76. Interestingly, Samuel Freeman notes that in his lectures Rawls, himself, indicated that affirmative action could be justified in order to remedy the present effects of past discrimination (*Rawls*, 90–91).

75 But for an argument that race-based affirmative action is compatible with Rawlsian ideal justice, see Meshelski, "Procedural Justice and Affirmative Action."

76 Taylor, "Rawlsian Affirmative Action," 494.

77 Taylor, "Rawlsian Affirmative Action," 498.

err on the side of caution. Because "at least for nonconsequentialist liberals, sins of commission should be of much greater concern than sins of omission—especially when the sinner is the state."[78] Therefore, at least in standard cases, we cannot tinker with the result of an admissions procedure, using soft or hard quotas, in order to bring it in line with an outcome that we antecedently judge to be fair. For we lack the epistemic capacity to judge what quotas would reflect the outcomes of genuinely fair equality of opportunity, and we should err on the side of caution.

My account sidesteps Taylor's challenge.[79] For the epistemic judgments required on my account are significantly easier to make in a crucial respect. Innocent group preferences may make it hard to predict the outcome of genuinely fair admissions procedures. But, in contrast, it is not plausible to think that any innocent group preference could account for the fact that some members of certain groups are disproportionately unable to exercise their basic liberties to a fully adequate degree. Essentially, the demands imposed by the liberty principle are considerably less opaque than the demands imposed by the FEO principle [80]

More generally, my account provides a robust rationale for suspending features of fair equality of opportunity precisely because it defends affirmative action in terms of basic liberties rather than fair equality of opportunity. For Rawlsians, there are different features of justice; however, the basic liberties have greater value than all other features of justice. Consequently, my defense of race-based preference in admissions procedures is given the strongest possible Rawlsian defense: it is necessary to promote certain people's ability to exercise their basic liberties.

7. CODA

7.1. Policy Refinement

Although the nonideal principle of justice supports something roughly like the affirmative-action policies in the contemporary US, it also requires some modification of these policies. As I noted at the outset, there is empirical evidence that affirmative action imposes the heaviest burdens on Asian Americans. Yet the nonideal principle of justice does not support this feature of actual affirma-

78 Taylor, "Rawlsian Affirmative Action," 501

79 For a direct response to Taylor, see Matthew, "Rawlsian Affirmative Action"; and Meshelski, "Procedural Justice and Affirmative Action."

80 Of course, this does not mean that my account does not require difficult epistemic judgements about the long-term (perhaps intergenerational) effects of candidate policies. But such epistemic challenges apply to many public policies; consequently, they do not support abolishing affirmative action policies *per se*.

tive-action policies. For it requires that the costs should be distributed *evenly* on comparatively advantaged people at the same level of advantage (clause 4/condition c). And there is no evidence that Asian Americans are in general more comparatively advantaged compared to white Americans, at least with respect to the ability to exercise their basic liberties. Therefore, there is no reason to think that they should have to bear greater burdens of affirmative-action policies than white Americans.

A particularly controversial feature of affirmative-action policies is that they impose benefits on the most privileged subset of African Americans: the upper middle class, many of whom are recent immigrants rather than descendants of slaves.[81] Indeed, many argue that affirmative action is particularly unfair because it gives preference to upper-middle-class African Americans over poor white Americans from Appalachia.[82] Note, however, that the nonideal principle provides a *pro tanto* justification of this comparative preference—at least if this privileged subset of African Americans is in the best position to reduce racial discrimination in certain educational institutions and, by extension, statistical discrimination in society as a whole.

In this context it is important to highlight, however, that although I have focused on the case of affirmative action for African Americans my nonideal principle leaves it as an open empirical question as to whether affirmative action should also be extended to other groups, such as poor white rural Americans. If "poor white rural American" is a social category that inhibits its members' ability to exercise their basic to a fully adequate degree—and affirmative action for such a group would satisfy conditions a–c—then it should be extended.

7.2. The Value of Nonideal Principles of Justice

In recent years, political philosophers have become increasingly self-conscious about philosophical methodology, and an enormous amount has been written on the so-called ideal versus nonideal theory debate.[83] In the context of racial injustice in particular, a number of philosophers, including Anderson, reject the ideal-theory paradigm. Anderson rejects this paradigm for two reasons. First, following Amartya Sen, she argues that it is not *necessary* to work out an ideal theory of justice in order to offer an adequate nonideal treatment of racial injus-

81 At Harvard University, for instance, it is estimated that only about one-third of African American students are from families in which all four grandparents were born in the US. See Rimer and Arenson, "Top Colleges Take More Blacks, but Which Ones?"

82 Hurst, Fitz Gibbon, and Nurse, *Social Inequality.*

83 Following Valentini, "Ideal vs. Non-ideal Theory," 654.

tice.[84] Second, she argues that political philosophers should take an empirically informed "bottom-up" approach that responds to the concrete problems that they face. Given that Anderson opposes ideal theory for this latter reason, she would likely disapprove of my innovation of nonideal principles of justice. Like Rawls's ideal principles of justice, the nonideal principles make very abstract claims about what justice requires; accordingly, they should be classified as a top-down approach to nonideal theorizing.

I suggest that the methodological question of what types of theorizing (e.g., ideal theory) and concepts (e.g., nonideal principles of justice) are practically valuable should be settled by using the following test.[85] Consider which types of theorizing and concepts are required to craft the most compelling philosophical accounts of first-order problems (e.g., when, if at all, affirmative action is just). Essentially, the second-order methodological question should be determined by what adequate treatment of the first-order problems requires.

As I have argued, Anderson is unable to overcome the unfairness objection because she faces the problem of under-theorization. Assuming that I am right that the second-order methodological question should be determined by what is required to address first-order problems, it is incumbent on philosophers like Anderson to show how a nonideal "bottom-up" methodology can provide an adequate response to the unfairness objection. I am not optimistic that this could be achieved: given that the debate about affirmative action hinges on such deep questions about fairness and justice, I suggest that something very like my conceptual innovation of nonideal principles of justice will be indispensable.[86]

Indiana University Bloomington
mra10@iu.edu

84 Anderson, *The Imperative of Integration*, 3. See also Sen, "What Do We Want from a Theory of Justice?" 218–22.

85 I leave open the possibility that some political theory has intrinsic, nonpractical value. See Estlund, "What Good Is It?"

86 Previous versions of this article were presented at the Philosophy, Politics, and Economics Society in New Orleans, San Francisco State University, Stanford University, University of Rochester, University of Virginia, and the Western Political Science Association in San Diego. Thanks to all of the audiences in attendance for their questions and suggestions. I am particularly grateful for incredibly helpful feedback from Marcia Baron, Colin Bird, Talbot Brewer, Eamonn Callan, Hannah Carnegy-Arbuthnott, Randal Curren, Harrison Frye, Laura Gillespie, Johannes Himmelreich, Donncha Maccoil, Kristina Meshelski, Anne Newman, Fay Niker, A John Simmons, and Rebecca Stangl. Finally, I thank the editors of the *Journal of Ethics and Social Philosophy* and a number of anonymous referees for helping me to improve the paper.

REFERENCES

Adams, Matthew. "The Value of Ideal Theory." In *John Rawls: Debating the Major Questions*, edited by Jon Mandle and Sarah Roberts-Cady, 73–88. Oxford: Oxford University Press, 2020.

Alexander, Michelle. *The New Jim Crow: Mass Incarceration in the Age of Colorblindness*. New York: New Press, 2012.

Allport, Gordon W. *The Nature of Prejudice*. Cambridge, MA: Perseus Books, 1954.

Anderson, Elizabeth. *The Imperative of Integration*. Princeton: Princeton University Press, 2013.

Anderson, Nick. "Federal Judge Rules Harvard Does Not Discriminate against Asian Americans in Admission." *Washington Post*, October 1, 2019. https://www.washingtonpost.com/local/education/federal-judge-rules-harvard-does-not-discriminate-against-asian-americans-in-admissions/2019/10/01/dc106b54-a8a1-11e9-a3a6-ab670962db05_story.html.

Appiah, Kwame Anthony. "'Group Rights' and Racial Affirmative Action." *Journal of Ethics* 15, no. 3 (September 2011): 265–80.

Arrow, Kenneth. "Models of Job Discrimination." In *Racial Discrimination in Economic Life*, edited by Anthony Pascal, 83–102. Lexington, MA: Lexington Books, 1972.

Arvan, Marcus. "First Steps toward a Nonideal Theory of Justice." *Ethics and Global Politics* 7, no. 3 (2014): 95–117.

Attas, Daniel. "A Transgenerational Difference Principle." In *Intergenerational Justice*, edited by Alex Gosseries and Lukas H. Meyer, 189–218. Oxford: Oxford University Press, 2012.

Beauchamp, Tom L. "In Defense of Affirmative Action." *Journal of Ethics* 2, no. 2 (1998): 143–58.

Berkey, Brian. "Rawlsian Institutionalism and Business Ethics: Does It Matter Whether Corporations Are Part of the Basic Structure of Society?" *Business Ethics Quarterly* 33, no. 2 (April 2021): 179–209.

Bowen, Deirdre M. "Brilliant Disguise: An Empirical Analysis of a Social Experiment Banning Affirmative Action." *Indiana Law Journal* 85, no. 4 (Fall 2010): 1197–254.

Bown, William G., and Derek Bok. *The Shape of the River: Long-Term Consequence of Considering Race in College and University Admissions*. Princeton: Princeton University Press, 1998.

Boxill, Bernard R. "The Morality of Preferential Hiring." *Philosophy and Public Affairs* 7, no. 3 (Spring 1978): 246–68.

Carroll, Jeffrey. "In Defense of Strict Compliance as a Modeling Assumption." *Social Theory and Practice* 46, no. 3 (July 2020): 441–66.

Center for Constitutional Rights. "Racial Disparity in NYPD Stops-and-Frisks: Preliminary Report on UF-250 Data from 2005 through June 2009." Center for Constitutional Rights, January 15, 2009.

Cohen, Carl. "Why Race Preference Is Wrong and Bad." In *Affirmative Action and Racial Preference: A Debate*, by Carl Cohen and James P. Sterba, 3–181. Oxford: Oxford University Press, 2003.

Dhont, Kristof, Alain Van Hiel, and Miles Hewstone. "Changing the Ideological Roots of Prejudice: Longitudinal Effects of Ethnic Intergroup Contact on Social Dominance Orientation." *Group Processes and Intergroup Relations* 17 (2014): 27–44.

Donaldson, Thomas. *Corporations and Morality*. Englewood Cliffs, NJ: Prentice Hall, 1982.

Donaldson, Thomas, and Thomas Dunfee. *Ties That Bind: A Social Contracts Approach to Business Ethics*. Boston: Harvard Business School Press, 1999.

Dworkin, Ronald. *A Matter of Principle*. Cambridge, MA: Harvard University Press, 1985.

Eastman, Terry. *Ending Affirmative Action: The Case for Colorblind Justice*. New York: Basic Books, 1997.

Espenshade, Thomas J., Chang Y. Chung, and Joan L. Walling. "Admission Preferences for Minority Students, Athletes, and Legacies at Elite Universities." *Social Science Quarterly* 85, no. 5 (December 2004): 1422–46.

Espenshade, Thomas J., and Alexandria Walton Radford. *No Longer Separate, Not Yet Equal*. Princeton: Princeton University Press, 2009.

Estlund, David. "What Good Is It? Unrealistic Political Theory and the Value of Intellectual Work." *Analyse und Kritik* 33, no. 2 (2011): 395–416.

Feagin, Joe R. "The Continuing Significance of Race: Antiblack Discrimination in Public Places." *American Psychology Review* 56, no. 1 (February 1991): 101–16.

Freeman, Samuel. *Rawls*. Oxford: Routledge, 2007.

Gaus, Gerald, and John Thrasher. "Rational Choice in the Original Position: The (Many) Models of Rawls and Harsanyi." In *The Original Position*, edited by Timothy Hinton, 256–91. Cambridge: Cambridge University Press, 2016.

Gerstein, Josh, and Jennifer Haberkorn. "It's Not Just Abortion: 5 Issues Likely to Be Affected by Kennedy's Exit." Politico, June 27, 2018. https://www.politico.com/story/2018/06/27/anthony-kennedy-retirement-supreme-court-cases-680104.

Goffman, Erving. *Stigma: Notes on the Management of Spoiled Identity*. New York: Touchstone, 2009.

Hodgson, Louis-Philippe. "Why the Basic Structure?" *Canadian Journal of Philosophy* 42, nos. 3–4 (September/December 2012): 303–34.

Hosein, Adam Omar. "Racial Profiling and a Reasonable Sense of Interior Political Status." *Journal of Political Philosophy* 26, no. 3 (September 2018): e1–e20.

Human Rights Watch. "Punishment and Prejudice: Racial Disparities in the War on Drugs." *HRW Reports* 12, no. 2 (May 2000).

Hurst, Charles E., Heather M. Fitz Gibbon, and Anne M. Nurse. *Social Inequality: Forms, Causes, and Consequences*. 9th ed. New York: Routledge, 2016.

Jones, Janelle, John Schmitt, and Valerie Wilson. "50 years after the Kerner Commission." Economic Policy Institute, February 26, 2018. https://www.epi.org/publication/50-years-after-the-kerner-commission/.

Korsgaard, Christine. *Creating the Kingdom of Ends*. Cambridge: Cambridge University Press, 1996.

Lippert-Rasmussen, Kasper. "Affirmative Action, Historical Injustice, and the Concept of Beneficiaries." *Journal of Political Philosophy* 25, no. 1 (2017): 72–90.

Lynch, Frederick R. *Invisible Victims: White Males and the Crisis of Affirmative Action*. New York: Praeger, 1991.

Mason, Andrew. *Levelling the Playing Field: The Idea of Equal Opportunity and Its Place in Egalitarian Thought*. Oxford: Oxford University Press, 2006.

Matthew, D. C. "Rawlsian Affirmative Action." *Critical Philosophy of Race* 3, no. 2 (2015): 324–43.

Mayer, Susan. *What Money Can't Buy: Family Income and Children's Life Chances*. Cambridge, MA: Harvard University Press, 1999.

Meshelski, Kristina. "Procedural Justice and Affirmative Action." *Ethical Theory and Moral Practice* 19 (2016): 425–44.

Miller, David. *Principles of Social Justice*. Cambridge, MA: Harvard University Press, 2001.

Mills, Charles W. *The Racial Contract*. New York: Cornell University Press, 1999.

Morris, Christopher W. "Existential Limits to the Rectification of Past Wrongs." *American Philosophical Quarterly* 21, no. 2 (April 1984): 175–82.

Moses, Michele S. "After *Fisher*: Lawsuit against Harvard Goes On." *New Boston Post*, July 19, 2016.

Nagel, Thomas. "Equal Treatment and Compensatory Discrimination." *Philosophy and Public Affairs* 2, no. 4 (Summer 1973): 348–63.

Pettigrew, Thomas F. "Intergroup Contact Theory." *Annual Review of Psychology* 49 (1998): 65–85.

Pettit, Philip. *Republicanism: A Theory of Freedom and Government*. Oxford: Oxford University Press, 1997.

Phelps, Edmund S. "The Statistical Theory of Racism and Sexism." *American Economic Review* 62, no. 4 (September 1972): 659–61.

Pojman, Louis P. "The Case against Affirmative Action." *International Journal of Applied Philosophy* 12, no. 1 (1998): 97–115.

Rawls, John. *Justice as Fairness: A Restatement*. Cambridge, MA: Belknap Press, 2003.

———. *The Law of Peoples*. Cambridge, MA: Harvard University Press, 2001.

———. *Political Liberalism: Expanded Edition*. New York: Columbia University Press, 2005.

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

Rimer, Sarah, and Karen W. Arenson, "Top Colleges Take More Blacks, but Which Ones?" *New York Times*, June 24, 2004. https://www.nytimes.com/2004/06/24/us/top-colleges-take-more-blacks-but-which-ones.html.

Sabbagh, Daniel. *Equality and Transparency: A Strategic Perspective on Affirmative Action in American Law*. New York: Palgrave MacMillan, 2007.

Satz, Debra. "Equality, Adequacy, and Education for Citizenship." *Ethics* 117, no. 4 (July 2007): 623–48.

Scalia, Antonin. "The Disease as Cure: 'In Order to Get beyond Racism, We Must First Take Account of Race.'" *Washington University Law Quarterly* 147 (1979): 147–57.

Schmidt, Peter. "A History of Legacy Preferences and Privilege." In *Affirmative Action for the Rich: Legacy Preferences in College Admissions*, edited by Richard D. Kahlenberg. New York: The Century Press, 2010.

Sen, Amartya. "What Do We Want from a Theory of Justice?" *Journal of Philosophy* 103, no. 5 (May 2006): 215–38.

Shelby, Tommie. "Integration, Inequality, and Imperatives of Justice: A Review Essay." *Philosophy and Public Affairs* 42, no. 3 (Summer 2014): 253–85.

———. "Race and Ethnicity, Race and Social Justice: Rawlsian Considerations." *Fordham Law Review* 72, no. 5 (2004): 1697–1714.

Simmons, A. John. "Ideal and Nonideal Theory." *Philosophy and Public Affairs* 38, no. 1 (Winter 2010): 5–36.

Sterba, James. *Affirmative Action for the Future*. Ithaca, NY: Cornell University Press, 2009.

———. "Defending Affirmative Action, Defending Preference." In *Affirmative Action and Racial Preference: A Debate*, by Carl Cohen and James P. Sterba, 189–267. Oxford: Oxford University Press, 2003.

Taylor, Robert S. "Rawlsian Affirmative Action." *Ethics* 119, no. 3 (April 2009): 476–506.

———. "Rawls's Defense of the Priority of Liberty: A Kantian Reconstruction." *Philosophy and Public Affairs* 31, no. 3 (Summer 2003): 246–71.

Thomson, Judith Jarvis. "Preferential Hiring." *Philosophy and Public Affairs* 2, no. 4 (Summer 1973): 364–84.

United States Commission on Civil Rights. "Civil Rights Issues Facing Asian Americans in the 1990s." February 1992. https://www.ojp.gov/pdffiles1/digitization/135906ncjrs.pdf.

Valentini, Laura. "Ideal vs. Non-ideal Theory: A Conceptual Map." *Philosophy Compass* 7, no. 9 (September 2012): 654–64.

Wiens, David. "Motivational Limitations on the Demands of Justice." *European Journal of Political Theory* 15, no. 3 (2016): 333–52.

Young, Iris Marion. *Justice and the Politics of Difference*. Princeton: Princeton University Press, 1990.

Journal of Ethics and Social Philosophy
VOL. 20, NO. 3 · NOVEMBER 2021
https://doi.org/10.26556/jesp.v20i3.1010

REALISM, METASEMANTICS, AND RISK

Billy Dunaway

Does realism about a subject matter entail that it is especially difficult to know anything about it? In broad outline the motivation for an affirmative answer to this question is a natural one: since realism (on a superficial gloss) holds that a domain exists independently of what we think, experience, or feel, it is possible for beliefs about the domain to diverge systematically from the facts. Realism about a subject matter, then, entails that the relevant facts are independent of us in a way that allows for widespread and systematic error in our beliefs about them.

This is just an initial gloss on a common and natural view about the epistemic consequences of realism. It is worth emphasizing at the outset that the alleged problem is a *substantive* connection between a kind of metaphysical thesis that is associated with the term "realism" and an epistemological claim according to which we lack—in a sense to be made more precise below—epistemic access to the facts, realistically construed. Thus I am conceiving of realism broadly as a metaphysical thesis. By way of contrast, some approaches to realism take the thesis to entail a lack of epistemic access *by definition*.[1] I am not working with these epistemic characterizations of realism here. Instead the question is whether the metaphysics of realism nontrivially implies that a range of beliefs are epistemically defective, and the aim is to investigate whether this argument constitutes a powerful epistemological consideration against the realist thesis.

"Realism" is a term that can apply (or fail to apply) to views about a wide variety of domains: physical objects, scientific unobservables, and mental states are all examples. The skeptical consequences of realism I sketched above allegedly follow from realism regardless of subject matter. I will focus, in what follows, on one particular domain where the skeptical consequences of realism are especially pressing and have received significant discussion. This is the domain of morality, and normativity more generally.[2] While the central characteristics of

1 See, e.g., Dummett, "Realism"; and Wright, *Truth and Objectivity*.

2 Examples of epistemological arguments against realism in this domain include Harman's claim that moral facts do not causally explain why we form moral judgments (Moral Expla-

normative language that give rise to epistemic difficulties for realism about normativity may also be present given realism about other domains, I will not raise this question here. It is, however, worth keeping in mind the question of whether the skeptical argument developed here applies to realism about other domains.

The argument I will develop has at its center the notion of *metasemantic risk*: this is the idea that our normative terms could have, in suitably different conditions, referred to something besides the normative properties they actually pick out.[3] I will outline how a particular kind of metasemantic risk follows from some core commitments of any plausible version of normative realism. And this kind of risk has consequences for knowledge and epistemic justification. I spell out these connections in an argument I call the *Argument from Risk*, and I will use the argument to explore the epistemology of realism.

1. Risk: Metasemantic and Epistemic

The short version of an argument that connects metasemantic risk and the absence of knowledge uses two technical terms, which I will describe briefly here, before adding more detail when evaluating the argument. *Metasemantic risk* refers to the possibility of a shift in the reference of a term. So while the normative term "ought" actually refers to obligation, there is some metasemantic risk in "ought" because the term could have referred to something else. That is, if "ought" is metasemantically risky, there is a possible linguistic community that speaks a language that is similar to English but differs from actual English enough that their "ought" refers to something distinct from obligation. Exactly how significant the risk is—that is, how similar this possible community where "ought" shifts reference is to our own—is a question I discuss below.

The second technical term is *epistemic risk*. A belief is subject to epistemic risk when it is at risk of being false in a way that is incompatible with that belief being knowledge. If I believe that I ought to keep my promise, but the belief is at risk in the relevant (epistemic) sense, then I could have had a false belief about my

nations of Natural Facts"); Street's argument that it is compatible with a naturalistic evolutionary process that we make moral judgments according to any of a wide variety of mutually incompatible moral systems ("A Darwinian Dilemma for Realist Theories of Value"); and Mackie's "argument from queerness," which (in one form) claims that moral properties are massively different in kind from any other property we know about (*Ethics*). I will not address these arguments here, and instead aim to present a distinct epistemological worry for realism.

3 Hawthorne makes use of a related notion ("*A Priority* and Externalism"). While the context of Hawthorne's discussion is a slightly different one—his focus is on a characterization of *a priori* knowledge, not epistemological arguments against realism—the present paper owes much to Hawthorne's discussion.

promise-keeping obligations. It follows, by definition, that I do not know that I ought to keep my promise.

The Argument from Risk connects metasemantic risk with epistemic risk for normative beliefs. According to the argument, metasemantic risk for normative terms is a commitment of realism about normativity. Epistemic risk in normative beliefs is an alleged consequence of metasemantic risk. In worlds where "ought" has shifted reference, some agents will have false normative beliefs, and this makes normative beliefs suffer from epistemic risk. So metasemantic risk implies that normative beliefs are not knowledge.

1.1. The Argument and an Illustration

The main premises in this argument are as follows:

1. "Ought" is metasemantically risky.
2. If "ought" is metasemantically risky, then one could easily be in a world where "ought" does not refer to obligation.
3. If one could easily be in a world where "ought" does not refer to obligation, then one could easily have had a false normative belief.
4. If one could easily have had a false normative belief, then one's actual normative beliefs are at epistemic risk and are not knowledge.

In the rest of this paper, I will spell out why the Argument from Risk is not a straightforward instance of a general argument that can be applied without modification to any domain. Instead, its premises are in a number of ways very plausible when their subject is *normative* belief, because of some unique features of our normative thought and language. Further, some of the premises in the Argument from Risk will be much more plausible to someone who adopts a *realist* view about the metaphysics of normativity: in fact, I will argue that realists must accept some of the premises in order to successfully respond to other arguments against realism in the literature. None of this is to deny that metasemantic risk may give rise to epistemological worries in other domains as well. But the rationale behind such worries will not necessarily be analogous to the support I offer for the premises in the Argument from Risk here.

Before turning to an evaluation of the Argument from Risk, we can begin with a concrete case where metasemantic risk appears to give rise to epistemic risk. (The case is loaded with theoretical assumptions that I will discuss later; the purpose here is only to provide an intuitive illustration of the epistemological problems that arise if the assumptions are correct.) Suppose that our community uses "ought" to refer to obligation, and moreover that among the obligatory actions is the act of giving 10 percent of one's annual income to charity. But giving 25 per-

cent of one's income to charity is not, we can suppose, obligatory. Because uses of "ought" are subject to metasemantic risk, there are worlds where our use of "ought" changes slightly. Among these worlds there are some where the linguistic situation is such that our use of "ought" refers to a property distinct from obligation—call it *obligation**—that has giving away 25 percent of one's income in its extension.

Our term "ought" refers to obligation. For reasons I will discuss below, it is plausible that "ought" also refers to obligation in many other possible worlds that we could easily have been in, which differ only slightly from our world with respect to how we use our word "ought." But the presence of metasemantic risk means that in some worlds, "ought" refers to a distinct property. Perhaps in these worlds it refers to a property that applies to acts that would be best, without regard to whether these are acts that an agent can realistically perform. In such worlds speakers insist that "ought" applies not only to acts that a speaker can reasonably be expected to perform but also to acts that, regardless of physical limitations of actual agents, would be best if they were to occur. Such acts are obligatory*. The claim that "ought" is metasemantically risky would be witnessed by a possible community that manages to use their normative "ought" to refer to obligation*.

Giving 25 percent of one's income to charity is, while not obligatory, obligatory*. In a world where "ought" refers to obligation*, one speaks falsely if one says "giving 25 percent of one's income is not something one ought to do." One also *believes* something false if one forms the belief in the proposition this sentence expresses in the world in which "ought" has shifted reference. Since we could be in such a world, if the Argument from Risk is sound, the normative belief that one ought to give 10 percent of one's income to charity is subject to knowledge-destroying epistemic risk.

1.2. A Program for Filling in the Argument from Risk

Every term in our language is capable of referring to something other than what it refers to in English, since there is no intrinsic connection between a string of letters or phonemes and the reference-determining features of the term. We could have used "ought" as we actually use "cat"; if we did, "ought" would not be a normative term in our language. This is not (an interesting form of) metasemantic risk. The Argument from Risk requires the possibility of semantic shifts that could easily have happened, and which, if they did happen, would give rise to false *normative* beliefs.

There are independently plausible theses about how we use normative language, and how we form our normative beliefs, that make the semantic shifts in normative language more interesting than a generic case of change in reference. In broad outline, one distinctive feature of normative language is that it

is possible to use normative language in many different ways, without making a conceptual mistake: there are, for example, possible communities of speakers that coherently apply their normative "ought" to acts of selfishly keeping one's money for oneself. This is an extreme difference between us and other coherent users of normative language, but there are all kinds of differences in between: some apply their "ought" to acts of keeping one's money when the benefit to oneself is extremely large; others make slightly less demanding exemptions, and so on. Each of these communities might still be motivated to do what they say they "ought" to do in the right way, and each community is not conceptually confused, so their "ought" will still have the role of a normative term. I will say that possible uses of normative language are *modally continuous*.

It is natural to add a second claim to this, which I will argue below is a commitment of any defensible version of normative realism on independent grounds. While uses of normative language can differ in all kinds of ways, many possible communities that use their "ought" as a normative term still manage to refer to *obligation*, rather than some distinct property. The community that says "one ought to be selfish and not donate any money to charity" manages to say that not donating has the property of *being obligatory*, the same property that we refer to when we say "one ought to give 10 percent of one's income to charity." They say something false about the same thing we are talking about, rather than saying something potentially true about a distinct property that fits their use better. I will say that normative terms are *semantically stable*.

However, there are limits to the amount of stability in any term, and normative terms are no exception. Thus I will say that normative terms are *moderately* semantically stable, since there are some possible communities that use their "ought" in ways that make it refer to something distinct from obligation. This is a significant assumption, and I return to it below.

Third, we and other possible linguistic communities can rely on the meaning of our public language term "ought" to form beliefs about what it refers to—whether this is obligation or some other property. That is, when one is in a community whose "ought" refers to obligation and accepts the sentence "one ought to give to the poor," one will typically have the corresponding normative belief that giving to the poor is obligatory. If one were to be in a part of a possible community whose normative "ought" refers to obligation*, then in accepting the sentence "one ought to give to the poor," one would typically have the belief that giving to the poor is obligatory*.

Just as the Argument from Risk needs a refined notion of metasemantic risk, it also needs refinements to the notion of *epistemic* risk. The generic notion of *risk* concerns what happens in nearby worlds, or worlds that could easily have ob-

tained.[4] Risking a false belief is, subject to refinements, incompatible with knowledge. Similarly, one way to lose justification for a belief is to learn that it could easily have been false in a sense that is incompatible with knowledge. So metasemantic risk will plausibly have important epistemological consequences for both normative knowledge and normative justification. But epistemic risk is not simply a matter of having a false belief in a nearby world. Once we add the needed refinements, the ways in which normative beliefs are metasemantically risky will make the Argument from Risk more compelling in the case of normative belief specifically.

2. Premise 1: Metasemantic Risk

Premise 1 in the argument from risk says:

1. "Ought" is metasemantically risky.

Metasemantic risk takes the following form for normative terms: they are stable, but the stability is only moderate. In this section I will sketch the motivations for both parts of this premise.

There are several considerations that suggest that stability is an explanatory *desideratum* for a realist view.[5] I will focus on the Moral Twin Earth case, but recent literature adds further considerations in favor of stability.

2.1. Moral Twin Earth

In a series of papers including "Troubles on Moral Twin Earth" and "Troubles for New Wave Moral Semantics," Horgan and Timmons argue that certain versions of realism cannot explain the range of disagreement between possible communities that use moral language. They argue against a version of realism that includes a causal theory of reference, due to Boyd, by describing two possible communities whose use of moral vocabulary is causally related to different properties but who nonetheless appear to disagree about morality:

> Earthlings' moral judgments and moral statements are causally regulated

4 Cf. "safety" principles in Sosa, "How to Defeat Opposition to Moore"; Williamson, *Knowledge and Its Limits*; and Pritchard, *Epistemic Luck*.

5 Note that even if stability is an explanatory *desideratum*, it does not follow that every realist view in fact explains it. Some views may acknowledge stability as an explanatory goal and treat it as a cost if they fail to explain its full range. Railton, "Moral Realism," is an example of a realist view that acknowledges the limits to the range of stability it predicts for moral terms. That a view has some theoretical costs is not, on its own, a decisive reason to reject it; Enoch emphasizes this methodological point (*Taking Morality Seriously*). However, I will assume that the realist does not have to concede stability as a point in favor of competing views.

> by some unique family of functional properties, whose essence is functionally characterizable via the generalizations of a single substantive moral theory. Suppose, too, that this theory is discoverable through moral inquiry employing coherentist methodology. For specificity, let this be some sort of consequentialist theory, which we will designate T^c.
>
> Now for Moral Twin Earth. Its inhabitants have a vocabulary that works very much like human moral vocabulary; they use the terms "good" and "bad," "right" and "wrong," to evaluate actions, persons, institutions, and so forth (at least those who speak twin English use these terms, whereas those who speak some other twin language use terms orthographically identical to the corresponding moral terms in the corresponding Earthly language). But on Moral Twin Earth, people's uses of twin-moral terms are causally regulated by certain natural properties distinct from those that (as we are already supposing) regulate English moral discourse. The properties tracked by twin English moral terms are also functional properties, whose essence is functionally characterizable by means of a normative moral theory. But these are *non-consequentialist* moral properties, whose functional essence is captured by some specific deontological theory; call this theory T^d.[6]

Horgan and Timmons think that when we consider communities like these, it is clear that they disagree about morality: "Here the question about what really is the fundamental right-making property seems to be an open question, and one over which Earthlings and Twin Earthlings disagree."[7]

There are a few details that are needed to turn the Moral Twin Earth case into an argument for stability for normative terms. First, although Horgan and Timmons are explicitly concerned with moral terms like "good," similar points apply to normative vocabulary like the all-things-considered "ought."[8] Second, the intuition of disagreement is not limited to the single case involving the Earthlings and Moral Twin Earthlings. As Horgan and Timmons emphasize elsewhere, similar cases can be described involving other pairs of possible communities whose use of moral vocabulary differs in other ways, aside from being causally regulated by different properties.[9] Finally, a realist should want to explain the

6 Boyd, "How to Be a Moral Realist"; Horgan and Timmons, "Troubles on Moral Twin Earth," 245.

7 Horgan and Timmons, "Troubles on Moral Twin Earth," 248.

8 Cf. Dunaway and McPherson, "Reference Magnetism as a Solution to the Moral Twin Earth Problem."

9 Horgan and Timmons, "Copping Out on Moral Twin Earth." How far the disagreements extend is an interesting question. I will address this point below.

disagreement by explaining how it is that each community in a Moral Twin Earth scenario is referring to the same property, and thereby makes a claim that is incompatible with claims made by speakers in the other community.

These details together motivate a realist view that entails a measure of semantic stability for normative terms. Since, for a realist, these disagreements should be explained in part by a metasemantic theory that entails that each community is referring to the same property, it follows from an adequate realist treatment of Moral Twin Earth cases that "ought" is semantically stable.[10]

Similar lessons emerge from more recent discussions of normative objectivity and knowledge, which I will mention only briefly. One comes from what I will call a "symmetry argument," found in Eklund's *Choosing Normative Concepts*. Eklund's "Bad Guy" is a possible user of a normative "ought" who, like the Twin Earthlings in the Moral Twin Earth scenario, uses the term differently from actual users. Bad Guy ends up saying different things than we say, applying "ought" for example to acts of stealing from the poor. Moreover, Bad Guy acts as we would expect for someone who applies a normative term to such acts.

If Bad Guy were speaking *truly*, by referring to a property that is distinct from obligation, there would be a kind of symmetry between him and us, since each of us speaks truly by using our normative "ought" and acts accordingly. This, according to Eklund, should be troubling for the realist: Bad Guy does things that he ought not to do. There should be some grounds for criticizing him. But, as Eklund points out, Bad Guy can make symmetrical criticisms of us: we fail to do some things that are obligatory*, and Bad Guy, in his language, speaks truly when he says "they fail to do some things that are obligatory."[11]

10 This is a commitment that is specific to realism: a noncognitivist or expressivist might explain the disagreements differently, cf. the notion of "disagreement in plan" in Gibbard, *Thinking How to Live*.

11 Here is Eklund:

> We can still say that Bad Guy doesn't do what he all-things-considered ought to do or has reason to do. But using his language, Bad Guy can say the corresponding things about us. Using his counterpart of "wrong"—the word in his vocabulary that has the role for him that "wrong" has for us—he can say that we do "wrong" things. And he is as correct in his verdict about us as we are in our verdict about him. The same would go for all other normative vocabulary.... Despite all the realist trappings that our normative language is supposed to have, there may still for all that has been said be *parity* between us and Bad Guy that the ardent realist would want to avoid. For all that has been said, Bad Guy is not objectively mistaken about anything; he just does not employ our notion of reason or our notion of what ought to be done but instead employs alternative normative notions. (*Choosing Normative Concepts*, 5)

Clarke-Doane, *Morality and Mathematics*, explores additional worries along similar lines.

I have not argued here that a realist can meet these explanatory *desiderata*, and I will, in what follows, simply assume that they can be met.[12] The Argument from Risk claims that it follows that realism faces a further problem: by adequately accounting for stability, realism introduces an epistemological difficulty—namely, the inability to know normative claims. Thus the Argument from Risk captures a distinct (alleged) problem for realism since it *assumes* that the realist can explain the moderate stability of normative language that is raised by Horgan and Timmons, and claims that it therefore fails to explain how normative knowledge is possible.[13]

2.2. Interlude: Realism and Stability

Stability is an explanatory *desideratum* for the realist. Nonrealists, such as expressivists who follow Gibbard, can also endorse stability for normative terms.[14] In this case the endorsement will be explained with nonrealist resources: for example, in Gibbard's terms, the explanation will involve the thesis that claims about reference are *plan-laden*.[15] So simply explaining stability does not make a view realist; whatever additional metaphysical claims are needed for realism will need to be compatible with stability for normative terms. The Gibbardian explanation in terms of normative beliefs as planning states will not do for the realist.

For the sake of illustration, here is one version of a theory that can explain the stability datum for normative terms for realists. The theory consists in the metaphysical claim that some properties are metaphysically privileged or *elite* properties and the metasemantic claim that elite properties are easier to refer to

12 Elsewhere I argue that the realist has the resources to meet this challenge, together with other explanatory *desiderata* I describe below. See Dunaway, *Reality and Morality*.

13 Eklund does raise an epistemological issue that emerges from his discussion of Bad Guy. The issue is not what he calls "run-of-the-mill" skepticism concerning knowledge of what we ought to do (*Choosing Normative Concepts*, 14). Instead, Eklund's "normative skepticism" is a special question, since he grants that we can know what we ought to do but is less sanguine about our ability to know, roughly, that we should care about what we ought to do (as opposed to, say, caring about what Bad Guy is talking about with his normative terms). This is just a gesture at what Eklund's epistemological issue is, since Eklund thinks that the real question may be "ineffable" and so knowing its answer, which settles what is at issue between us and Bad Guy, will be difficult if not impossible (25).

This is an interesting question, but it is not the one I will address in what follows: I am concerned with "run-of-the-mill" skepticism, which is our inability to know what we ought to do. Moreover, I will not be presupposing any of Eklund's discussion of the existence of a further, ineffable issue that is under dispute between us and Bad Guy, as Eklund does. The epistemological problems I raise here can be raised even if we reject the existence of such an issue.

14 Gibbard, *Meaning and Normativity*.

15 See Dunaway, "Expressivism and Normative Metaphysics," for elaboration.

than nonelite properties. These claims together are sometimes labeled *reference magnetism* and are outlined by David Lewis.[16] This is the view that terms refer to what they do not solely in virtue of how they are used but also in virtue of how the world is: some speakers' use of a term best fits a particular property *P*, but there is an elite property *P**, similar to *P*, that fits their use pretty well and is metaphysically privileged over *P*. These speakers are, according to a metasemantic theory that includes reference magnetism, referring to *P**. This is the sense in which elite properties are easy to refer to: we can refer to elite properties without tailoring our usage to fit them precisely.

If the normative realist holds that obligation is a metaphysically elite property, then obligation (modulo the additional assumption that there are no additional elite normative properties in the vicinity of it) will be easy to refer to for speakers that use normative language. Views along these lines have been sketched by Van Roojen, Edwards, and Dunaway and McPherson.[17] I will not rehash the details here, but the theory illustrates one substantive set of claims that appears to explain stability for the realist. It is, plausibly, not an unmeetable explanatory *desideratum*.

2.3. *Moderate Stability*

The normative term "ought," I will assume, is metasemantically risky because the stability is only *moderate*. "Ought," as used by some possible communities, refers to some property other than obligation. But what makes the stability merely moderate is the fact that some of these possible communities that use their "ought" to refer to something other than obligation still use it as a *normative* term. Roughly, this means that they use the term to settle what to do and to close off deliberation. When one uses "ought" in this way, it is not coherent to conclude that one "ought" to act in a certain way and yet fail to do so.[18] This is to use "ought" with a *normative role*.

Some additional notes about this assumption are in order.

First, the standard motivations for stability do not motivate a thesis that is stronger than moderate stability. The standard examples of possible communities that disagree in Moral Twin Earth scenarios all involve communities that appear to accept different substantive theories about what wrongness, or obligation, consists in. (The usual example involves consequentialists and deontolo-

16 Lewis, "New Work for a Theory of Universals" and "Putnam's Paradox."

17 Van Roojen, "Knowing Enough to Disagree"; Edwards, "The Eligibility of Ethical Naturalism"; and Dunaway and McPherson, "Reference Magnetism as a Solution to the Moral Twin Earth Problem."

18 Cf. Gibbard, *Meaning and Normativity*.

gists, but examples can be multiplied by imagining that the communities accept different first-order ethical theories.) However, there are other ways for possible communities to differ while still using their "ought" with a normative role. Generalization off the usual cases will not give us any reason to think that all of these possible communities will be referring to obligation.

As a second point, we can give an example of this kind of community that appears to (i) use their "ought" with a normative role and (ii) refer to something distinct from obligation. Suppose that, as a matter of fact, our "ought" does not apply to single-handedly ending a large famine. There are a number of possible communities that do apply their "ought" to ending the famine. Some of these communities fit the model of the Horgan and Timmons Moral Twin Earth case, where the community simply has a different substantive theory of obligation. But there are other possibilities: the community could value the same things we value, but not use their normative "ought" as if it were governed by the principle "ought" implies "can." If the difference between English speakers and speakers in this possible community is that the latter say "one ought to end the famine single-handedly," and this difference arises because they do not restrict application of their "ought" to actions that can be performed, then the difference in how they use their "ought" need not be a difference over substantive first-order theory. The difference between them and actual English speakers does not constitute a dispute over what makes an action obligatory.[19]

All of this is compatible with both communities—including the community that does not have an "ought" that is governed by the "ought" implies "can" principle—using their "ought" with a normative role.[20] Cases like the community in the previous example suggest that it is not at all clear that a community should be interpreted as referring to obligation simply because they use "ought" with a normative role. English speakers can agree that options they are unable to perform have a property that resembles the property that makes actions obligatory. It is not inconceivable that some other possible communities have normative terms that refer to such normatively significant properties that are distinct from obligation.

A final point is that in order to deny that "ought" is only moderately stable, one has to adopt an extreme metasemantic view. Such views exist in the literature: Wedgwood and Williams hold a version of the "conceptual role determinism for wrongness" thesis.[21] Likewise, Eklund takes such a thesis seriously

19 See Dunaway, *Reality and Morality*, chs. 1–2, for more discussion.

20 They will conclude that they "ought" to do some things that they inevitably fail to do. But this is not incompatible with the term having a normative role in their language; they are simply forced to conclude that they regularly display a kind of incoherence.

21 Wedgwood, "Conceptual Role Semantics for Moral Terms" and *The Nature of Normativity*;

without outright endorsing it.[22] I will not argue against these views here, but it is worth noting that they require a highly unusual metasemantic contribution from a term's normative role. This is because on each view, the fact that "ought" is used with a normative role is sufficient for "ought" to refer to obligation; all other facts about how the term is used are irrelevant to what it refers to.[23]

Perhaps normative language is special in this way. But theories that entail that normative role determines reference are more ambitious than what the intuitive data supports, and they treat normative role as a privileged reference-determining feature of normative terms; I will assume without further argument here that the stability of normative terms does not extend as far as these theories imply.[24] However, for readers who are not inclined to grant this assumption, the Argument from Risk promises to provide additional motivation for such a strong metasemantic assumption. If the argument is successful, it shows that there are *epistemic* considerations that favor a strong metasemantic commitment to stability, in addition to the usual arguments that motivate such views.

3. Premise 2: Risk of Shifted Reference

The first step in the Argument from Risk is to show that metasemantic risk for normative terms exists. The second step is to claim that a reference shift could easily have happened. This is what premise 2 claims:

2. If "ought" is metasemantically risky, then one could easily be in a world where "ought" does not refer to obligation.

It is worth emphasizing that while reference shifts are possible for other nonnor-

Williams, "Normative Reference Magnets" and *The Metaphysics of Representation*. Quote from Williams, "Normative Reference Magnets," section 2.3.

22 Eklund, *Choosing Normative Concepts*.

23 For example, with color terms, reference is partly determined by something like conceptual role: it is part of the conceptual role of "red" that it refers to a color that is darker than what "orange" refers to and lighter than what "purple" refers to. But this alone does not settle that "red" refers to redness—cf. the "permutation problem" in Smith, *The Moral Problem*. Additional facts about how "red" is used by English speakers, including the fact that they regularly apply their "red" to red objects, are relevant.

24 Note that conceptual determinism for wrongness (or obligation) is extreme from a realist perspective, when obligation and other normative properties are real properties that serve as candidate referents for a term or concept. In this case there are alternative properties that are candidate referents; explaining why a term used with a normative role never refers to these alternative candidate referents, regardless of the additional features of its use, is a substantial task. Eklund calls such a view "metasemantically radical" (*Choosing Normative Concepts*, 167).

mative portions of our vocabulary, these shifts are much less threatening from an epistemological perspective.

Reference shifts for color terms like "red" are, in a sense, much more widespread than shifts for "ought." Any community-wide shift in usage is guaranteed to produce a different referent; color terms are not stable, but rather are semantically *plastic*.[25] If we had applied "red" to a slightly different range of light wavelengths (and made corresponding adjustments in our use of other color terms), then we would have referred to something other than redness. The analogue of semantic stability for "red"—that we would have continued to refer to redness and would be making false claims about which objects are red—is not plausible.

These shifts do not threaten our knowledge of colors. In most cases I will not be prevented from knowing that my coffee mug is red on the grounds that my term "red" could easily have referred to a different property.[26] One reason is that these are *known* shifts. Since we know that whenever a community-wide change in use occurs, the reference of "red" changes as well, worlds where "red" refers to some property besides redness are also worlds where we are in a position to update our belief-forming practices to reflect the change in subject matter. For instance, in worlds where we apply "red" to some slightly orangey shades, we refer to a different property—call it red*. But we will also reliably believe, of those orangey shades, that they have the property red*, since we are aware that reference shifts with this change in use. We will not systematically have false beliefs where semantically plastic terms are involved.[27]

25 I am borrowing this terminology from Dorr and Hawthorne, "Semantic Plasticity and Speech Reports."

26 See Clarke-Doane, *Morality and Mathematics*, ch. 6, for a claim along similar lines that rejecting stability in favor of a plenitudinous ontology of set-like entities solves epistemological problems for set-beliefs. Clarke-Doane explicitly considers the prospects for an analogous plenitudinous ontology of obligation-like properties. But these advantages are available only to someone who explicitly rejects stability for "ought," a consequence I am assuming the realist wants to avoid.

27 Another reason is that competent users of color terms will allow for a wide range of borderline cases. For shades on the borderline between red and orange, a competent user of "red" will refuse to apply "red" to these shades and will refuse to apply "not red" to them. Likewise, she will not believe of these borderline shades that they are red. If "red" in her community shifts reference to a slightly different property, even if she is not aware of the shift, she will not have a false color belief. Rather, even though her term refers to the color redness* rather than redness, redness* will include some borderline cases of redness about which she withholds belief—she will not falsely believe that these shades are red*. It would take an extremely large semantic shift for "red" to refer to a property that includes shades that, in the actual world, she believes are not red.

Our color beliefs are, of course, not infallible. We can be mistaken about what the term "red" in our own language refers to. But there is no systematic risk here, as the errors will be

Semantic shifts for stable normative terms appear to be much more threatening epistemologically. When we form beliefs about what we ought to do, having a good belief-forming disposition requires being resistant to changing one's beliefs whenever one's community uses normative terminology slightly differently. One should not be disposed to change one's beliefs about what one ought to do just because one's community uses "ought" slightly differently. In general, being disposed to change normative beliefs to match the way one's community uses "ought" is a way to have *false* beliefs about obligation, since our obligations do not for the most part depend on our linguistic habits.

While these semantic shifts do occur, it is unlikely that one will know the precise point at which this happens. The difference between stable and plastic terms does not concern whether we can know how our community uses them. Rather, the difference lies in whether we will be able to tell that the reference of these terms has shifted for a community that uses their terms differently. With a semantically stable term like "ought," we cannot simply assume that any unanimous shift in use in our community is accompanied by a shift in reference. As a result the disposition to (correctly) treat "ought" as semantically stable will, in some worlds where the term has shifted reference, yield a false positive: one will, in virtue of having the disposition, continue to treat "ought" as referring to obligation.

To sum up, premise 2 in the Argument from Risk is true, in the following sense: there are nearby worlds where "ought" refers to something distinct from obligation because these worlds are not easily distinguishable from worlds where it refers to obligation. These are worlds where one will continue to use "ought" as if it refers to obligation, by speaking as if there are no meaningful differences between the usage of "ought" in one's community and the usage of the term in a community that refers to obligation. The differences between such worlds are small but significant: *ex hypothesi*, the usage of "ought" in one's own community, when a semantic shift has occurred, makes it the case that "ought" does not refer to obligation.

4. Premise 3: From Metasemantic Risk to Risk of a False Belief

Premise 3 in the Argument from Risk says:

3. If one could easily be in a world where "ought" does not refer to obligation, then one could easily have had a false normative belief.

This is a conditional claim with an antecedent that concerns normative *language*,

attributable to speakers having impoverished or misleading information, or failing to treat their color terms as sufficiently plastic. I return to the significance of this in sections 4 and 5.

and a consequent that is about the propositional attitude of *belief* about normative matters. According to this premise, shifts in the reference of a piece of language have consequences for the content of our beliefs. The connection is not obvious, so in this section I will outline a background picture of the connection between language and belief that makes premise 3 plausible.

The previous section outlines how metasemantic risk gives rise to worlds with false normative beliefs because, given the moderate semantic stability of normative vocabulary, we could have been in a world where "ought" refers to something besides obligation, but we do not know that this shift has occurred. As a schematic example, take a world where the term "ought" shifts reference by referring not to obligation, but rather to obligation*.[28] Premise 3 claims that the content of normative beliefs of language users in a community that uses "ought" to refer to obligation* will change. Believers will usually rely on the referents of terms in their public language in order to form beliefs about the world around them. And so in worlds where the term "ought" shifts reference to refer to obligation*, speakers in these worlds will be forming normative beliefs about obligation* as well.

A shift in the referent of "ought" in a public language does not entail that one has the same linguistic dispositions as all other members of the community one is a part of. In a world where usage of "ought" by a community has shifted enough to make the term refer to obligation*, members of that community will form normative beliefs about obligation* by relying on the referent of "ought" in their language. For instance, suppose Suzy forms a belief about the normative status of giving 25 percent of her income to charity. This is the belief that has the content she would express by uttering the sentence "it is not the case that one ought to give 25 percent of one's income to charity." If Suzy is in a world where the referent of "ought" has shifted, the normative belief she forms in that world does not have the content *giving 25 percent of one's income to charity is not obligatory*. Instead she refers to—and forms beliefs about—something else, namely obligation*. Shifts in the usage of normative language by Suzy's community can affect shifts in what Suzy's normative thought is *about*, without Suzy changing how she forms her own beliefs. The shift can be a result of changes in her surrounding linguistic community only.

Moreover, these shifts can affect whether Suzy's normative beliefs are *true*. In a world where "ought" has not shifted reference, and refers to obligation, Suzy's

28 Earlier I suggested that if the normative "ought" refers to the property of producing the most good out of the options an agent has a reasonable ability to perform, the referent of "ought" could shift in some worlds to refer to the property of producing the most good out of all possible actions, whether performable or not. The schematic example can be filled in by identifying this property with obligation*.

normative belief has the content of her sentence "it is not the case that one ought to give 25 percent of one's income to charity." She forms a true belief. The belief she forms in the world where her community uses "ought" slightly differently, and thereby refers to obligation*, need not be true. In such a world, Suzy need not be aware that in this world she refers to something different from what her counterpart in an obligation-referring world refers to, and she will continue to believe the content of the sentence "it is not the case that one ought to give 25 percent of one's income to charity." Thus she might continue to form beliefs in the same way. But in the shifted world, "ought" refers to obligation*. Giving 25 percent of one's income to charity *is* obligatory*. Suzy then has a false belief, and the only difference between this world where she has the false belief and a world where she has a true belief is that the usage of normative vocabulary by others in her linguistic community has changed.

As I have emphasized, reference shifts are possible for nonnormative vocabulary as well, but there are some distinguishing features that make the possibility of false beliefs owing to these nonnormative reference shifts less threatening. Return to the example of color terms, like "red," as a contrast case. These terms are much more widespread than shifts for "ought," as almost any community-wide shift in usage is guaranteed to produce a different referent. If we had applied "red" to a slightly different range of light wavelengths (and made corresponding adjustments in our use of other color terms), then we would have referred to something other than redness. This is what makes color terms semantically plastic.

The consequences of plasticity for color terms give rise to a difference in status of color *beliefs* in worlds where reference has shifted. In particular, the increased chance of a reference shift in color terms has the opposite effect: speakers will be less likely to form false beliefs when there are reference-shifting changes in their community's usage. The semantic shifts in plastic terms are *known* shifts, and so we know (roughly) that whenever a community-wide change in use occurs, the reference of "red" changes as well. Moreover, since one should not be disposed to change one's beliefs about what one ought to do in the possible situation where one's community uses "ought" slightly differently, someone with the right dispositions will be further susceptible to false normative beliefs when shifts in the reference of "ought" do occur. These points suggest that when semantic shifts for "ought" do occur, we will be at risk not only of using "ought" accordingly but also of having false normative beliefs.[29]

29 Of course in other nonnormative areas, the relevant vocabulary might not be as plastic as color terms, and instead will display some degree of stability. Steadfast dispositions might be appropriate to some degree as well. These are questions that will need to be answered in extending the Argument from Risk to nonnormative areas.

5. PREMISE 4: FALSE BELIEFS AND EPISTEMIC RISK

The final step in the Argument from Risk claims not only that moderate stability gives rise to possible false normative beliefs but also that the presence of these false beliefs is incompatible with *knowledge*. The risk of false belief is an *epistemic* risk, which prevents even those normative beliefs that manage to be true from being knowledge. This is captured by premise 4:

4. If one could easily have had a false normative belief, then one's actual normative beliefs are at epistemic risk and are not knowledge.

The relevant notion of epistemic risk is captured broadly by a "safety" condition on knowledge, as developed in Sosa, Williamson, and Pritchard: beliefs are subject to epistemic risk when they fail to be safe.[30] My aim here is not to defend the safety condition. Instead, I aim simply to show that plausible refinements on such a condition, which are motivated by plausible intuitions about knowledge, are compatible with the claim that metasemantic risk gives rise to epistemic risk. For those who do not wish to think in terms of a connection between knowledge and safety, the motivating examples are still relevant, as they will provide constraints on alternative views of what is required for knowledge.

Below I will sketch two refinements that any plausible version of a safety condition will need to incorporate. These are a restriction of nearby false beliefs—the kind that make actual beliefs subject to distinctively epistemic risk—to beliefs that are (i) similar in content and (ii) the products of similar causal processes. Normative beliefs are plausibly subject to epistemic risks owing to semantic shifts, even when these refinements are in place. It is not obvious that the same goes for nonnormative beliefs.

The first qualification on the notion of epistemic risk is that a belief is at risk only if *similar* beliefs are false in nearby worlds.[31] I can know that I had breakfast

30 Sosa, "How to Defeat Opposition to Moore"; Williamson, *Knowledge and Its Limits*; Pritchard, *Epistemic Luck*.

31 What makes a world *nearby* in the sense relevant to epistemic risk? Williamson, who accepts a general connection between knowledge and the absence of risk, says that the concept of knowledge cannot be *defined* in terms of the absence of false belief in nearby (or "sufficiently similar") worlds:

> If one believes *p* truly in a case *a*, one must avoid false belief in other cases sufficiently similar to *a* in order to count as reliable enough to know *p* in *a*. The vagueness in "sufficiently similar" matches the vagueness in "reliable" and "know." . . . We need not assume that we can specify the relevant degree and kind of similarity without using the concept *knows*. (*Knowledge and Its Limits*, 100)

While I am not pursuing the question of whether knowledge can be analyzed in terms of

this morning even if there is a nearby world where I misremember the name of a new acquaintance and have a false belief about her name.

While misremembering an acquaintance's name does not put my beliefs about my breakfast at risk, this is not simply because the beliefs are different. Knowledge-destroying false beliefs do not need to be identical in content. If one is guessing at the answer to questions about the sums of moderately large numbers, then one's correct guesses will not be such that there are nearby worlds where the same belief is false. If one correctly guesses that 634 + 399 = 1,033, then one has a true belief, and moreover this very belief is not false in any nearby world (in all nearby worlds, 634 + 399 = 1,033). So one cannot have a false belief that 634 + 399 = 1,033 in any nearby world. Correctly guessing does not, however, bring knowledge. If one is guessing at the relevant sums, then even if one actually gets the answer right, there is a nearby world where one instead comes to believe a related but false claim—for instance, that 634 + 399 = 893. This belief is quite similar to one's actual belief. Since there are nearby worlds where one has false beliefs like this when one is guessing, one's actual true belief that 634 + 399 = 1,033 is subject to risk.

Return to the case of normative beliefs that are subject to metasemantic risk. We can call a world in which one has false normative beliefs owing to semantic shifts in the normative "ought" a *shifty world*. The content of a normative belief in a shifty world is not identical to the content of our actual normative beliefs. Normative beliefs in the actual world are about obligation, but beliefs in the shifty worlds are about a different property.

The false beliefs in shifty worlds will, however, be similar enough to put our actual normative beliefs at risk. If our normative beliefs in the actual world are about obligation, shifty worlds where we instead form normative beliefs about the distinct property obligation* are still worlds where our normative beliefs are very similar: both beliefs involve reference to properties that broadly bear on what to do, or what is best. Moreover, these are formed with the use of concepts that bear on motivation to act, and so are plausibly still normative beliefs rather than beliefs about a radically different subject matter. So if beliefs about obligation* in the shifty world are false, they will be beliefs that are not disqualified

epistemic risk here, it is worth emphasizing that Williamson's line is one that is consistent with the approach to the central questions of this paper. That is, in asking whether normative beliefs can be knowledge for the realist, I will make claims about certain normative beliefs being false in nearby worlds. Williamson may be right that these judgments rest (partly) on our judgments about whether those that hold the relevant beliefs *know*. It is not my aim here to settle this question; rather the point is to deploy the framework of epistemic risk, which has been developed elsewhere, to draw attention to some connections between the metasemantic aims of the realist and normative knowledge.

from being beliefs that prevent our actual normative beliefs from being knowledge simply in virtue of being beliefs with distinct contents.

Here is a second qualification: not all similar false beliefs in nearby worlds are incompatible with knowledge. Some nearby similar false beliefs are arrived at in a suitably different way, and so do not put one's actual beliefs at risk in the relevant sense. If I happen to see a friend who usually lives in another city walk by, I know they are in town (we can suppose this is true even if I have no other evidence that they are visiting). This belief is not at risk just because there is a nearby world where our paths never cross and I continue to believe that my friend is not in town.[32] The reason is that the beliefs are formed by very different processes. The causal processes leading up to the formation of beliefs about my friend's location are very dissimilar, as one involves perception and the other involves an inference on the basis of my knowledge of my friend's usual place of residence. It is thus very natural to conclude that if a belief is subject to epistemic risk, there must be nearby false beliefs that are both similar in content *and* similar in respect to the causal processes that produce them.

Recall the difference between true normative beliefs and their counterparts in shifty worlds that are false owing to metasemantic risk: the only difference between them need be that the latter are formed in a world where the surrounding community of language users deploys normative vocabulary differently. Community-wide changes like this need not be accompanied by a change in one's own belief-forming methods; one can continue to use "ought" and form normative beliefs as if this term referred in one's language to obligation.

Moreover, there is no relevant difference in the aptness of the process one uses in the world where "ought" has shifted reference. Because steadfast dispositions are appropriate for normative terms and beliefs, there is no obvious sense in which the false normative belief is formed by a defective process. Instead, the process by which one forms normative beliefs in each world involves dispositions one should have in each world, since one should not change one's normative beliefs simply in response to changes in one's community's use of normative language.[33] Thus, having a false belief owing to a semantic shift is compatible

32 Cf. Pritchard, *Epistemic Luck*.

33 An anonymous referee raises the possibility that metasemantic risk raises analogous skeptical worries for beliefs about natural kinds, such as arthritis. This may be—I am not taking a stand on the extent of the epistemic consequences of metasemantic risk here—but the distinctive relevance of steadfast dispositions to normative belief formation should not be overlooked. In a world where speakers use "arthritis" to refer to a different natural kind, arthritis*, one might form false beliefs about arthritis* by using the relevant concepts as if they refer to arthritis. If one does this while knowing roughly how one's community (including the relevant experts) uses "arthritis," one will be making a mistake by taking the change in

with nearly everything in the process that produces the false belief being identical to the belief-forming process in a world where the shift has not happened.

Take for example Suzy's false belief about the nonobligatoriness of giving 25 percent of one's income to charity. The token process that produces this belief—which includes the practical reasoning Suzy employs to arrive at her belief as well as other psychological factors that can influence normative belief formation—can be nearly identical in a shifty world, where "ought" refers to obligation*. The difference in content, and the difference in the truth conditions for these beliefs, depends only on factors that lie outside the causal chains that produce these beliefs. We need only to change how the surrounding community uses their normative vocabulary in order to describe a world where Suzy's normative beliefs refer to obligation* and are therefore false.

These points together suggest that we can mount a strong case that possible normative beliefs that are false owing to metasemantic risk entail that our actual normative beliefs are subject to epistemic risk.

6. Limits to the Argument from Risk

So far I have assembled the credentials for metasemantic risk as the key notion in an epistemological argument against normative realism. Since there are independent constraints that entail that normative terms should, on the realist view, be moderately stable, metasemantic risk exists on the realist view. This yields possible false beliefs that have certain distinguishing features of epistemic risks for our actual normative beliefs. These false beliefs are the products of changes in community usage that are not distant possibilities; they are similar in content to our actual normative beliefs, and they are the products of very similar causal processes. These considerations ward off any general strategies for dismissing the Argument from Risk.

The Argument from Risk does not, however, establish sweeping skepticism for the normative realist. Rather, its ambitions will have to be scaled back: while some normative beliefs do fail to be knowledge owing to metasemantic risk, the skeptical consequences of realism do not extend to every normative belief. Whether this restricted skepticism is problematic for realism, and vindicates traditional epistemological worries about the view, is a question I will revisit in the conclusion.

---

usage to be irrelevant to the reference of "arthritis." There is a significant difference between the nondefective method one might use to form beliefs about arthritis in a normal world and the defective beliefs one would use to form beliefs about arthritis* in a world where "arthritis" has shifted reference. But there is no analogous difference in the methods one uses in forming normative beliefs in worlds where "ought" shifts reference.

6.1. Contingent Epistemic Risk

Premise 2 in the Argument from Risk assumes that the core semantic features of normative language for the realist—moderate semantic stability and modal continuity—imply that "ought" could easily have shifted referents.

A shifty world is a world where our normative beliefs are false owing to a semantic shift in the referent of "ought." One could *easily* have been in a shifty world, in the following sense: differences between our world and a shifty world are potentially very small and are not necessarily changes that we can tell will constitute a shift in the reference of "ought." Moderate semantic stability guarantees the existence of the shifty world; our dispositions to treat "ought" as stable make us likely to treat "ought" as referring to obligation even in some shifty worlds. In one sense the shifty world is one we could easily have been in, since in a shifty world one might not know that one is in a shifty world.

If we are in a nonshifty world, we can say that the world we are in is *stable*. Premise 4 claims that shifty worlds are close enough to stable worlds, and so the normative beliefs we hold in stable worlds are subject to epistemic risk. But it is not sufficient for distinctively *epistemic* risk that one not be able to tell the difference between the stable and shifty worlds. The kind of epistemic risk that is incompatible with knowledge requires that the causal process that produces a belief must be similar to one that gives rise to false beliefs in worlds that are metaphysically close to the actual world. Guessing gives rise to epistemic risk, since one's guess could easily have produced a different belief. Even though hallucination is possible, relying on visual impressions to form beliefs does not give risk to epistemic risk: a world where one hallucinates need not be metaphysically close to the actual world.

Nothing in premise 2 of the Argument from Risk guarantees that all stable worlds are close to shifty worlds in the sense that is relevant to epistemic risk. It is consistent with our world being one where "ought" refers to obligation that all of the nearby worlds where usage of normative terms differs only slightly from ours are also stable. All premise 2 tells us is that if we were in a (possibly distant) shifty world, we would not necessarily know that we are in a shifty world. In this case, the risk of false belief brought about by shifty worlds is not threatening to knowledge, since the shifty worlds are modally distant and cannot generate the kind of epistemic risk that prevents actual normative beliefs from being knowledge.[34]

There are, however, some stable worlds that are metaphysically close to

34 Williamson emphasizes a general version of the point that simply because we would not know that we are in the shifty world if we were in one, it does not follow that the world is a close one that threatens our knowledge in the actual world (*Knowledge and Its Limits*, sec. 8.2). If we were to assume that our inability to discriminate the shifty worlds where we have

shifty worlds. In such worlds, the motivations for premise 2 will be relevant to our normative knowledge in these stable worlds. Premise 2 could be filled out appropriately so it is true in these stable worlds: that is, worlds that are both difficult for us to distinguish from shifty worlds and also modally close to shifty worlds. With appropriate modifications to the other premises in the Argument from Risk, we will have a version of the argument that is *contingently* sound. Its premises will not be true in all worlds but will be true in some.

Nonetheless this reveals a distinctive epistemic commitment in virtue of the core features of the realist view. Owing to metasemantic risk, we will need to admit that it is a contingent possibility that we lack normative knowledge. If we are in a world where a shifty world is nearby, then we will lack normative knowledge. The presence of knowledge-destroying metasemantic risk is an epistemic limitation distinctive to realism but not a necessary feature of the view.

6.2. Higher-Order Knowledge

There is a related skeptical conclusion that the realist must accept for every stable world, including the actual one. Call the worlds that could be reached by just small changes in use from our actual world the worlds that are in the *immediate vicinity* of ours. If there is a shifty world in the immediate vicinity of ours, we lack *first-order* normative knowledge. But as the previous subsection argues, there is nothing in the commitments of the realist that entails that it is necessary that there should be a shifty world in the immediate vicinity of a world where normative beliefs are formed. What is necessary is merely that there is a shifty world somewhere in modal space.

The existence of a shifty world somewhere in modal space does not by itself threaten our first-order normative knowledge. But it does threaten our higher-order normative beliefs. Suppose the shifty world is not in the immediate vicinity of the actual world but is in the immediate vicinity of a world in our immediate vicinity. We can then know that we ought to give 10 percent of our income to charity (this is first-order knowledge). We do not, however, *know* that we know that we ought to give 10 percent of our income to charity. The higher-order belief is at risk, epistemically. There are worlds in our immediate vicinity where it is false—these are the worlds that have the shifty world in their immediate vicinity. (In a world with a shifty world in its immediate vicinity, we do not know that we ought to give 10 percent of our income to charity. And in worlds with a world where we do not know that we ought to give 10 percent of our income to charity in their immediate vicinity, we do not know that we know that we ought to give

false beliefs guarantees their closeness, skepticism would follow for reasons having nothing to do with moderate stability.

10 percent of our income to charity.) So, even if we are not at immediate risk of forming false normative beliefs, the realist will have to concede that some of our higher-order normative beliefs will fail to be knowledge.[35]

Whether the loss of merely higher-order normative knowledge is damaging to the realist is an open question. It does, however, point to one respect in which a realist view that entails the moderate semantic stability of normative terms will have a distinctive (and limited) epistemic profile.

6.3. Imprecise Knowledge

So far we have seen that the Argument from Risk is limited in what it establishes, in certain respects. It is only contingently sound at best when it targets our first-order normative beliefs. All we can be certain of is that, owing to metasemantic risk, we will lack some higher-order normative knowledge.

In addition, the Argument from Risk is not equally threatening to all of our normative beliefs. Even if we are (as a contingent matter) in a world that is very close to a shifty world, there are kinds of normative belief that will have a claim to survive as normative knowledge. Begin with the beliefs that are most threatened by the Argument from Risk. The example I have been working with is the belief that it is obligatory to give 10 percent of one's income to charity. This is just for illustrative purposes, and there are many beliefs like it that will be similarly threatened: for instance, the belief that one ought to avoid eating any kind of animal regardless of the effects for future animal suffering, or the belief that lying to avoid personal embarrassment is impermissible. Even if these beliefs are true, if a certain type of shifty world lurks nearby, they will not be knowledge.

But what they all have in common is that they are extremely *specific* beliefs. A specific belief is one that is incompatible with beliefs that make only slightly different cutoff points for obligatory action. For example, believing that one ought to give 10 percent of one's income to charity is incompatible with the belief that one is only obligated to give exactly 9 percent of one's income to charity. (If the latter belief is true, one is not obligated to give 10 percent, though it would perhaps be commendable to do so.) Likewise, we could have believed that we are only obligated to avoid eating mammals, or that we are only obligated to refrain from lying when the consequences are more serious than mere personal embarrassment.

Specific beliefs like these go false in shifty worlds. There are ways for "ought" to shift its reference while still being used as a normative term that make the

35 We might have to go even further away from the actual world to find worlds where we do not know that we ought to give 10 percent of our income to charity, owing to the proximity of shifty worlds. Regardless, we will not know that we know that we . . . that we ought to give 10 percent of our income to charity for the same reasons; some iteration of knowledge will fail.

belief with the content (in the shifty world) of "one ought to give 10 percent of one's income to charity" false in that world. If giving 10 percent of one's income to charity is obligatory but not obligatory*, then acts of giving 10 percent do not have the property "ought" refers to in the shifty world. The same problem does not necessarily affect nonspecific beliefs. Take for instance the belief that one ought to give between 1 and 30 percent of one's income to charity (where this belief is equivalent to the disjunctive belief that one ought either to give exactly 1 percent, or one ought to give exactly 2 percent, and so on). This belief has a *margin for error* built in—unlike the specific belief that one ought to give 10 percent of one's income, it is compatible with a variety of specific facts about one's obligations. Some beliefs with sufficiently wide margins for error will not have false counterparts in nearby shifty worlds.

Even if one's belief that one ought to give 10 percent of one's income to charity has a false counterpart in a nearby shifty world, it does not follow that the belief that one ought to give between 1 and 30 percent of one's income to charity does as well. This is because of the built-in margin for error: even though what one believes when one forms a normative belief in the shifty world is different—one does not believe that one *ought* to give between 1 and 30 percent of one's income to charity, since one's term "ought" does not refer to obligation—one forms a true belief anyway. If a normative "ought" shifts reference to refer to obligation*, one still forms a true belief, since there is some percentage of one's income between 1 and 30 percent that is such that donating it is also obligatory*.

This points to a further respect in which the Argument from Risk is limited. Premise 3 does not distinguish between which normative beliefs go false in a shifty world. But it should, since some normative beliefs with sufficiently wide margins do not go false even when "ought" shifts reference. Even in a setting where a specific normative belief is at risk, and thereby cannot be knowledge, a belief with margins for error built in might, in the very same setting, be knowledge. So a suitable modified Argument from Risk will not threaten all normative knowledge equally.[36]

7. CONCLUDING REMARKS

There are, then, some limits to the amount of skepticism that the Argument from Risk entails. It does not entail a wholesale rejection of knowledge about the target domain, although it does point to distinctive ways in which our knowledge may be limited owing to a realist metaphysics that implies that normative terms

36 Thanks to Levi Spectre for discussion of this point.

are moderately semantically stable. Where do the limits leave realism's epistemological credentials?

From one perspective, it saves realism from the damning conclusion that realism entails wholesale skepticism. The skeptical conclusion is obviously unpalatable, since if it followed from the core commitments of realism that all normative beliefs are not knowledge, we would have strong reason to look elsewhere for a normative metaphysics that makes room for normative knowledge.

However, from another perspective, the Argument from Risk does not leave realism on a par with its competitors when epistemological considerations are in play. It is worth emphasizing the respects in which settling for knowledge only for claims that have sufficiently wide margins for error is not entirely satisfactory. At first glance it seems to be a small concession to hold that while we might not know specific normative propositions—such as the claim that we ought to give 10 percent of our income to charity—we can know related propositions with margins for error built in. But in fact our normative knowledge is connected in a variety of important ways to other aspects of our cognitive and practical lives, and our beliefs with wide margins for error cannot sustain all of these connections. I will close with two examples.

One example of the way this skepticism ramifies comes from the consequences for knowledge of other normative claims that are believed in typical ways. It is plausible that when we form beliefs by deducing them from other beliefs, the output of the deduction is known only if the premises from which it was deduced were known. (This principle is sometimes known as *counter-closure*.) Typical agents who hold the wide-margin-for-error belief that they ought to give between 1 and 30 percent of their income to charity often do so on the basis of a deduction from a specific normative belief—say, the belief that one ought to give 10 percent of one's income to charity. But if the specific beliefs are not knowledge, then by counter-closure the wide-margin-for-error beliefs that are deduced from them will not be knowledge either. So by conceding that some normative claims are not known owing to metasemantic risk, we might be forced to hold that other in-principle-knowable claims are not in fact known.

The other example is from the practical side: it is common to hold that an agent's reasons, of a certain kind, can only be claims that she knows or has some epistemic access to. For instance, if Bob does not know that the police are arriving at the crime scene, it is infelicitous to say that Bob's reason for fleeing the crime scene was that the police were arriving.[37] If Suzy gives 10 percent of her income to charity out of her sense of moral duty, it is tempting to say that her rea-

37 Hyman, "How Knowledge Works." This is sometimes called an ascription of a *personal* reason for action.

son for giving the money is that she is obligated to give 10 percent of her income to charity. But this cannot be her personal reason for acting, if she does not know specific normative propositions like this. And Suzy might not be thinking about a proposition with margins for error built in, much less take such a claim to be her reason—most people do not give 10 percent of their income merely on the basis of the fact that they are obligated to give between 1 and 30 percent of their income to charity. So even a lack of knowledge of some normative propositions can deprive agents of some important personal reasons for action.

These considerations are not decisive, but they point to some respects in which even a limited skepticism about normativity might be troubling for realism. How troubling the limited skepticism is remains to be seen, but I have sketched two reasons in closing for holding that the skepticism cannot be dismissed as wholly irrelevant simply because it is not wholesale skepticism. Proponents of realism can of course avoid these results by denying moderate semantic stability for normative terms. There are, however, strong motivations elsewhere in the literature for realists to accept moderate semantic stability. What (some version of) the Argument from Risk shows is that there may be no version of realism that is entirely problem free. Endorsing moderate semantic stability, even if a satisfactory metaphysics and metasemantics can be found to support it, faces its own problems in light of its epistemological consequences.[38]

University of Missouri–St Louis
dunawayw@umsl.edu

REFERENCES

Boyd, Richard N. "How to Be a Moral Realist." In *Essays on Moral Realism*, edited by Geoffrey Sayre-McCord, 181–228. Ithaca, NY: Cornell University Press, 1988.

Clarke-Doane, Justin. *Morality and Mathematics*. Oxford: Oxford University Press, 2020.

Dorr, Cian, and John Hawthorne. "Semantic Plasticity and Speech Reports." *Philosophical Review* 123, no. 3 (July 2014): 281–338.

38 Thanks to audiences at the Stockholm June Workshop in Philosophy; Kansas State University; the Midwest Epistemology Workshop in Madison, Wisconsin; and the Saint Louis Ethics Workshop, and two anonymous referees for very helpful comments on earlier versions of this paper.

Dummett, Michael. "Realism." In *Truth and Other Enigmas*, 145–65. London: Duckworth, 1978.

Dunaway, Billy. "Expressivism and Normative Metaphysics." In *Oxford Studies in Metaethics*, vol. 11, edited by Russ Shafer-Landau, 241–64. Oxford: Oxford University Press, 2016.

———. *Reality and Morality*. Oxford: Oxford University Press, 2020.

Dunaway, Billy, and Tristam McPherson. "Reference Magnetism as a Solution to the Moral Twin Earth Problem." *Ergo* 3, no. 25 (2016): 639–79.

Edwards, Douglas. "The Eligibility of Ethical Naturalism." *Pacific Philosophical Quarterly* 94, no. 1 (March 2013): 1–18.

Eklund, Matti. *Choosing Normative Concepts*. Oxford: Oxford University Press, 2017.

Enoch, David. *Taking Morality Seriously: A Defense of Robust Realism*. Oxford: Oxford University Press, 2011.

Gibbard, Allan. *Meaning and Normativity*. Oxford: Oxford University Press, 2013.

———. *Thinking How to Live*. Cambridge, MA: Harvard University Press, 2003.

Harman, Gilbert. "Moral Explanations of Natural Facts—Can Moral Claims Be Tested against Moral Reality?" *Southern Journal of Philosophy* 24, no. S1 (Spring 1986): 57–68.

Hawthorne, John. "*A Priority* and Externalism." In *Internalism and Externalism in Semantics and Epistemology*, edited by Sanford C. Goldberg, 201–18. Oxford: Oxford University Press, 2007.

Horgan, Terry, and Mark Timmons. "Copping Out on Moral Twin Earth." *Synthese* 124, no. 1–2 (July 2000): 139–52.

———. "Troubles for New Wave Moral Semantics: The 'Open Question Argument' Revived." *Philosophical Papers* 21, no. 3 (1992): 153–75.

———. "Troubles on Moral Twin Earth: Moral Queerness Revived." *Synthese* 92, no. 2 (August 1992): 221–60.

Hyman, John. "How Knowledge Works." *Philosophical Quarterly* 49, no. 197 (October 1999): 433–51.

Lewis, David. "New Work for a Theory of Universals." *Australasian Journal of Philosophy* 61, no. 4 (December 1983): 343–77.

———. "Putnam's Paradox." *Australasian Journal of Philosophy* 62, no. 3 (September 1984): 221–36.

Mackie, J. L. *Ethics: Inventing Right and Wrong*. London: Pelican Books, 1977.

Pritchard, Duncan. *Epistemic Luck*. Oxford: Oxford University Press, 2004.

Railton, Peter. "Moral Realism." *Philosophical Review* 95, no. 2 (April 1986): 163–207.

Smith, Michael. *The Moral Problem*. Oxford: Wiley-Blackwell, 1994.

Sosa, Ernest. "How to Defeat Opposition to Moore." *Philosophical Perspectives* 13, Epistemology (1999): 141–53.

Street, Sharon. "A Darwinian Dilemma for Realist Theories of Value." *Philosophical Studies* 127, no. 1 (January 2006): 109–66.

Van Roojen, Mark. "Knowing Enough to Disagree: A New Response to the Moral Twin Earth Argument." In *Oxford Studies in Metaethics*, vol. 1, edited by Russ Shafer-Landau, 161–94. Oxford: Oxford University Press, 2006.

Wedgwood, Ralph. "Conceptual Role Semantics for Moral Terms." *Philosophical Review* 110, no. 1 (January 2001): 1–30.

———. *The Nature of Normativity*. Oxford: Oxford University Press, 2007.

Williams, J. Robert G. *The Metaphysics of Representation*. Oxford: Oxford University Press, 2020.

———. "Normative Reference Magnets." *Philosophical Review* 127, no. 1 (January 2018): 41–71.

Williamson, Timothy. *Knowledge and Its Limits*. Oxford: Oxford University Press, 2000.

Wright, Crispin. *Truth and Objectivity*. Cambridge, MA: Harvard University Press, 1992.

Journal of Ethics and Social Philosophy
VOL. 20, NO. 3 · NOVEMBER 2021
https://doi.org/10.26556/jesp.v20i3.1249

MORAL FETISHISM AND A THIRD DESIRE FOR WHAT'S RIGHT

Nathan Robert Howard

PHILOSOPHERS who agree that the morally best kind of motivation requires a desire for what's right nevertheless disagree about how to understand that requirement.[1] This disagreement originates in an ambiguity in the phrase "a desire for what's right," leading philosophers to defend different readings of the requirement. This debate seems to presuppose that the phrase has only two readings. However, outside of this debate, it is widely recognized that scope-ambiguous phrases like "a desire for what's right" have *three* distinct readings. The debate has wholly ignored the third reading. This paper describes it and argues that it claims some of the appealing features of the other two readings.

1. IDENTIFYING THE THIRD READING

In the context of the debate about whether moral judgments intrinsically motivate, Michael Smith reminds us that two very different kinds of desires can be described as desires to do what's right.[2] The ambiguity in this description is a general feature of expressions that combine quantifiers, including, following Hintikka, attitude verbs like "desire," and phrasal complements. For example, "Jules wants to eat a Big Kahuna burger" is true of two distinct scenarios. In the first, Jules sees a specific burger and comes to want it, unaware that the burger is

1 Because the question of which motives are morally good is ancient, I focus on the discussion succeeding and informed by Smith, *The Moral Problem*. Earlier discussions from Copp ("Moral Obligation and Moral Motivation"), Lillehammer ("Smith on Moral Fetishism"), and Dreier ("Dispositions and Fetishes") focus on this question as it relates to moral judgment externalism. More recent discussions focusing on the nature of good moral motivation itself include Olson, "Are Desires De Dicto Fetishistic?"; Carbonnell, "De Dicto Desires and Morality as Fetish"; Sliwa, "Moral Worth and Moral Knowledge"; Aboodi, "One Thought Too Few"; and Isserow, "Moral Worth and Doing the Right Thing by Accident" and "Moral Worth."

2 Smith, *The Moral Problem*.

a Big Kahuna burger. This scenario corresponds to the expression's *de re* reading. On the second, Jules has heard of Big Kahuna burgers and thinks of or *cognizes* the burgers he wants as Big Kahuna burgers. This scenario corresponds to the expression's *de dicto* reading.

The phrase "a desire to do what's right" also has *de re* and *de dicto* readings. The former implies a desire to do a particular action, which happens to be right. The latter implies a desire to do what the agent takes to be right, as such. Correspondingly, philosophers who agree that morally good motives require a desire to do what's right can still disagree about whether that claim is to be understood *de dicto* or *de re*. However, arguments in this debate often seem to presuppose that "a desire to do what's right" has *only* these two readings. This presupposition is false.

As Janet Fodor first showed, expressions like "Jules wants to eat a Big Kahuna burger" have a *third* reading, which she calls the nonspecific reading, contrasting this with the "specific *de re*" and "nonspecific *de dicto*" readings characterized above.[3] The more familiar specific *de re* reading requires that Jules wants a specific burger. The *de dicto* reading requires that Jules cognizes his desired burgers as Big Kahuna burgers. However, "Jules wants to eat a Big Kahuna burger" can be true on the third reading even if neither of these conditions is met. For example, suppose that to be a Big Kahuna burger is to be a burger garnished with a pineapple slice and yellow hibiscus sauce—the pineapple slice and hibiscus sauce are essential to and sufficient for being a Big Kahuna burger.[4] Jules tries a Big Kahuna burger for the first time, falling for its unique taste. If Jules craves any burger with pineapple and hibiscus, and, necessarily, only Big Kahuna burgers combine those ingredients, then we can accurately report Jules's craving with "Jules wants to eat a Big Kahuna burger." But suppose that Jules does not want a particular burger nor do his desires deploy the concept *Big Kahuna burger*—that is, Jules does not cognize his desired burgers as Big Kahuna burgers. So the report is not true on either its specific *de re* or *de dicto* reading. But it is true on its *nonspecific reading* for Jules has a *de re* desire for a particular *kind* of burger, of which only certain burgers are members.[5]

3 Fodor, *The Linguistic Description of Opaque Contexts*. For further discussion of the nonspecific reading, see Bäeurle, "Pragmatisch-semantisch Aspekte der NP-Interpretation"; Von Fintel and Heim, "Beyond De Re/De Dicto"; Égré, "Hyperintensionality and De Re Beliefs"; and Keshet and Schwartz, "De Re/ De Dicto."

4 In this sense, "Big Kahuna burger" operates more like "shepherd's pie," denoting the result of a particular combination of ingredients, than it does "Brand *X* burger," a burger by Brand *X*.

5 Although there is widespread consensus about how to represent nonspecific *de dicto* and specific *de re* readings, representing the nonspecific reading is more controversial. Von Fintel and Heim offer the following logical form as the "standard" intentional approach to representing the reading's logical form:

Since "Jules wants to eat a Big Kahuna burger" and "Jules desires to do what's right" share the relevant scopal features, the latter also allows a nonspecific reading.[6] The former is true on this reading when Jules wants a particular kind of burger but he does not want a particular burger and he does not desire that kind of burger under the description "Big Kahuna burger." Likewise, the latter is true on this reading when Jules desires to perform a particular kind of act but he does not want to perform any act in particular and he does not cognize the kind of act that he wants to perform as the morally right kind.

For example, suppose (strictly for the purposes of illustration) that act utilitarianism is true so that an act is morally right just when and because it maximizes the balance of pleasure over pain. Imagine that Jules performs a pleasure-maximizing act and is overcome with a warm tingle. Because of the tingle, Jules wants to do more acts of *that pleasure-maximizing kind*, which happen to be all and only the morally right acts. "Jules desires to do what's right" is true on its nonspecific reading in this situation. Jules desires to perform a particular kind of action—the morally right kind—even though he does not desire it under that description.[7]

As a result, there is cause to recognize a *third* position in the debate about which desires underlie morally good motivation: morally good motivation might require a nonspecific desire to do what's right. This unacknowledged position is significant for two reasons. First, as I have already suggested, debate

Jules wants$_w$ [$\lambda w'$ [a Big-Kahuna-burger$_w$] λx [PRO to eat$_{w'}$ x]]

The formalization expresses, roughly, that in all of the worlds where Jules gets what he wants, he eats any burger that is actually a Big Kahuna burger.

6 That the latter sentence involves an embedded wh question clouds the scopal similarities of the two sentences somewhat. However, Groenendijk and Stokhof ("Semantic Analysis of 'Wh' Complements" and *Studies in the Semantics of Questions and the Pragmatics of Answers*) demonstrate that interrogative phrases like "what's right" produce scopal ambiguities when embedded in opaque contexts. In particular, "Jules desires to do what's right" has the same scopal ambiguities as "Jules desires to do an act that is right." Using von Fintel and Heim's "standard" approach, we can represent the latter claim's logical form as:

Jules wants$_w$ [$\lambda w'$ [a right-act$_w$] λx [PRO to do$_{w'}$ x]]

Roughly, when Jules satisfies the reading corresponding to this form, in all the worlds where he does what he desires, he does any act that is actually a morally right act. See also Sharvit ("Embedded Questions and 'De Dicto' Readings") for a particularly nuanced discussed of the scopal properties of embedded questions. For a related discussion concerning logic, see Woods, "Logical Indefinites."

7 Of course, utilitarianism is probably not the true moral theory. There is likely a plurality of potentially right-making features, which includes being pleasure maximizing along with being fair, being just, and being honest. Consequently, the motivational structure required to satisfy the nonspecific reading will also have to be pluralistic. I have assumed utilitarianism only for the purposes of exposition.

about moral motivation often seems to presuppose that individuals can desire to do what's right in only two ways. Second, the third reading combines some of the most attractive features of the first two readings, providing an attractive new basis for theorizing. I will briefly explore these two features after articulating a controversial assumption on which I will rely throughout.

2. NONSPECIFIC DESIRES AND OPAQUE CONTEXTS

Attitude terms create so-called opaque linguistic environments where the substitution of intensionally equivalent terms fails to preserve truth. For example, "Clark Kent" and "Superman" are intensionally equivalent, denoting the same individual in all possibilities.[8] However, Lois Lane might believe that Clark Kent wears glasses without believing that Superman wears glasses. Substituting "Clark Kent" with an intensionally equivalent term, like "Superman," under the scope of an attitude verb affects the sentence's truth conditions.

According to an orthodox explanation of this phenomenon, whether an attitude report is true depends on whether the sense of a given claim matches the sense in which the believer accepts it. The description or "sense" that Lois associates with the claim that Clark Kent wears glasses differs from the sense that she attaches to the claim that Superman wears glasses.[9] That sense matches the sense expressed by "Clark Kent wears glasses" but not "Superman wears glasses" even though those claims are intensionally equivalent.

Like beliefs, desires are widely assumed to be propositional attitudes. Unsurprisingly then, desire ascriptions, like belief ascriptions, do not seem to support the inter-substitution of intensionally equivalent terms under the scope of "desire." Just as above, Lois may want to kiss Superman even if she does not want to kiss Clark Kent. That is because the sense in which she desires to kiss someone involves a description that she attaches to "Lois kisses Superman" but not to "Lois kisses Clark."

Rather, a desire attribution is true, I will assume, only if the agent desires the relevant object under the description associated with the attribution. On an appealingly simple (but optional) way of thinking about this phenomenon, desires must deploy the concepts expressed by a given description to desire an object under that description.[10] On this approach, even if woodchucks are necessarily groundhogs, Lois can desire to adopt a woodchuck but not a groundhog (each *de*

8 I will assume that a single individual can belong to distinct worlds.

9 I am using "sense" broadly and loosely. I explicitly disavow Fregean commitments inessential to the account, whichever those are.

10 I am assuming this approach simply for concreteness. The account that I propose is consis-

dicto) if her desires deploy the concept *woodchuck* but not *groundhog*. Likewise, even if utilitarianism is true and what's right is necessarily what's pleasure-maximizing, Jules can desire to do what's pleasure-maximizing but not what's right (each *de dicto*), if his desires deploy *is pleasure-maximizing* but not *is right*.

As a special case of this phenomenon, unless Jules desires what's right under the right description, he does not satisfy "Jules desires to do what's right," on its nonspecific reading. What is the right description? I will assume, controversially, that the description picks out the kind through the features distinctive of that kind. For example, a burger is a Big Kahuna burger in virtue of its pineapple slice and yellow hibiscus sauce. Consequently, this assumption implies that Jules satisfies "Jules wants a Big Kahuna burger," on its nonspecific reading, only if Jules desires any burger that is garnished by a pineapple slice and yellow hibiscus sauce.[11] More generally, I will assume that an agent *A* satisfies "wants what's *K*," on its nonspecific reading, only if (a) something belongs to *K* in virtue of being *F* and (b) *A* wants anything that is *F* in virtue of its being *F*, for a suitably restricted domain of *F*s. In short, I am assuming that an agent nonspecifically desires a member of a kind when the features that explain why an object belongs to a kind *match* the features that explain why the agent desires the object.

This assumption implies that Jules nonspecifically desires to do what's right only if (a) an act is right in virtue of being *F* and (b) Jules wants to do anything that is *F*, for a suitably restricted domain of *F*s. It is widely assumed that something is right just when and because its right-making features outweigh its wrong-making features—just when its right-making features are sufficient. Consequently, only if Jules is attracted to the kind of actions whose right-making features are sufficient in virtue of their sufficiency does Jules satisfy "Jules desires to do what's right" on its nonspecific reading, according to the view that I am sketching.

These claims are plausible but highly controversial. They presuppose not only that attitude ascriptions are hyperintensional but also that nonspecific desire attributions are sensitive to the features in virtue of which an object belongs to the desired kind. A deeper examination of nonspecific desire attributions may very well cast doubt on these assumptions. Nevertheless, there is room for them at

tent with other accounts of what it takes to have an attitude under a description that does not rely on concepts.

11 This domain of burgers, over which "any burger" quantifies, is plausibly restricted: even though Jules wants a Big Kahuna burger, in the nonspecific sense, it goes without saying that he does not want a burger that has been dropped on the floor, even if it has a pineapple slice and yellow hibiscus sauce. The domain of burgers that Jules desires is restricted to those that have not been dropped on the floor. Consequently, "anything" in (b) quantifies over only a restricted set of *F*s. For more, see Fara, "Specifying Desires"; Braun, "Desiring, Desires, and Desire Ascriptions"; and Philips-Brown and Grant, "Getting What You Want."

this relatively early stage of inquiry: no sustained discussion of the hyperintensional features of nonspecific attitude ascriptions yet exists. As such, although my claims may be controversial, they are also uncontested. I will therefore assume them given the apparent need for at least a *minimal* theory of sense for a plausible theory of nonspecific desire attribution.

3. Applying the Third Reading

I have just offered a brief sketch of the truth conditions for attributions of nonspecific moral motivation. An agent *A* satisfies "desires to do what's right" on the nonspecific reading, just when *A* desires to do the morally right kind of action in virtue of the features that make those acts morally right. Note that desiring a kind of object in virtue of a certain feature shared by that kind does not require believing that the kind shares that feature. Jules can desire a Big Kahuna burger in virtue of its hibiscus sauce without believing anything about hibiscus sauce. Jules can just think, "Mmhmm! This is a tasty burger!" without reflecting on what makes the burger tasty. Rather, Jules satisfies the attitude ascription because what explains Jules's desire for the burger is the fact that it is topped with pineapple and hibiscus sauce. Likewise, I need not believe that an act is right in virtue of its right-making features being sufficient in order to satisfy "Nathan desires to do what's right" on the nonspecific reading. It suffices that my desire for that kind of act is explained by the fact that those acts possess features that make them right.

As we will now see, nonspecific moral motivation, so construed, offers an attractive middle ground between the two dominant positions in the debate over what constitutes good moral motivation. The first position holds that good moral motivation requires a noninstrumental desire for what's right *de re*.[12] The second position holds that good moral motivation requires a noninstrumental desire for what's right *de dicto*. Each position has clear virtues: the first but not the second entails that good moral motivation requires doing what is in fact right. The second but not the first entails that good moral motivation requires acting out of respect for morality. The rest of the paper sketches a case for the claim that nonspecific moral motivation exemplifies each of these virtues without succumbing to their vices, which I will now discuss.

Smith rejects desiring to do what's right, *de dicto*, on the grounds that it inappropriately "fetishes" rightness itself, which risks alienating the agent from the

12 I am restricting my claims to noninstrumental motivation, desires that are not in the service of promoting further desires. *De re* or *de dicto* motivation that is strictly instrumental on nonmoral concerns is plainly not the morally best kind of motivation, so irrelevant for my purposes.

features that make acts right, such as that they are just or fair.[13] Smith supports this position by rehearsing a point from Bernard Williams.[14] Williams describes a case from Fried where a husband can save either his wife or a stranger from drowning. Williams observes that, with sufficient philosophical ingenuity, we can wring the clearly correct implication that the husband may save the wife rather than the stranger from impartial moral theories like Kantian deontology or utilitarianism. These Kantian and utilitarian justifications, however, involve one thought too many. It should not matter to the husband that the stranger lost a fair lottery or that the husband's maxim of action can be universalized, argues Williams; it matters only that the husband's wife is drowning. On Smith's reading, bizarre moral deliberations like those apparently recommended by Kantianism and utilitarianism seem motivated by a *de dicto* desire for what's right. It would be inappropriate or fetishistic for the husband to have this desire because it seems to displace a desire for his wife's well-being. *De dicto* desires to do what's right are thus fetishistic because they supplant desires for what actually matters more generally, such as obligations to our loved ones, our promises, generalized benevolence, and so on.

This first criticism is complemented by a second. While it may seem bad to displace the husband's desire for his wife's well-being, that is genuinely worrying only if the husband's new desire is in some sense worse than the old one. Williams's case does not entail that a desire for rightness as such is bad; it merely suggests it. Smith regards its badness as common sense. I agree. However, we could reasonably ask for an argument for this claim.

Drawing on observations from Philip Stratton-Lake, Jonathan Dancy, and, ultimately, Philippa Foot concerning the "verdictive" nature of rightness, we might hold that goodness is a measure of value.[15] Something is good when it is of sufficient value, just as something is tall when it is of sufficient height. But just as it is a mistake to think that tallness has a certain quantity of height rather than being a *measure* of a certain quantity of height, it is a mistake to think that goodness has a certain quantity of value rather than being a *measure* of a certain quantity of value. If goodness merely represents a degree of value but does not itself possess value (just as tallness does not have a height), and if it is not good to desire the valueless, then it is not good to desire goodness itself, apart from desiring good things. Similarly, moral rightness can be understood as simply a measure of support by moral reasons. Desiring the property of moral rightness

13 Smith, *The Moral Problem*.

14 Smith, *The Moral Problem*, 75; Williams, "Persons, Character, and Morality."

15 Stratton-Lake, *Kant, Duty, and Moral Worth*; Dancy, *Ethics without Principles* and "Should We Pass the Buck?"

itself, as suggested by the *de dicto* reading, desires the *measure* and not what is measured, which is what genuinely deserves moral concern. This is why many philosophers find the *de dicto* account of moral motivation unappealing.

However, the competing *de re* account is not without its faults either. In particular, a *de re* desire for what's right does not suffice for the morally best kind of motivation. Plausibly, such motivation requires being motivated to perform a morally right action. Acting with a *de re* desire for what's right entails acting rightly. But it does not suffice for the morally best kind of motivation: "morally worthy" motivation. Famously, the shopkeeper in Kant's *Groundwork* wants to treat his customers honestly.[16] Treating his customers honestly is the right thing to do. Consequently, the shopkeeper desires to do what's right, read *de re*. However, it turns out that this shopkeeper is indifferent to morality; he does not act out of respect for the moral law as such. Rather, he is completely selfish. Ordinarily, this selfishness would lead him to defraud his customers. However, the shopkeeper believes that his shop will succeed only if it has a reputation for honesty. Consequently, the shopkeeper's selfishness leads him to desire to do what's right, *de re*. But his motives are clearly morally deficient. As a result, *de re* moral motivation is insufficient for good moral motivation.

While some, most explicitly Markovits, defend the *de re* approach to moral motivation against challenging cases like Kant's shopkeeper through a match between the noninstrumental moral reasons for an action and the agent's noninstrumental motivating reasons for the action, others argue that *de re* moral motivation is incompatible with moral worth.[17] According to Paulina Sliwa and likeminded philosophers such as Zoe Johnson King, *de re* motivation only ever produces accidentally right actions because the right-making features that sometimes underlie *de re* desires are defeasible.[18] For example, just because an act is, for example, *fair* does not entail that it is right—the fair division of disputed land may have catastrophic effects that make the fair division wrong. As such, right-making features, as opposed to rightness itself, are only contingently right-making. Sliwa argues that *de re* motivation premised on right-making features therefore produces only contingently—indeed, argues Sliwa, *only acciden-*

16 Kant, *Groundwork for the Metaphysics of Morals*.

17 See Markovits, "Acting for the Right Reasons." I think that challenges to Markovits's account flow from misunderstanding the ontology of reasons. I defend an analysis that closely resembles hers by rethinking the ontology of reasons in Howard, "The Goals of Moral Worth." For a competing but similar approach, see Portmore, "Moral Worth and Our Ultimate Moral Concerns."

18 Sliwa, "Moral Worth and Moral Knowledge."

tally—right action. As such, *de re* motivation cannot be the morally best kind of motivation.[19]

To sum up, both (noninstrumental) *de re* and *de dicto* moral motivation have attractive features: the former entails that the performed act is right, the latter entails that the act is performed out of explicitly moral concern. However, each approach is flawed. The former allows for the accidental overlap of morally right action with nonmoral motivation, as in the case of Kant's shopkeeper. The latter risks alienating the agent from the features of right actions that merit nonderivative concern, such as fairness, justice, the promotion of equality, honesty, and the like.

As I will now argue, nonspecific moral motivation—at least the version that I have laid out above—appears to share these two strengths without the two weaknesses. Conclusively establishing that this is true requires more space than I have here. So I will make a provisional case for nonspecific moral motivation, which I hope will spur further interest from philosophers interested in virtue and moral motivation.

De re moral motivation is appealing partly because it entails that the performed action is morally right. Nonspecific moral motivation shares this virtue. Just as "Jules wants a Big Kahuna burger" is true on its nonspecific reading only if the burgers that Jules wants are in fact Big Kahuna burgers, "Jules wants to do what's right" is true on its nonspecific reading only if what Jules wants to do is in fact morally right. *De dicto* motivation does *not* have this implication. Agents with false beliefs about what's right desire to do what's right, *de dicto*, but their desires require them to break the moral law, not follow it.

In spite of this flaw, *de dicto* moral motivation is attractive because it entails a concern for morality. It involves moral concern because it deploys the concept *morally right*. Nonspecific moral motivation also seems to entail a concern for morality, although of a different sort. For example, Jules has a nonspecific desire for a Big Kahuna burger in virtue of desiring a particular kind of burger. His desire for the Big Kahuna kind of burger, rather than another kind of burger, is explained by the very features that make a burger a Big Kahuna burger—the pineapple slice and hibiscus sauce. Likewise, Jules has a nonspecific desire for what's right in virtue of desiring a particular kind of act without desiring that kind under the description *morally right*. His desire for the morally right kind of action, rather than any other kind of action, is explained by the strength of the features that make the act right, such as its fairness or honesty. Consequently, Jules desires what's right, on the nonspecific reading, only if he is concerned with

19 I show that this argument is flawed in Howard, "One Desire Too Many."

the features that make acts right. So nonspecific moral motivation clearly entails a concern for morality.

It also appears that nonspecific moral motivation lacks the flaws associated with *de re* and *de dicto* motivation, respectively. As we saw with Kant's shopkeeper, *de re* motivation allows for only accidentally right action when doing what's morally right accidentally overlaps with what promotes the agent's nonmoral desires. Because nonspecific moral motivation, in contrast, implies that the agent desires a *kind* of action, it rules out many cases of accidental overlap. The shopkeeper desires to do what's right, *de re*, but he does not desire to do the right kind of action. So he lacks a nonspecific desire for what's right. To be clear, there is still room for some accidental overlap in cases of nonspecific moral motivation, as when an agent (nonspecifically) desires to do what's right in order to be praised. But it seems that we can rule out this kind of overlap by restricting our attention to noninstrumental, nonspecific moral motivation.

Finally, nonspecific moral motivation, unlike *de dicto* motivation, is not subject to the charge that it embodies a kind of moral fetishism that risks alienating an agent from the features of actions that make them right, which merit nonderivative concern. As we have seen, just as motivation by the features that make something a Big Kahuna burger is necessary for satisfying "Jules wants a Big Kahuna burger" on its nonspecific reading, motivation by the features that make actions morally right is necessary for satisfying "Jules wants to do what's right" on its nonspecific reading. Consequently, nonspecific moral motivation does not alienate agents from the features of actions that make them right. On the contrary, it implies motivation by them.

I fall short of full-throated endorsement of nonspecific moral motivation. The features that make nonspecific moral motivation attractive are at least partly grounded in contestable claims about hyperintensionality and how it constrains desire attributions. Nevertheless, nonspecific moral motivation clearly deserves attention from theorists working on questions of moral worth and moral fetishism, beyond discussion here.[20]

Texas A&M University
nrhoward@tamu.edu

20 This paper originated in a reading group with Caleb Perl and Jonathan "Disco" Wright on Von Fintel and Heim's *Intensional Semantics* during summer 2014 at the University of Southern California. Particular thanks to them for many happy memories. Thanks also to, in no particular order, Mark Schroeder, Steve Finlay, Ralph Wedgwood, Nate Charlow, Nicholas Laskowski, Renee Bolinger, Maegan Fairchild, Joe Horton, Alex Dietz, Kenneth Silver, Max Hayward, and Kenny Easwaran.

REFERENCES

Aboodi, Ron. "One Thought Too Few: Where De Dicto Moral Motivation Is Necessary." *Ethical Theory and Moral Practice* 20, no. 2 (April 2017): 223–37.

Bäuerle, Rainer. "Pragmatisch-semantisch Aspekte der NP-Interpretation." In *Allgemeine Sprachwissenschaft, Sprachtypologie und Textlinguistik: Festschrift für Peter Hartmann*, edited by Manfred Faust, Roland Harweg, Werner Wienold, and Götz Lehfeldt, 121–31. Tübingen, Germany: Gunter Narr, 1983.

Braun, David. "Desiring, Desires, and Desire Ascriptions." *Philosophical Studies* 172, no. 1 (January 2015): 141–62.

Carbonell, Vanessa. "De Dicto Desires and Morality as Fetish." *Philosophical Studies* 163, no. 2 (March 2013): 459–77.

Copp, David. "Moral Obligation and Moral Motivation." *Canadian Journal of Philosophy Supplementary Volume* 21 (1995): 187–219.

Dancy, Jonathan. *Ethics without Principles*. Oxford: Oxford University Press, 2000.

———. "Should We Pass the Buck?" In *Recent Work on Intrinsic Value*, edited by Toni Rønnow-Rasmussen and Michael J. Zimmerman, 33–44. Dordrecht: Springer, 2005.

Dreier, James. "Dispositions and Fetishes: Externalist Models of Moral Motivation." *Philosophy and Phenomenological Research* 61, no. 3 (November 2000): 619–38.

Égré, Paul. "Hyperintensionality and *De Re* Beliefs." In *Epistemology, Context, and Formalism*, edited by Franck Lihoreau and Manuel Rebuschi, 213–44. New York: Springer, 2014.

Fara, Delia Graff. "Specifying Desires." *Noûs* 47, no. 2 (June 2013): 250–72.

Fodor, Janet Dean. *The Linguistic Description of Opaque Contexts*. 1979. Abingdon, UK: Routledge, 2014.

Groenendijk, Jeroen, and Martin Stokhof. "Semantic Analysis of 'Wh' Complements." *Linguistics and Philosophy* 5, no. 2 (1982): 175–233.

———. *Studies in the Semantics of Questions and the Pragmatics of Answers*, PhD dissertation, University of Amsterdam, 1984.

Howard, Nathan Robert. "The Goals of Moral Worth." *Oxford Studies in Metaethics*, forthcoming.

———. "One Desire Too Many." *Philosophy and Phenomenological Research* 102, no. 2. (2021): 302–317.

Isserow, Jessica. "Moral Worth and Doing the Right Thing by Accident." *Australasian Journal of Philosophy* 97, no. 2 (2019): 251–64.

———. "Moral Worth: Having It Both Ways." *Journal of Philosophy* 117, no. 10 (October 2020): 529–56.

Kant, Immanuel. *Groundwork for the Metaphysics of Morals*. Translated and edited by Allen W. Wood. New Haven: Yale University Press, 2018.

Keshet, Ezra, and Florian Schwartz. "De Re/De Dicto." In *The Oxford Handbook of Reference*, edited by Jeanette Gundel and Barbara Abbott, 168–202. Oxford: Oxford University Press, 2019.

Lillehammer, Hallvard. "Smith on Moral Fetishism." *Analysis* 57, no. 3 (July 1997): 187–95.

Markovits, Julia. "Acting for the Right Reasons." *Philosophical Review* 119, no. 2 (April 2010): 201–42.

Olson, Jonas. "Are Desires *De Dicto* Fetishistic?" *Inquiry* 45, no. 1 (2002): 89–96.

Philips-Brown, Milo, and Lyndal Grant. "Getting What You Want." *Philosophical Studies* 177, no. 7 (July 2020): 1791–1810.

Portmore, Douglas W. "Moral Worth and Our Ultimate Moral Concerns." *Oxford Studies in Normative Ethics*, forthcoming.

Sharvit, Yael. "Embedded Questions and 'De Dicto' Readings." *Natural Language Semantics* 10, no. 2 (2002): 97–123.

Sliwa, Paulina. "Moral Worth and Moral Knowledge." *Philosophy and Phenomenological Research* 93, no. 2 (September 2016): 393–418.

Smith, Michael. *The Moral Problem*. Malden, MA: Blackwell, 1994.

Stratton-Lake, Philip. *Kant, Duty, and Moral Worth*. London: Routledge, 2000.

Von Fintel, Kai, and Irene Heim. "Beyond *De Re/De Dicto*: The Third Reading." *Intensional Semantics*. Unpublished manuscript, 2011.

Williams, Bernard. "Persons, Character, and Morality." In *Moral Luck*, 1–19. Cambridge: Cambridge University Press, 1981.

Woods, Jack. "Logical Indefinites." *Logique et Analyse* 57, no. 227 (September 2014): 277–307.

Journal of Ethics and Social Philosophy
VOL. 20, NO. 3 · NOVEMBER 2021
https://doi.org/10.26556/jesp.v20i3.1438

HOW WE CAN MAKE SENSE OF CONTROL-BASED INTUITIONS FOR LIMITED ACCESS CONCEPTIONS OF THE RIGHT TO PRIVACY

Björn Lundgren

THERE IS a long-standing discussion on whether privacy and/or the right to privacy should be conceptualized in terms of *limited access* or *control (of access)* to certain private matters (whatever those are).[1] Although control-based conceptions are among the most popular conceptions of privacy and the right to privacy, such conceptions suffer from various counterexamples.[2] However, those who argue against control-based conceptions of privacy or the right to privacy rarely attempt to explain how competing conceptions of privacy or the right to privacy can make sense of some arguably strong control-based intuitions. For example, while presenting a dilemma against control-based conceptions of privacy, I myself have acknowledged that there are strong intuitions in favor of control-based conceptions of privacy, noting—somewhat summarily—that perhaps these intuitions can be better explained by contextual accounts.[3] This is potentially problematic because, even if we find the counterexamples convincing, we may similarly find the control-based intuitions to be strong. Moreover, many recent counterexamples are only concerned with control-based conceptions of privacy, not the right to privacy. Indeed, the dilemma I present is only a dilemma for control-based conceptions of privacy. However, given that privacy is the object of the right to privacy, it follows that if privacy

1 I introduced the term "private matters" in "A Dilemma for Privacy as Control." This term refers to the things we should have *control* over or things that others should have *limited access* to (e.g., personal information and our bodies).

2 See, e.g., Parent, "Privacy, Morality, and the Law," "Recent Work on the Concept of Privacy," and "A New Definition of Privacy for the Law"; Macnish, "Government Surveillance and Privacy in a Post-Snowden World"; Lundgren, "A Dilemma for Privacy as Control"; and Solove, *Understanding Privacy*.

3 Lundgren, "A Dilemma for Privacy as Control," 165–66n1.

cannot be defined in terms of control, neither should the right to privacy.[4] This raises the question of whether conceptual consistency is more important than intuitions in determining the right way to conceptualize the right to privacy.

In this article, I aim to remedy this situation by showing how limited access conceptions of the right to privacy can satisfy control-based intuitions while providing a satisfactory alternative explanation for these intuitions. The focus is only on the right to privacy, not privacy as such.[5] Furthermore, the focus is only on the moral right to privacy, not the legal right to privacy. Moreover, my intention is not to defend control-based intuitions; rather, I will present and explain these intuitions and then show how a limited access conception can be modified to address these intuitions and yield the same conclusion as a control-based conception. The question of whether we *should* make such a modification is not something I aim to settle in this paper, although I will make brief comments on this subject.

The modifications I will propose herein are based on the idea that risk taking can violate or infringe upon the right to privacy. I proposed this idea previously, based on Sven Ove Hansson's more general idea of a *pro tanto* right against risks.[6] I also defended this idea more explicitly in a recent publication, in which I argued that it makes sense to think of the right to privacy as being violated or infringed upon in cases where someone attempts to access private matters. Moreover, I qualified these attempts in terms of substantial risks.[7] The publication also critiqued an analysis of the right to privacy in terms of negative control by Jakob Thraine Mainz and Rasmus Uhrenfeldt.[8] I argued that my presumptions followed from Mainz and Uhrenfeldt's arguments, and these presumptions can

4 Lundgren, "A Dilemma for Privacy as Control," 166.

5 There are at least three reasons for this. First, as I noted in "A Dilemma for Privacy as Control," control-based conceptions are mostly popular due to control-based conceptions of the right to privacy, not due to control-based conceptions of privacy. Second, intuitions in favor of control-based conceptions are arguably stronger for conceptions of the right to privacy, rather than conceptions of privacy as such. Third, as noted above, some substantial arguments against control-based accounts of privacy deal primarily with privacy rather than the right to privacy.

6 Lundgren, "Against AI-Improved Personal Memory"; Hansson, "Ethical Criteria of Risk Acceptance." Hansson called this right *prima facie*, but as I have noted (Lundgren, "Against AI-Improved Personal Memory," 229n39), what he proposed is better described as a *pro tanto* right.

7 Lundgren, "Confusion and the Role of Intuitions in the Debate on the Conception of the Right to Privacy."

8 Mainz and Uhrenfeldt, "Too Much Info."

also be used to create a counterexample against their proposed definition of the right to privacy.[9]

Lauritz Aastrup Munchalso proposed the idea that the right to privacy should provide moral protection against risks.[10] Munch's work focuses on the abuse of information. Moreover, similar to myself, he proposed this idea based on works concerning the normative aspects of risks, albeit by a different author (John Oberdiek). More recently, Munch provided an extensive defense of the idea that probabilistic inferences can violate the right to privacy.[11]

The aim of the present article differs from the aims of these prior works. More specifically, my aim is to show how control-based intuitions can be explained based on the idea that the right to privacy protects against certain forms of risk.

To simplify this discussion and show how these ideas can be generalized, I will introduce a *simplified schema* of necessary and jointly sufficient conditions for when a standard control-based conception (*C*) and a standard, limited access-based conception (*L*) are infringed upon or violated:[12]

> *C*: A control (access) conception of the right to privacy implies that an individual *A*'s right to privacy is infringed upon or violated if and only if another individual controls (access to) part of *A*'s private matters.
>
> *L*: A limited access conception of the right to privacy implies that an individual *A*'s right to privacy is infringed upon or violated if and only if another individual accesses part of *A*'s private matters.

Note that *C* only covers certain control-based conceptions of the right to privacy; later (in section II), I will turn to alternatives. I will begin section I by discussing

9 It should be noted that an unfortunate formulation in Lundgren, "Confusion and the Role of Intuitions in the Debate on the Conception of the Right to Privacy," may make it appear as if that article was the first source of this idea, ignoring both Lundgren, "Against AI-Improved Personal Memory"; and Munch, "The Right to Privacy, Control Over Self-Presentation, and Subsequent Harm." What I should have said was that the idea was relatively new and perhaps reasonably unknown to Mainz and Uhrenfeldt.

10 Munch, "The Right to Privacy, Control Over Self-Presentation, and Subsequent Harm."

11 Munch, "Privacy Rights and 'Naked' Statistical Evidence"; I will not discuss this article, as it was published after the present article was accepted.

12 By merely providing a schema, I am setting aside many issues that a complete conception of the right to privacy must consider (e.g., the role of informed consent, right forfeiture, and conditions—if any—when the right is overridden). What I am interested in here is whether we can modify the limited access conception (*L*) to yield the same conclusion as control-based conceptions of the right to privacy in a situation where control-based intuitions appear to be very strong.

one example of control-based intuitions from my previous writings.[13] The aim of this example is to show how we can modify *L* to yield the same evaluation as an application of *C* while ensuring that the modification of *L* does not turn into a control-based conception of the right to privacy. In section II, I will modify this example to consider alternative formulations of control-based conceptions of the right to privacy and to show how *L* can be further modified. These examples are intended to show that generalizability of the idea that a limited access account of the right to privacy can be modified to explain control-based intuitions. Finally, in the last section, I will conclude and summarize.

I

In recent work, I mentioned an example in which a prisoner is held in a cell with a hatch that someone else controls (henceforth, the "prison hatch example").[14] If we accept *C* (and set aside the issue of whether the right to privacy might have been forfeited or overridden), we should conclude that the prisoner's right to privacy is infringed upon or violated because the hatch is controlled by someone else and it follows that someone else controls (access to) the prisoner's private matters. Furthermore, proponents of control-based conceptions of the right to privacy would hold that even if the hatch is never opened, it may seem as if the prisoner's right to privacy is infringed upon or violated because the presence of the hatch affects the prisoner's control over (access to) their private matters. In contrast, if we accept *L*, we must recognize that if the hatch is not opened, then there is no infringement or violation of the right to privacy (i.e., the right to privacy would only be infringed upon in situations where the hatch is actually opened, allowing access to the prisoner's private matters). Proponents of limited access conceptions may argue that although the presence of the hatch may have negative effects on the prisoner, such as changes in their behavior (due to the potential of being surveilled), these effects can fully be explained by how it affects their autonomy, rather than their privacy. Nevertheless, this idea could also be used to speak in favor of control-based conceptions because of the close link between privacy and autonomy. Hence, I will show how it is possible to modify limited access conceptions to yield the same conclusion as control-based conceptions while explaining control-based intuitions in a way that retains a limited access conception.

As I noted in the introduction, the modification I will focus on herein recognizes that the right to privacy includes a right not to have others put one's

13 Lundgren, "A Dilemma for Privacy as Control."

14 Lundgren, "A Dilemma for Privacy as Control."

privacy at substantial risk. Hansson argued for a general right against risk exposure, which can be overridden only under certain circumstances, such that "this [risk] exposure is part of an equitable social system of risk-taking that works to her advantage."[15]

The idea that the right to privacy should include protection against risk impositions can be defended in a variety of ways. Consider a situation in which Jane voluntarily performs an action @, knowing that @ has a high probability (e.g., > 99%) of exposing a very private matter of Joe's. If we grant that exposing this very private matter would violate or infringe upon Joe's right to privacy, then should we not also grant that it would violate or infringe upon his right even if the risk is not actualized? I believe that the answer is yes, although it is important to note that the specific probability of the actualization of the risk may matter in this scenario, and I will consider this in my modification of *L*.

To more clearly understand the idea that the right to privacy should morally protect you against certain risks to your privacy, it may be useful to consider a practical example. Suppose, for example, that you store very personal information on a secure cloud service. Furthermore, suppose that someone hacks the security protection of this information so that it is accessible by anyone. Even if no one actually accesses this information, it would be reasonable to hold that the hacker has infringed upon or violated your right to privacy. Although one could alternatively argue that the hackers infringed upon or violated rights other than privacy, it would then be difficult to explain the difference between making privacy-sensitive information accessible and making privacy-insensitive information accessible. More importantly, the goal here is not to defend control-based intuitions but to show how a limited access conception can provide alternative explanations for control-based intuitions and reach the same result. While proponents of control-based conceptions would argue that risk impositions affect the right-bearer's control over their private matters, an alternative explanation is that an action that puts access under risk is an action that risks delimiting access and, hence, is an action that infringes or violates upon the right to limited access to one's private matters.[16] If we accept this idea, then we can grant that proponents of *C* are correct to consider the hatch as an infringement or violation of the

15 Hansson, "Ethical Criteria of Risk Acceptance," 305, bracketed text added.

16 Of course, control and risk taking can come apart, but this does not mean that control-based intuitions speak against intuitions in favor of the idea that a right to privacy should protect against privacy risks. In fact, as I have argued before, intuitions about risks to or attempted access may sometimes be in a better position to explain what are commonly taken to be intuitions in favor of control; see Lundgren, "Confusion and the Role of Intuitions in the Debate on the Conception of the Right to Privacy."

right to privacy but wrong in their explanation for why the hatch is an infringement or violation. We can easily modify *L* to satisfy this alternative explanation:

> *L-risk*: A limited access conception of the right to privacy implies that an individual *A*'s right to privacy is infringed upon or violated if and only if another individual *B* either (1) accesses part of *A*'s private matters *p* or (2) makes it so that *p* is at risk of being accessed by some individual *C* (such that that *C* is not *A*).

This would resolve the prison hatch example because the hatch puts *A*'s private matters at risk. Risk exposure would also provide an alternative explanation to control-based intuitions while remaining congruent with the lack of control.

However, a few possible complications may arise with this conception. First, the meaning of "*B* makes it so that" is not entirely clear. Setting aside complications about how to understand the concept of causation, there is a question as to who creates this risk. For example, is it the door maker; the state, which put *A* in this situation; or the prison ward, who makes decisions about the whether to open the hatch? However, given that a similar problem arises with control-based conceptions, we can set this issue aside (given the limited purpose of this paper). Second, *A*'s private matters are always at risk in a strict sense, making L-risk somewhat vague. What does it mean for *B* to make it so that *A*'s private matters are at risk if these matters are already at risk? We can solve this quite simply by saying, "*B* makes it so that p is at (greater) risk," thus relativizing the proposal. However, this may raise another problem, as *B* may be performing an action that only slightly increases the level of risk. That is, we may unintentionally say that one can violate or infringe upon another's right to privacy by simply performing an action that indirectly increases the level of risk to the other person's privacy by a mere fraction of a percentage. While some would agree that this is the correct understanding, we could also resolve this by requiring a substantial increase in risk. Alternatively, we could note that the privacy right against risk exposure is overridable, perhaps by modifying it according to Hansson's criteria for situations in which the right against risk exposure is overridden.[17]

II

In the previous section, I argued that, if we accept *C*, we recognize that the presence of the prison hatch infringes upon or violates the prisoner's privacy, even if it is never opened. Similarly, I showed how a limited access conception can be modified to reach the same conclusion. However, what if the hatch is not

17 Hansson, "Ethical Criteria of Risk Acceptance."

only never opened but—unbeknownst to the prisoner—cannot be opened? To understand the intuition around this case it may be more illustrative to consider a dummy camera rather than a dummy prison hatch (henceforth, the "dummy prison hatch example").[18]

Note that, according to *C*, the dummy prison hatch would not infringe upon or violate the prisoner's right to privacy. However, some proponents of control-based conceptions of the right to privacy may claim that what matters is not only control over private matters but that this control is a form of the individual's self-control. Although it may, again, be argued that such ideas conflate privacy with autonomy, I will nevertheless consider this claim herein.[19]

> *C-self*: A control (access) conception of the right to privacy implies that an individual *A*'s right to privacy is infringed upon or violated if and only if *A*'s self-control over (access to) part of *A*'s private matters is reduced.

According to *C*-self, the dummy prison hatch infringes upon or violates the prisoner's privacy because it affects the prisoner's self-control over their private matters. To see this spelled out in greater detail, consider Andrei Marmor's idea that the right to privacy is grounded in an interest to control how we present ourselves to others.[20] While we may be skeptical of spelling out the right to privacy in terms of this interest, this idea can be used to illustrate how the dummy prison hatch may infringe upon the right to privacy.[21] Simply put, the presence of a dummy hatch (or dummy camera) can affect how a person behaves. However, such an effect is explainable only if the presence of the hatch affects the person's beliefs or knowledge of potential surveillance by others. Thus, let us consider how *L*-risk can be further modified to address such intuitions.

Based on the arguments just considered, I will introduce a distinction be-

18 The comparison to the previous example would still hold. If we modify the previous example, the camera surveillance footage is not watched but nevertheless implies the substantial risk that it *could* be viewed. In the current example, a mounted dummy camera affects people's behavior because they do not know that the camera is a dummy.

19 Below, I will exemplify this view with ideas from Marmor, "What Is the Right to Privacy?" However, this idea has recently been defended in analyses of privacy (rather than the right to privacy) by Menges, "A Defense of Privacy as Control" and "Three Control Views on Privacy." As Menges notes (in "Three Control Views on Privacy") his arguments can be spelled out in defense of a conception of the right to privacy (as he plans to do). Moreover, Menges explains how his analysis of privacy differs from an analysis of autonomy ("A Defense of Privacy as Control").

20 Marmor, "What Is the Right to Privacy?"

21 For counterexamples, see Lundgren, "A Dilemma for Privacy as Control," 172n17. For a detailed critique, see Munch, "The Right to Privacy, Control Over Self-Presentation, and Subsequent Harm."

tween belief in an actual possibility (risk) and an agent's epistemic or doxastic uncertainty as to whether something is at risk. To modify *L*-risk, we have two alternatives. First, we may consider *A*'s privacy to be infringed upon or violated when *B makes A believe that their private matters are at risk of being accessed by some individual C (such that C is not A)*. Second, we may consider *A*'s privacy to be infringed upon or violated when *B makes A uncertain as to whether their private matters are at risk of being accessed by some individual C (such that C is not A)*. We could also further qualify these alternatives (e.g., by requiring the degree of uncertainty to be substantial). Adding both conditions would produce the following:

> *L-risk and epistemic and doxastic uncertainty*: A limited access conception of the right to privacy implies that an individual *A*'s right to privacy is infringed upon or violated if and only if another individual *B* either (1) accesses part of *A*'s private matters, (2) makes it so that *p* is at risk of being accessed by some individual *C* (such that *C* is not *A*), (3) makes it so that *A* believes that their private matters are at risk of being accessed by some individual *C* (such that *C* is not *A*), or (4) makes it so that *A* is (substantially) uncertain as to whether their private matters are at risk of being accessed by some individual *C* (such that *C* is not *A*).

Modifying *L*-risk accordingly would resolve the dummy prison hatch example, either because *A* believes that the hatch can be opened and thus poses a risk or because *A* is uncertain as to whether the hatch poses a risk. Arguably, if *A*'s self-control (over how she presents herself) is affected, it is affected because *A* does not know whether the hatch may be used (and similarly in the case of a dummy camera). That is, if we accept the basic intuitions here, then the right to privacy protects against manipulation of an agent's belief in (or knowledge of) the risk that others may access their private matters.

Keep in mind that the goal here is not to defend the conclusions about privacy rights infringements and violations that follow from accepting *C*-self or *L*-risk and epistemic and doxastic uncertainty. While I believe that such a conception of the right to privacy would conflate the right to privacy with the right to know that one's privacy is retained or protected, that discussion is beyond the aim and scope of this paper. The goal of this paper was merely to show that it is possible to provide an alternative explanation for control-based intuitions and a limited access conception of the right to privacy that satisfies these intuitions. By considering these alternatives, I hope to have shown that it is very likely that these type of modifications can be generalized.

CONCLUSIONS

Above, I have shown how limited access conceptions of the right to privacy can provide alternative explanations to control-based intuitions about the right to privacy. Indeed, we can modify a limited access conception of the right to privacy to yield the same conclusions as variants of control-based conceptions of the right to privacy. First, we saw how one can make sense of some very common control-based intuitions by adapting the idea that the right to privacy (in terms of limited access) also includes a right against substantial risk impositions to one's privacy (or private matters). Next, we turned to conceptions of the right to privacy that suppose a closer link between privacy and autonomy. I showed how the idea that we can infringe upon or violate someone's right to privacy by affecting their self-control can be addressed by using a limited access conception of the right to privacy and introducing criteria related to the agent's knowledge or beliefs. Specifically, an agent's right to privacy can be infringed upon or violated if their beliefs (or knowledge) about access to their private matters are (substantially) affected. The question of whether we should accept these revisions is partly beyond the scope of this paper, but I hope these examples can show how this defense of limited access analyses of the right to privacy can be generalized and adapted to address more specific control-based intuitions (as I did in previous writings).[22]

Umeå University, Institute for Futures Studies, and Stockholm University
bjorn.lundgren@umu.se

REFERENCES

Hansson, Sven Ove. "Ethical Criteria of Risk Acceptance." *Erkenntnis* 59, no. 3 (November 2003): 291–309.

Lundgren, Björn. "Against AI-Improved Personal Memory." In *Aging Between Participation and Simulation: Ethical Dimensions of Socially Assistive Technologies in Elderly Care*, edited by Jochen Vollmann, Johanna Hovemann, and Joschka Haltaufderheide, 223–34. Berlin: De Gruyter, 2020.

22 Lundgren, "Confusion and the Role of Intuitions in the Debate on the Conception of the Right to Privacy." Lastly, I want to thank an anonymous reviewer for *Journal of Ethics and Social Philosophy*. Moreover, I want to acknowledge that this work was supported by the Wallenberg AI, Autonomous Systems and Software Program—Humanities and Society (WASP-HS), funded by the Marianne and Marcus Wallenberg Foundation and the Marcus and Amalia Wallenberg Foundation (grant number: MMW 2018.0116).

———. "Confusion and the Role of Intuitions in the Debate on the Conception of the Right to Privacy." *Res Publica* (2021). https://doi.org/10.1007/s11158-020-09495-9.

———. "A Dilemma for Privacy as Control." *Journal of Ethics* 24, no. 2 (June 2020): 165–75.

Macnish, Kevin. "Government Surveillance and Privacy in a Post-Snowden World." *Journal of Applied Philosophy* 35, no. 2 (May 2018): 417–32.

Mainz, Jakob Thrane, and Rasmus Uhrenfeldt. "Too Much Info: Data Surveillance and Reasons to Favor the Control Account of the Right to Privacy." *Res Publica* 27, no. 2 (May 2021): 287–302.

Marmor, Andrei. "What Is the Right to Privacy?" *Philosophy and Public Affairs* 43, no. 1 (Winter 2015): 3–26.

Menges, Leonhard. "A Defense of Privacy as Control." *Journal of Ethics* 25, no. 3 (September 2021): 385–402.

———. "Three Control Views on Privacy." *Social Theory and Practice* (forthcoming).

Munch, Lauritz Aastrup. "Privacy Rights and 'Naked' Statistical Evidence." *Philosophical Studies* 178, no. 1 (November 2011): 3777–95.

———. "The Right to Privacy, Control over Self-Presentation, and Subsequent Harm." *Journal of Applied Philosophy* 37, no. 1 (February 2020): 141–54.

Parent, W. A. "A New Definition of Privacy for the Law." *Law and Philosophy* 2, no. 3 (December 1983): 305–38.

———. "Privacy, Morality, and the Law." *Philosophy and Public Affairs* 12, no. 4 (Autumn 1983): 269–88.

———. "Recent Work on the Concept of Privacy." *American Philosophical Quarterly* 20, no. 4 (October 1983): 341–55.

Solove, Daniel J. *Understanding Privacy*. Cambridge, MA: Harvard University Press, 2008.