JOURNAL *of* ETHICS
& SOCIAL PHILOSOPHY

VOLUME XXVI · NUMBER 2
*December 2023*

EDITORIAL

ARTICLES

journal *of* ethics & social philosophy
*http://www.jesp.org*

The *Journal of Ethics and Social Philosophy* (issn 1559-3061) is a peer-reviewed online journal in moral, social, political, and legal philosophy. The journal is founded on the principle of publisher-funded open access. There are no publication fees for authors, and public access to articles is free of charge and is available to all readers under the creative commons attribution-noncommercial-noderivatives 4.0 license. Funding for the journal has been made possible through the generous commitment of the 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. 26, no. 2 · december 2023

*https://doi.org/10.26556/jesp.v25i3.3083*

*Editorial*

# TRIALS AND TRIUMPHS OF UNIVERSITY-FUNDED OPEN-ACCESS PUBLISHING

The *Journal of Ethics and Social Philosophy* (jesp) was founded in 2005 with the conviction that universities must be in the business of disseminating knowledge, not of handing it for free to private stakeholders who profit by locking it away. In the years since, this mission has proven to be even more important, as for-profit publishers have co-opted the concept of "open access" as a name for charging authors exorbitant fees to make their articles accessible on top of the existing profit stream from university libraries that are still paying for subscription access. Our model has proven again and again that it does not cost $2,500 in "open-access fees" to copyedit, publish, and preserve a thirty-page philosophy article, and that great financial and epistemic benefits to universities arise from ditching the middleman and directly funding publication and dissemination of academic research—not only in making the fruits of research accessible to everyone, but also in giving everyone access to publish their research on its own merits.

Still, as we have also learned over the years, there are many obstacles to running a university-funded open-access journal. While the barriers to entry are low—anyone who can spare the time and has access to a web server can in principle run such a journal—the costs of growth and institutionalization are high. University budgets are divided into research budgets and library budgets; the research budgets are designed to be spent on paying someone else to publish and the library budgets are built around paying someone else for access. This feature of institutional design leaves no one below the level of university provost or president with the power and scope of decision-making to enact substantive change, and it turns out that university provosts and presidents have many other priorities.

Every journal editor, of course, has a challenging job. It includes recruiting willing and reliable associate editors and referees, obtaining quality submissions, and dealing with the inevitable fact that it is flatly impossible to make everyone happy. Even the most optimal submission-evaluation process will, like any medical test, be subject to both false positives and false negatives. Their journal's process will, of course, depend on the cooperation of too many people to be anything close to optimal. And the challenges of occupying this role are incessant—authors can submit to your journal at any time of day or night, potential referees can decline requests, committed referees can go AWOL, and, as the journal's editor, you live and breathe the fact that every gap between one person doing their step in moving a submission forward and the next taking it up is another day (or week) that the submitting author has to wait for a verdict—a pain that you have felt on the other side many times.

But the structures of twenty-first-century academic publishing also present many additional challenges to would-be open-access editors with the energy and enthusiasm to overcome the lack of institutional initiative. They must navigate a wide range of issues, including how to implement a robustly triple-blind process, often using tools that are not well-designed for it, designing and implementing typesetting and journal style, learning how to assign and register DOIs, archiving publications against the risk of possible future collapse of the journal's funding, hiring and managing copy editors, handling tech issues with the journal's content management system, and resetting authors' forgotten passwords—all of which fall directly onto their plate.

And for the would-be open-access journal editor, these substantial difficulties also come with the further challenge of securing ongoing funding to support their journal's operations. The initial funding commitment from a favorable dean may evaporate when the dean's successor goes looking for soft spots in their budget, or the journal's success may fuel growth that outpaces the resources originally envisioned. They may spend substantial time and effort writing for grant support for their journal, petitioning university librarians, navigating the technical and legal issues to create web-based donation portals, or resorting to fees or "suggested donations" to authors when all of the institutions whose ostensible function is to support and disseminate research cannot find room for it in their budgets because they need to spend those dollars paying for access to paywalled research or "open-access fees." And they do all of these things as a volunteer.

Finally, even the successful open-access editor must, at the end of all of this work, confront their most important obstacle. Because the institution of their journal is not larger than themselves, they cannot simply resign and trust the owners or operators of the journal to secure an able successor. Instead, their most important and also most difficult job is to find and secure their own successor—someone they trust to carry forward their open-access mission, protect the reputation of the journal even while adding their own editorial vision, and secure the funding support required to sustain the journal, which, if they are successful and submissions continue to grow, will only become more difficult over time. And they must somehow recruit this successor with clear eyes about the kinds of frustrations and challenges inherent to operating a fully open-access journal.

Through all of these challenges, *JESP* has by all accounts thrived. The number of submissions that we receive has grown, the quality of submissions that we receive is up, and the amount that we publish is higher as well. The pool of authors who submit to and publish in *JESP* has grown more diverse. *JESP* continues to be the best venue in the field to publish discussion notes in ethics or related areas. And I am proud to report that *JESP* is now living up to its title as a journal of not just *ethics*, but also of *social philosophy*, publishing exciting new work on race, gender, disability, relationships, parenting, the family, and more, in addition to continuing to publish great work in normative ethics, metaethics, practical reason, moral responsibility, legal philosophy, political philosophy, and value theory. Though we have faced challenges over these years and not every decision that I have made has been the correct one, each of these is, I hope, an improvement for the experience of readers, authors, referees, and editors working with the journal. And my reward for the time commitment and challenges of running the journal has been the kind words that you all have shared with me about how much you appreciate the journal and recognize the quality of what is published in its pages.

I assumed the editorship of *JESP* on December 1, 2014, from my colleague and *JESP*'s founding editor, Andrei Marmor, who was at that time moving from USC to take up a position at Cornell. I took it with the goal of keeping *JESP* at USC, which had funded the journal through its first ten years and promised to continue to do so, and with the expectation that it could be up to a ten-year commitment. It has now been just over nine years, during which I have updated the look and feel of the journal, moved to a new and more robust online content-management system, integrated the journal's publication system more thoroughly into the modern publication system including registration of DOIs and improved archiving, grown the team of associate editors, transitioned to a new workflow better suited to the growth in submission volume, and overseen a growth from publishing about three issues of three articles each per year to now publishing as many as ten issues per year with seven full articles plus discussion notes in each issue.

But it is now time for me to fulfill the hardest part of this job. It is time for new leadership to take the journal to yet higher levels.

I am therefore delighted to announce that, as of January 1, 2024, editorial leadership of and institutional support for *JESP* will pass to Sarah Paul and Matthew Silverstein of New York University Abu Dhabi (NYUAD). NYUAD's logo features a torch as a symbol of light cast into darkness, and there is no better manifestation of a university's commitment to cast light into the darkness than institutional support for access to a publication that is free at point of access to both authors and readers. But it is also a symbol, for me, of the torch that I am passing to Paul and Silverstein, and that USC is passing to NYUAD. The financial support that NYUAD is providing to *JESP* and its mission represents a major commitment and shows true leadership among world universities in protecting *JESP*'s vision for open knowledge. I hope for a world in which more universities follow their lead.

I am grateful as well for the trusted hands into which I am able to pass the journal. In addition to being distinguished scholars, capable administrators, and deeply familiar with editing and journal processes, both Paul and Silverstein are longtime, enthusiastic supporters of *JESP* as readers, authors, and referees. Paul's article "Deviant Formal Causation" appeared in volume 5, issue 3 in 2011. Silverstein's "Inescapability and Normativity" appeared in volume 6, issue 3 the following year, and his "Reducing Reasons" appeared in volume 10, issue 1 in 2016. Since 2017, Silverstein has painstakingly typeset every article that has appeared in *JESP*, starting with volume 12—by my count, 255 articles in total by the time this note is published, fully 60 percent of all articles published in the history of the journal.

There is no one—or ones!—whose judgment or commitment I could trust more, and they have a shared vision and purpose to continue to take *JESP* to new and better places in the kinds and quality of work that it publishes and in the experiences that it offers readers, authors, and everyone else whose hard work makes the journal tick. They will face trials, it is true, but no one is better suited to triumph. I can't wait to see where they take it.

*Mark Schroeder*
EXECUTIVE EDITOR

*Journal of Ethics and Social Philosophy*
vol. 26, no. 2 · december 2023

*https://doi.org/10.26556/jesp.v26i2.2449*

# ALLIES AGAINST OPPRESSION

## INTERSECTIONAL FEMINISM, CRITICAL RACE THEORY, AND RAWLSIAN LIBERALISM

*Marcus Arvan*

Liberalism is often claimed to be at odds with feminism and critical race theory (crt). On the one hand, many feminists and critical race theorists criticize liberalism for inadequately addressing oppression.[1] On the other, some contend that feminism and crt conflict with liberal commitments to objectivity, fallibility, and pluralism.[2] In response, some argue that liberalism can be deracialized and feminist.[3] Still, the most influential contemporary liberal political theory—John Rawls's theory of justice as fairness—has been criticized by feminists and crt as a pernicious ideology that problematically abstracts away from historical and present-day injustice.[4] These criticisms have been challenged.[5] However, they remain common and have disseminated into popular discourse.[6]

[1] Crosthwaite, "Feminist Criticism of Liberalism"; Delgado and Stefancic, *Critical Race Theory*, 21–25; Lawrence et al., "Introduction," 3; Nussbaum, *The Feminist Critique of Liberalism*; Schwartzman, *Challenging Liberalism*; and Young, *Justice and the Politics of Difference*, 166–67, 228.

[2] Sullivan, "Removing the Bedrock of Liberalism." See also *Economist*, "The Threat from the Illiberal Left"; Kapoor, "Feminism Is Illiberal"; Powers, "Illiberal Feminism Is Running Amok"; and Rauch, *The Constitution of Knowledge*.

[3] Hartley and Watson, "Is a Feminist Political Liberalism Possible?"; Hay, *Kantianism, Liberalism, and Feminism*; Mills, *Black Rights/White Wrongs*, 201–16, and "Occupy Liberalism!"

[4] Kang, "Can Rawls's Nonideal Theory Save His Ideal Theory?"; Mills, "'Ideal Theory' as Ideology," "Rawls on Race/Race in Rawls," and "Retrieving Rawls for Racial Justice?"; Okin, "Justice and Gender" and "Justice, Gender, and the Family"; and Young, *Justice and the Politics of Difference*, chs. 1 and 2, esp. pp. 16–18, 20, 104–5.

[5] Arvan, "First Steps toward a Nonideal Theory of Justice," "Nonideal Justice as Nonideal Fairness," and "Educational Justice and School Boosting"; Matthew, "Rawlsian Affirmative Action," "Rawls's Ideal Theory," and "Rawls and Racial Justice"; Shelby, "Race and Social Justice," and "Racial Realities and Corrective Justice."

[6] Barndt, "The God Trick"; Britton-Purdy, "What John Rawls Missed"; Forrester, *In the Shadow of Justice*; Kang, "Can Rawls's Nonideal Theory Save His Ideal Theory?"; Mills,

This paper argues that Rawlsian liberalism, far from being at odds with feminism and CRT, lends additional support to their central commitments. To be clear, I do not mean that feminism and CRT need a Rawlsian justification. My argument is merely that Rawlsian liberalism should be understood as their *ally*. This paper also recognizes that intersectional feminism and CRT are diverse, such that certain theoretical lenses sometimes utilized in these fields—such as Marxism or postmodernism—may not entirely align with Rawlsian liberalism.[7] My argument does not erase or denigrate these differences. Instead, its point is merely that the central commitments of feminism, CRT, and Rawlsian liberalism converge far more than commonly recognized. Furthermore, or so I shall contend using recent findings in moral and social psychology, this may be of real practical importance. It may be vital to build bridges between liberalism, feminism, and CRT to achieve greater solidarity on the political left for dismantling oppression and undermining right-wing narratives opposed to social justice activism.

Section 1 argues that Rawlsian liberalism supports the central commitments of intersectional feminism. Section 2 argues that the same is true of CRT. Whereas sections 1 and 2 make these arguments programmatically, section 3 uses Iris Marion Young's "Five Faces of Oppression"—a classic work widely utilized in feminism and CRT to theorize and contest diverse oppressions—to illustrate how Rawlsian liberalism supports similar goals and projects, and why it may be critical to achieve solidarity between feminism, CRT, and Rawlsian liberalism. Finally, section 4 responds to five objections, including concerns that my argument may violate requirements of justice related to "speaking for others," allyship, epistemic appropriation, and intellectual gentrification.[8] While I take these concerns very seriously, I contend that my argument only supports the work of marginalized scholars, activists, and groups in ways that may beneficially broaden solidarity and allyship in pursuit of eliminating all forms of oppression.

*Black Rights/White Wrongs*, chs. 8 and 9; and Petri, "Sorry I Can't Comment on the President's Actions, I Just Remembered I'm Turning into a Bird."

[7] Schneider, "Integrating Critical Race Theory and Postmodern Implications of Race, Class, and Gender"; Stefano, "Marxist Feminism"; Young, "Post Race Posthaste"; and Young, *Justice and the Politics of Difference*, 3, 7, 10, 36. Cf. Federici, *Caliban and the Witch*; and Leeb, "Marx and the Gendered Structure of Capitalism."

[8] Curry, "Racism and the Equality Delusion"; Davis, "On Epistemic Appropriation"; Edwards, "Aspiring Social Justice Ally Identity Development"; Minh-ha, *Woman, Native, Other*; and Trebilcot, "Dyke Methods."

## 1. Intersectional Feminism as a Liberal Requirement of Fairness

Liberalism, as a rough approximation, takes "protecting and enhancing the freedom of the individual to be the central problem of politics."[9] Liberals generally agree on some things, such as individuals' rights to free speech, freedom of religion, to vote, and so on.[10] However, liberalism also has many variants, ranging from classical liberalism (which defends *laissez-faire* free markets), to liberal egalitarianism (which mandates fair distributions of socioeconomic goods), to cosmopolitan egalitarianism (which mandates global fairness).[11] This means that "liberalism is more than one thing."[12] Nevertheless, many feminists and critical race theorists object to liberalism's individualism and to John Rawls's liberal-egalitarian theory for merely giving an ideal theory of a "fully just society" that abstracts away from injustices.[13]

I believe there is real merit in criticisms of classical free-market liberalism, which Rawls's liberal-egalitarian theory also opposes.[14] Critics are also correct that Rawls never adequately addresses serious real-world injustices, including injustices concerning the Global South.[15] However, as Rawls explains and others emphasize, ideal theory arguably plays a critical role in social-political philosophy: it provides a *measure* of how unjust a society (and the world more generally) is and has been in the past.[16] Second, although Rawls recognized that addressing injustice is a further question of "nonideal theory," other authors

[9] Girvetz, Dagger, and Minogue, "Liberalism."

[10] Girvetz, Dagger, and Minogue, "Liberalism"; see especially the section titled "Rights."

[11] Blake and Smith, "International Distributive Justice," sec. 1; Courtland, Gaus, and Schmidtz, "Liberalism," sec. 2.

[12] Courtland, Gaus, and Schmidtz, "Liberalism."

[13] Delgado and Stefancic, *Critical Race Theory*, 28–29; Goodhart, *Injustice*; Mills, *Black Rights/White Wrongs*, "'Ideal Theory' as Ideology," and *The Racial Contract*; Young, *Justice and the Politics of Difference*, 36, 74–76, 228. Cf. Rawls, *A Theory of Justice*, 4–5, 216–17.

[14] Rawls, *A Theory of Justice*, 62–63.

[15] In *A Theory of Justice*, Rawls does address civil disobedience to unjust laws and conscientious refusal to obey unjust legal injunctions (319–46), and in *The Law of Peoples*, Rawls addresses just war theory and assisting "burdened societies" (pt. 3). However, Rawls fails to adequately address domestic, international, and global injustices more generally. See Arvan, "A Non-Ideal Theory of Justice," "First Steps toward a Nonideal Theory of Justice," and "Nonideal Justice as Nonideal Fairness"; Mills, "'Ideal Theory' as Ideology" and *The Racial Contract*. See also Phillips, "Reflections on the Transition from Ideal to Non-Ideal Theory."

[16] Rawls, *A Theory of Justice*, 7–8, 216–17, and sec. 53. See Matthew, "Rawls and Racial Justice," "Rawlsian Affirmative Action," and "Rawls's Ideal Theory"; Shelby, "Race and Social Justice" and "Racial Realities and Corrective Justice"; and Simmons, "Ideal and Nonideal Theory." See also Erman and Möller, "Is Ideal Theory Useless for Non-Ideal Theory?" and "Three Failed Charges against Ideal Theory."

have taken up the task of extending Rawlsian liberalism to nonideal theory.[17] We will now see that Rawlsian ideal and nonideal theory together support the central commitments of intersectional feminism.

### *1.1 Intersectionality and Inclusivity as Liberal Requirements of Fairness*

Intersectionality is widely recognized in feminism and CRT as an important tool for recognizing, understanding, and dismantling injustice.[18] However, its nature remains contested, and there is "incredible heterogeneity" in how it is understood.[19] Whereas some interpret intersectionality as a theory of social kinds, experience, or oppression, others understand it in terms of multifactor analyses or causal modeling, and others still understand it as a critical praxis or advocacy strategy to inform inclusive social activism and solidarity politics.[20] Fortunately, irrespective of these disagreements, intersectionality clearly sets a regulative ideal: it "requires activists and inquirers to treat existing classification schemes as if they are indefinitely mutually informing, with the specific aim of revealing and resisting inequality and injustice."[21] Intersectionality's central insight is that social identities are interconnected, revealing intersecting axes of discrimination, disadvantage, and privilege faced by members of different social groups.[22] For example, Black boys and men face specific oppressions—such as police profiling, violence, and mass incarceration—not simply as members of one oppressed social category (being Black) but instead due to specifically being *Black males*.[23] This is important for many reasons, including because it reveals that a social category (being male) that confers unjust privilege to members of some categories (e.g., White heterosexual cisgender men)

[17] Rawls, *A Theory of Justice*, 215–16. See Arvan, "First Steps toward a Nonideal Theory of Justice." See also Arvan, "A Non-Ideal Theory of Justice" and "Nonideal Justice as Nonideal Fairness." Cf. Adams, "Nonideal Justice, Fairness, and Affirmative Action."

[18] Delgado and Stefancic, *Critical Race Theory*, 10–12 and ch. 2; Evans and Lépinard, "Confronting Privileges in Feminist and Queer Movements"; and Gasdaglis and Madva, "Intersectionality as a Regulative Ideal."

[19] Collins and Bilge, *Intersectionality*, 2.

[20] Bright, Malinsky, and Thompson, "Causally Interpreting Intersectionality Theory"; Collins and Bilge, *Intersectionality*, 50–55; Crenshaw, "Demarginalizing the Intersection of Race and Sex"; Dubrow, "Why Should We Account for Intersectionality in Quantitative Analysis of Survey Data?"; Evans and Lépinard, "Confronting Privileges in Feminist and Queer Movements," 5–10; Roth, "Intersectionality and Coalitions in Social Movement Research," sec. 2; and Ruíz, "Framing Intersectionality."

[21] Gasdaglis and Madva, "Intersectionality as a Regulative Ideal," 1288.

[22] Coleman, "What's Intersectionality?"

[23] Curry, *The Man-Not*. See also Alexander, *The New Jim Crow*.

can generate unique forms of oppression for members of other identities (e.g., LGBTQIA+ men who are BIPOC [Black, Indigenous, or People of Color], etc.).

Intersectionality is also thought to support particular methods for understanding and combating injustice. First, it is thought to support standpoint epistemology.[24] Because members of different intersecting groups experience different forms of oppression on a daily basis in ways that may be obscured from individuals in other social categories, members of particular oppressed groups appear to be better situated to recognize and understand those forms of oppression than members of other groups, particularly unjustly privileged groups.[25] Second, intersectionality is thought to require inclusivity, such as trans-inclusive feminism and transfeminism.[26] For, if members of different social identities face different but overlapping forms of oppression and have epistemically privileged standpoints on those oppressions, then understanding and effectively combating all forms of injustice requires *including* members of all oppressed groups in theorizing and activism, without any forms of domination or exclusion.[27]

However, as important as intersectionality is, one common concern is that it lacks a clear definition or criteria for distinguishing genuine forms of intersectional oppression from ersatz claims that may uphold unjust privilege.[28] First, there is again "tremendous heterogeneity" in how intersectionality is understood, such that "if we were to ask … [scholars, policy advocates, practitioners, and activists], 'What is intersectionality?', we would get varied and sometimes contradictory answers."[29] As another book surveying the field explains:

> When is intersectionality achieved …? Is it a process, a challenge, or an objective that can be measured?… While intersectionality has become a central way to define and analyse feminist and queer movements, determining how to measure or capture, when, where, how, whether, and why intersectionality has been achieved, attained, or performed, remains an open, and debatable question.[30]

To take two cases of problems these disagreements can generate, "gender-critical" feminists allege that trans-inclusive activism oppresses children and ciswomen, and men's rights activists allege that "toxic feminism" oppresses White

[24] Yuval-Davis, "Dialogical Epistemology."

[25] Collins, "Learning from the Outsider Within"; and McKinnon, "Trans*formative Experiences," sec. 4.

[26] Lépinard, "Impossible Intersectionality?"; Koyama, "The Transfeminist Manifesto."

[27] See hooks, *Feminist Theory*.

[28] Davis, "Intersectionality as Buzzword"; and Nash, "Re-Thinking Intersectionality."

[29] Collins and Bilge, *Intersectionality*, 1.

[30] Evans and Lépinard, "Confronting Privileges in Feminists and Queer Movements," 6.

cisgender men.[31] While many (rightly) find such arguments unpersuasive, other intersectional debates—such as whether Islamic veiling oppresses Muslim girls and women—remain "divisive and conflictual in the feminist movement."[32] Second, intersectional oppression is closely related to unjust privilege—since "for every oppressed group there is a group that is *privileged* in relation to that group."[33] However, while privilege is "usually taken to be intimately associated with ideas surrounding power, oppression, and inequality," it is "also clear that the term is frequently deployed without any specificity and, moreover, … is often elided with 'power.'"[34] What is widely accepted is that "privilege is broadly understood as referring to 'unearned' advantages or benefits which society grants to individuals and specific groups … for example, white privilege or male privilege."[35] Yet this means that to fully understand intersectional privilege and oppression, we must know what makes socially conferred advantages *unearned* (an issue I return to shortly). Finally, insofar as some feminists follow Young in "displacing the distributive paradigm"—rejecting the notion that justice is primarily a matter of distributing rights, opportunities, and socioeconomic resources—some commentators "have been especially troubled by the decreasing focus on social inequality within intersectionality's scholarship."[36]

Of course, some intersectional feminists have offered resolutions to these issues. For example, Elena Ruíz distinguishes "operative intersectionality"—which focuses on abstract, academic examinations of "the *operation of power*" and "identifying primary features of social identity subject to power variances in culture"—from intersectionality as a liberation epistemology, which focuses on decolonization and "critical examinations of lived experience … *for the purposes of liberation from oppression*."[37] Ruíz then contends that "criticisms of intersectionality are largely criticisms of operative intersectionality" and, thus,

[31] Ditum, "Trans Rights Should Not Come at the Cost of Women's Fragile Gains"; Joyce, *Trans*; and Salzman, "Toxic Feminism."

[32] Higgins, "Three Hypotheses for Explaining the So-Called Oppression of Men"; Lépinard, *Feminist Trouble*, 32 and ch. 3; and Zanghellini, "Philosophical Problems with the Gender-Critical Feminist Argument against Trans Inclusion."

[33] Young, *Justice and the Politics of Difference*, 42.

[34] Evans and Lépinard, "Confronting Privileges in Feminist and Queer Movements," 12–13. Cf. McIntosh, "Reflections and Future Directions for Privilege Studies."

[35] Evans and Lépinard, "Confronting Privileges in Feminist and Queer Movements," 13.

[36] Collins and Bilge, *Intersectionality*, 227. Cf. Young, *Justice and the Politics of Difference*, ch. 1. Cf. Enslin and Tjiattas, "Educating for a Just World without Gender"; and Fricker, *Epistemic Injustice*, 155. Fricker understands oppression in *discriminatory* rather than distributive terms.

[37] Ruíz, "Framing Intersectionality," 336–37, and 342. Cf. Roth, "Intersectionality and Coalitions in Social Movement Research." Roth distinguishes "structural" from "political" intersectionality.

that "intersectional social theory is an important analytic tool ... but not in its current academic usage."[38] Other feminist nonideal theorists have offered detailed analyses of particular forms of unjust privilege and disadvantage—such as racial segregation, White feminism, and transnational missionary feminism—in efforts to clearly distinguish genuine from ersatz oppressions, often in ways that link oppression to distributional inequalities.[39]

Still, because intersectionality's nature remains contested, it would be a strong mark in favor of a theory of justice if it provided a compelling account of unearned benefits and clear principles for identifying, distinguishing, and evaluating different forms of intersectional oppression and privilege. As we will now see, Rawlsian liberalism not only supports intersectionality but can help with these issues *through* distributive justice arguments that support rather than supplant feminist analyses of power, privilege, and oppression.

Let us begin with the idea of unjust privilege as unearned social advantages.[40] Rawls presents his liberal model of justice as fairness—the "original position"—as an account of precisely this. The original position's "veil of ignorance" prevents citizens from using knowledge of their own identity (e.g., their race, gender, religion) to tailor principles of justice to their own unique advantage.[41] It is thus a device that "ensures that *no one* is advantaged or disadvantaged ... by the outcome of *natural chance or social contingencies*."[42] This means that whichever principles of justice the parties to Rawls's model agree to, those principles will specify what society must be like to ensure that no one is unjustly privileged.

We can see this further by examining the two principles that Rawls derives from the original position to define a just society:

> (a) Each person has the same indefeasible claim to a fully adequate scheme of equal basic liberties, which scheme is compatible with the same scheme of liberties for all; and
> (b) Social and economic inequalities are to satisfy two conditions: first, they are to be attached to offices and positions open to all under conditions of fair equality of opportunity; and second, they are to be to the greatest benefit of the least-advantaged members of society (the difference principle).[43]

[38] Ruíz, "Framing Intersectionality," 335.

[39] See, e.g., Anderson, *The Imperative of Integration*; Khader, *Decolonizing Universalism*; and Lépinard, *Feminist Trouble*.

[40] Evans and Lépinard, "Confronting Privileges in Feminist and Queer Movements," 13.

[41] Rawls, *A Theory of Justice*, ch. 3.

[42] Rawls, *A Theory of Justice*, 11 (emphases added).

[43] Rawls, *Justice as Fairness*, 42–43.

Rawls's first principle—the equal basic liberty principle—entitles all members of society to equal legal protections of various liberties, including rights to free speech, freedom of association, freedom to run for political office, to vote, and so on. This principle also entitles everyone to the *fair value* of political liberties (the right to vote and run for political office), such that these liberties must have the same usefulness for each person.[44] Rawls's second principle then has two parts. Its first part, Rawls's fair equality of opportunity (FEO) principle, holds that "in all parts of society there are to be roughly the same prospects of culture and achievement for those similarly motivated and endowed."[45] Its second part, the difference principle, then holds that all other social and economic inequalities (principally, income and wealth) must be to the maximum advantage of society's least-advantaged class.[46] Finally, Rawls's first principle has lexical priority over the second principle, such that inequalities of basic liberties cannot be justified by greater adherence to the second principle.[47] Similarly, the FEO principle has lexical priority over the difference principle.[48]

Bearing this and the role that Rawls argued that these principles should play as ideals in mind—as measures of how just a society is to guide social reform—let us return to intersectionality.[49] Notice that Rawls's principles provide clear grounds for determining which groups are unjustly privileged and to what extent and, conversely, which groups suffer which intersectional injustices and the relative severity thereof. This is not to say that Rawls's principles are the only or best way to recognize intersecting axes of privilege and oppression—as identifying oppression often comes not from theory but from those who experience it directly.[50] It is merely to say that Rawlsian ideal theory can help us understand how forms of intersectional oppression are *also* violations of liberal requirements of fairness.

For, consider Rawls's equal basic liberty principle, which again holds that a just society would ensure that everyone enjoys equal basic rights and liberties and fair value of political liberties. This principle is nowhere close to satisfied

[44] Rawls, *Justice as Fairness*, 46–53. See also Rawls, *A Theory of Justice*, 197, and *Political Liberalism*, 358. Cf. Krishnamurthy, "Completing Rawls's Arguments for Equal Political Liberty and Its Fair Value" and "Reconceiving Rawls's Arguments for Equal Political Liberty and Its Fair Value."

[45] Rawls, *Justice as Fairness*, 44.

[46] Rawls, *A Theory of Justice*, 65–68.

[47] Rawls, *A Theory of Justice*, 132, 175, 220–24.

[48] Rawls, *A Theory of Justice*, 266.

[49] Rawls, *A Theory of Justice*, 215–16.

[50] I thank Laura Wildemann Kane for encouraging me to highlight this.

in the United States.[51] First, voting suppression and stark racial, gender, and socioeconomic inequalities in attaining political office violate the fair value of political liberties.[52] Second, the war on drugs and curtailment of the basic liberties of Black people and other persons of color—including but not limited to racial profiling, pretext stops, and mass incarceration—show that these groups do not enjoy the same basic liberties to drive or walk down the street as more privileged groups.[53] Third, racial disparities in sentencing and false convictions indicate that Black Americans do not enjoy equal protections of basic liberties in courts of law.[54] Third, gays, lesbians, queer, and trans folk live in daily fear of violence against them on the basis of their identities and are underrepresented in political offices.[55] Fourth, gender and sexuality disparities in sexual violence and the need for the #MeToo movement to hold perpetrators accountable indicate that women and LGBTQIA+ groups have not enjoyed the same protections of basic liberties to be free from sexual violence as men of privileged identities.[56]

Rawls's first principle not only recognizes these as injustices: it supports *intersectional analyses* of them. For example, Black men are profiled, arrested, and imprisoned at vastly higher rates than other groups.[57] Similarly, although LGBTQIA+ folk face unjust violence, empirical studies show that different intersectional groups face different kinds and levels of it—with, for example, lesbians facing the highest levels of lifetime sexual-assault victimization but gay men the highest levels of childhood sexual assault.[58] Rawls's equal basic liberty principle supports recognizing these *as* intersectional oppressions—as different ways that persons of different intersecting identities are denied equal protections of basic rights and liberties.

[51] Arvan, "Nonideal Justice as Nonideal Fairness," 211.

[52] Bentele and O'Brien, "Jim Crow 2.0?"; Schoen and Dzhanova, "These Two Charts Show the Lack of Diversity in the House and Senate"; and Zippia, "President Demographics and Statistics in the US."

[53] Alexander, *The New Jim Crow*.

[54] Schmitt, Reedt, and Blackwell, "Demographic Differences in Sentencing"; and Stevens, "Race and Wrongful Convictions."

[55] Dinno, "Homicide Rates of Transgender Individuals in the United States"; and Rothman, Exner, and Baughman, "The Prevalence of Sexual Assault against People Who Identify as Gay, Lesbian, or Bisexual in the United States."

[56] Coulter et al., "Prevalence of Past-Year Sexual Assault Victimization among Undergraduate Students."

[57] Tucker, "The Color of Mass Incarceration."

[58] Rothman, Exner, and Baughman, "The Prevalence of Sexual Assault against People Who Identify as Gay, Lesbian, or Bisexual in the United States."

Rawls's second principle also supports intersectional analyses. Rawls's FEO principle again holds that "in all parts of society there are to be roughly the same prospects of culture and achievement for those similarly motivated and endowed."[59] As Leif Wenar explains, this means that

> within any type of occupation (generally specified) we should find that roughly one quarter of people in that occupation were born into the top 25% of the income distribution, one quarter were born into the second-highest 25% of the income distribution, one quarter were born into the second-lowest 25%, and one-quarter were born into the lowest 25%.[60]

However, the FEO principle does not merely apply to socioeconomic class: it applies to *all* social categories, holding, for example, that justice requires society "to insure that the life prospects of racial minorities are not negatively impacted by the economic legacy of racial oppression."[61] Since, for example, Black trans women are underrepresented in positions of corporate leadership, Rawls's FEO principle identifies this as a distinct form of oppression. Similarly, Rawls's difference principle entails that if members of some intersecting social identities are disproportionally disadvantaged by unjust economic inequality (as indeed they are), then these too are forms of intersectional socioeconomic oppression.[62]

Rawls's principles also provide an attractive normative framework for comparing different forms of oppression. For, although all forms of oppression are unjust, Rawls' theory holds that protecting the fair value of equal basic liberties is lexically more important than fair equality of opportunity and economic injustice.[63] This means, for example, that even if we grant that the US is economically unjust, Rawls's ideal theory entails that rectifying deprivations of equal basic liberties should be our highest priority, inequalities of opportunity our second-highest priority, and economic justice our third-highest priority.[64] Yet this coheres with feminism and CRT, which generally recognize that intersectionality requires prioritizing the most oppressed.[65]

[59] Rawls, *Justice as Fairness*, 44.

[60] Wenar, "John Rawls," sec. 4.3.

[61] Shelby, "Race and Social Justice," 1711, sec. 5. As Matthew notes, the FEO Principle "does not differentiate between disadvantages based on their source" ("Rawls and Racial Justice," 247).

[62] Michener and Brower, "What's Policy Got to Do with It?"; cf. Piketty and Saez, "The Evolution of Top Incomes."

[63] Rawls, *A Theory of Justice*, 53–54.

[64] Rawls, *A Theory of Justice*, 266–67.

[65] Disch and Hawkesworth, *The Oxford Handbook of Feminist Theory*, 1.

It is important to dispel here a common misconception about how Rawlsian liberalism understands society's least-advantaged group. Although Rawls's difference principle understands society's least advantaged in purely economic terms, this is merely how Rawls understands the least advantaged in ideal theory—as Rawls takes it for granted that everyone in a just society would enjoy equal basic liberties and fair equality of opportunity.[66] In contrast, in nonideal theory, Rawls holds that "we have a natural duty to remove any injustices, beginning with the most grievous as identified by the extent of the deviation from perfect justice."[67] This means that under unjust conditions, the most disadvantaged in Rawlsian liberalism are those who are denied *equal basic liberties* and are *multiply marginalized* (suffering, additionally, the worst forms of unfair equality of opportunity and socioeconomic injustice), namely, BIPOC and LGBTQIA+ groups. Yet it is widely recognized that justice requires prioritizing the most marginalized as such.[68] Consequently, Rawls's ideal theory provides an attractive liberal framework for understanding the nature and comparative severity of different intersecting forms of oppression.

We can also see here that another complaint about Rawls's ideal theory is mistaken. Echoing Onora O'Neill's complaint that ideal theories are a "grotesque parody" of the way the world is, Michael Goodhart writes:

> Conceiving of injustice as the absence or opposite of justice renders distant, static, or cerebral something that many people experience as immediate, dynamic, and visceral.... Theorizing injustice as an aberration or departure from ideal justice fundamentally mischaracterizes people's sense and experience of injustice and misses or misapprehends its political character and significance.[69]

However, our discussion suggests that Mariame Kaba has a more accurate take when writing:

> Let's begin our abolitionist journey not with the question, "What do we have now, and how can we make it better?" Instead, let's ask, "What can we imagine for ourselves and the world?" If we do that, then boundless possibilities of a more just world await us.[70]

[66] Wolff, "Equality"; and Rawls, *A Theory of Justice*, 4–5, 215–16.

[67] Rawls, *A Theory of Justice*, 216.

[68] Táíwò, "Being-in-the-Room Privilege."

[69] Goodhart, *Injustice*, 28. See also O'Neill, *Bounds of Justice*, 180.

[70] Kaba, *We Do This 'Til We Free Us*, 3.

It is an open question (well worth investigating) whether Rawlsian liberalism might support Kaba's prison and police abolitionism.[71] If abolitionism is indeed necessary for ensuring equal basic liberties, then under Rawls's first principle of justice, *liberalism* requires it. But, although we cannot resolve this here, the point is that Rawls's rationale for ideal theory coheres with Kaba's advocacy for locating abolitionist activism in utopian imaginary thought. In "Justice: A Short Story," Kaba imagines a planet without police or prisons, "Small Place," that is visited by an astonished "Earth visitor."[72] This story conveys—in a vivid, visceral, and systematic way—Kaba's vision of the vast gap between our world and a *just* world: an ideal world to realize *through* abolitionist activism. But this is directly analogous to Rawls's rationale for beginning with ideal theory:

> Obviously the problems of ... [nonideal theory] are *the pressing and urgent matters*. These are the things that we are faced with in everyday life. The reason for beginning with ideal theory is that it provides ... the only basis for the *systematic* grasp of these more pressing problems.[73]

Indeed, if we want to know *how far we need to go* to achieve true justice, and if we want to know who has been oppressed, how badly, and who is unjustly privileged and to what extent—things that intersectionality's critics allege that its "murkiness" is ill-suited to do—then Rawls's ideal theory provides a clear, principled, and attractive framework for doing so.[74] According to Rawlsian liberalism, insofar as Black males are killed, arrested, and imprisoned at astonishingly disproportionate rates (depriving them of their basic liberties), Black males endure some of the worst, most systemic, and long-lasting injustices of any social group. In addition, insofar as Black Americans face some of the most serious deprivations of health care and worst health-related mortality rates, insofar as Black women and Indigenous groups face uniquely serious health care disparities, and insofar as access to health care is increasingly recognized as a basic liberty, Rawlsian liberalism entails that these intersectional oppressions should be among our highest priorities to rectify as well.[75] Although Rawlsian

[71] N.b.: in what follows (and more broadly), I have sought to avoid unsound epistemic practices of reductive inclusion, i.e., interpolation and ossification, as defined by Dotson and Spencer, "Another Letter Long Delayed."

[72] Kaba, *We Do This 'Til We Free Us*, 157–63.

[73] Rawls, *A Theory of Justice*, 8 (emphases added).

[74] See Davis, "Intersectionality as Buzzword"; and Nash, "Re-Thinking Intersectionality."

[75] Manuel, "Racial/Ethnic and Gender Disparities in Health Care Use and Access"; Nesbitt and Palomarez, "Increasing Awareness and Education on Health Disparities for Health Care Providers"; and Orgera and Artiga, "Disparities in Health and Health Care."

liberalism explains these injustices in distributive terms, we will see in section 3 that its analysis substantially converges with feminist accounts that understand oppression in terms extending beyond the "distributive paradigm."[76]

Finally, Rawlsian theory also supports intersectionality in nonideal theory. To determine what Rawlsian liberalism requires in unjust conditions, Rawls's original position must be reformulated as a "nonideal original position."[77] Next, as I have argued previously, the parties to this model should seek remedial "nonideal primary goods" that empower oppressed groups and their allies to rectify injustices.[78] These goods include remedial social, political, and economic institutions ranging from the Civil Rights Act to the National Association for the Advancement of Colored People (NAACP), National Labor Relations Act (NLRA), and educational equity reforms (including the Women's Educational Equity Act), as well as grassroots activism that confers compensatory bargaining power on the oppressed and disseminates skills and information for effectively and equitably combating oppression.[79] Yet these too are the kinds of things that feminism and CRT advocate: creating sociopolitical conditions that *center and amplify* the perspectives, voices, knowledge, and interests of intersectionally oppressed groups, particularly the most oppressed.[80]

None of this is to say that Rawlian liberalism should be understood as "the" justification for intersectionality, nor does it imply that a Rawlsian approach to intersectionality should displace distinctly feminist ones (I return to this in section 3). It is merely to say that Rawlsianism is a theoretical and practical *ally* of intersectional feminism.

### *1.2. Standpoint Epistemology, Allyship, and Epistemic Justice as Liberal Requirements of Fairness*

Rawlsian liberalism has been alleged to be problematically ahistorical, abstracting away from historical and present-day injustices and the lived experience of

[76] Young, *Justice and the Politics of Difference*; and Fricker, *Epistemic Injustice*.

[77] Arvan, "First Steps toward a Nonideal Theory of Justice" and "Nonideal Justice as Nonideal Fairness"; Mills, *Black Rights/White Wrongs*, 234; and Adams, "Nonideal Justice, Fairness, and Affirmative Action." See also Arvan, "A Non-Ideal Theory of Justice," ch. 1.

[78] Arvan, "First Steps toward a Nonideal Theory of Justice," 108–14, and "Nonideal Justice as Nonideal Fairness," sec. 3.

[79] Arvan, "Educational Justice and School Boosting," "First Steps toward a Nonideal Theory of Justice," 112, and "Nonideal Justice as Nonideal Fairness," 220–25.

[80] Goodkind and Deacon, "Methodological Issues in Conducting Research with Refugee Women"; and Tuggle, "Towards a Moral Conception of Allyship."

oppressed groups.[81] Following the realization that injustice is intersectional, feminists and CRT argue that it is vital to center marginalized experiences.[82] Specifically, because individuals of different social identities directly experience different forms of oppression on a daily basis—including how elements of society engage in and perpetuate those injustices—there are grounds for thinking that different oppressed groups occupy *privileged epistemic standpoints* on these matters that give their members access to truths that may be deeply obscured from individuals occupying other social categories.[83]

However, are Rawls's critics correct that Rawlsian liberalism problematically abstracts away from lived experience and the epistemic value of intersectional standpoints? Although in ideal theory Rawls reasons abstractly using the original position, in nonideal theory Rawls explicitly focuses on oppressed standpoints: "I have assumed that it is always those with the lesser liberty who must be compensated. We are always to appraise the situation *from their point of view*."[84] Second, while Rawls never developed this much further, Rawlsian nonideal theorists have argued that justice as fairness *does* require centering the lived experiences of the oppressed precisely because of privileged epistemic features rooted in social situatedness.

Specifically, I have argued previously that the parties to a Rawlsian nonideal original position would treat opportunities to be involved in open, inclusive, and equitable grassroots movements in pursuit of just ideals (equal basic liberties, fair equality of opportunity, etc.) as a nonideal primary good for combating injustice.[85] The basic rationale for this is, first, that oppressed individuals living under unjust conditions directly experience the daily costs of injustice, and the parties to a nonideal original position know behind its veil of ignorance that *they* may turn out to be oppressed.[86] Second, because oppressed individuals experience costs of injustice and social reform based on their positionality, the parties to a nonideal original position have grounds to treat the *standpoints* of individuals oppressed by injustice as epistemically privileged with respect

[81] Goodhart, *Injustice*, 28; Kang, "Can Rawls's Nonideal Theory Save His Ideal Theory?"; and Mills, "'Ideal Theory' as Ideology" and *The Racial Contract*. See also Farrelly, "Justice in Ideal Theory." Cf. Erman and Möller, "Three Failed Charges Against Ideal Theory," sec. 4.

[82] Ruíz, "Framing Intersectionality."

[83] Grasswick, "Feminist Social Epistemology," sec. 2.1. Cf. Collins, "Learning from the Outsider Within"; and McKinnon, "Trans*formative Experiences."

[84] Rawls, *A Theory of Justice*, 218 (emphasis added).

[85] Arvan, "First Steps toward a Nonideal Theory of Justice," 108–10.

[86] Arvan, "First Steps toward a Nonideal Theory of Justice," 105, 109–11. See also Arvan, "Nonideal Justice as Nonideal Fairness," sec. 3.

to these phenomena.[87] Third, because oppression comes in degrees and the parties know that they could turn out to be oppressed, Rawlsian nonideal theory supports prioritizing the perspectives, voices, and interests of the *most* oppressed.[88] Yet these conclusions cohere with what feminist perspectives on standpoint epistemology and allyship have long advocated.[89]

Finally, Rawlsian nonideal theory also provides liberal support for feminist insights into epistemic justice.[90] As Byskov details, epistemic injustice comprises five conditions of *unfairness*: a disadvantage condition (unfair outcome), prejudice condition (unfair judgments about epistemic capacities), stakeholder condition (unfair denial of stakeholder rights), epistemic condition (unfair denial of knowledge), and social justice condition (unfair existing vulnerability).[91] Insofar as Rawlsian liberalism holds that justice *is* fairness—and Rawlsian nonideal theory holds that fairness under unjust conditions requires prioritizing rather than denigrating the perspectives, voices, knowledge, and interests of the oppressed—Rawlsian liberalism can help explain why epistemic injustice *is* a form of unfairness that serves to uphold and compound preexisting forms of unjust unfairness (unequal basic liberties, unequal opportunities, and economic injustice) already faced by oppressed groups.

Thus, Rawlsian liberalism not only supports intersectionality: it provides a *liberal* justification for feminist standpoint epistemology, allyship, and epistemic justice.

### 2. Critical Race Theory as a Liberal Requirement of Fairness

Critical race theorists have criticized liberalism for "color-blindness" and "ignoring the problem of intersectionality" and Rawlsian liberalism for "whitewashing" history.[92] Mills, in particular, has argued that liberalism problematically abstracts away from the history of colonialism, slavery, and racial oppression and that Rawls's ideal theory of justice constitutes a problematic ideology that obscures how liberal ideals can support the unjust *status quo*.[93] However, in

[87] Arvan, "A Non-Ideal Theory of Justice," 39–43, 193–99.

[88] Arvan, "First Steps toward a Nonideal Theory of Justice," 115.

[89] Devadson, "Allyship"; Ghabra and Calafell, "From Failure and Allyship to Feminist Solidarities"; Grasswick, "Feminist Social Epistemology"; Táíwò, "Being-in-the-Room Privilege"; and Tuggle, "Towards a Moral Conception of Allyship."

[90] Fricker, *Epistemic Injustice*.

[91] Byskov, "What Makes Epistemic Injustice an 'Injustice'?" esp. 3.

[92] Delgado and Stefancic, *Critical Race Theory*, 26–30, 64; and Mills, "The Whiteness of John Rawls."

[93] Mills, "'Ideal Theory' as Ideology" and *The Racial Contract*.

more recent work, Mills expresses optimism that Rawlsian liberalism can be adapted to correct for these problems.[94] We will now see that he is right and, indeed, that Rawlsian liberalism coheres with the central commitments of CRT as it is understood today.

As with all theoretical frameworks, there may be significant disagreement over exactly what CRT's commitments are, and it has been contended by some proponents that "critical race theory cannot be understood as an abstract set of ideas or principles."[95] At the same time, these proponents have enumerated the following "defining elements" of CRT:

1. Critical race theory recognizes that racism is endemic to American life.
2. Critical race theory expresses skepticism toward dominant legal claims of neutrality, objectivity, color blindness, and meritocracy.
3. Critical race theory challenges ahistoricism and insists upon a contextual/historical analysis of the law.
4. Critical race theory insists on recognition of the experiential knowledge of people of color and our communities of origin in analyzing law and society.
5. Critical race theory is interdisciplinary and eclectic.
6. Critical race theory works toward the end of eliminating racial oppression as part of a broader goal of ending all forms of oppression.[96]

Other "hallmark critical race theory themes" have been claimed by CRT proponents to include:

7. The thesis of interest convergence: that civil rights advances always coincide with and advance the self-interest of whites.[97]
8. Revisionist history: replacing comfortable historical narratives with "ones that square more accurately with minorities' experiences."[98]
9. Structural determinism: the view that structural elements of society (such as legal practice) result in predictable outcomes, such as slowing down social change, imposing costs of progress predominantly (and inequitably) on marginalized races, and upholding white supremacy.[99]

[94] Mills, *Black Rights/White Wrongs*, epilogue (as prologue).

[95] Lawrence et al., "Introduction," 3.

[96] Lawrence et al., "Introduction," 6.

[97] Bell, "Racial Remediation"; and Delgado and Stefancic, *Critical Race Theory*, 18.

[98] Delgado and Stefancic, *Critical Race Theory*, 20.

[99] Delgado and Stefancic, *Critical Race Theory*, 25–32. Cf. Mills, *The Racial Contract*.

We will return to diversity in CRT scholarship and activism in section 3—but for now, let us ask programmatically: Does Rawlsian liberalism support or oppose these defining or hallmark elements of CRT?

Let us begin with 1: whether Rawlsian liberalism recognizes racism as endemic to American life. As illustrated in section 1 and in the work of others applying Rawls to race, Rawls's two principles of justice clearly entail that racism is and always has been endemic to American life. BIPOC groups have *never* enjoyed fully equal basic liberties (viz., Rawls's first principle), fair equality of opportunity (viz., Rawls's FEO principle), or economic justice (viz., Rawls's difference principle).[100] According to Rawls's principles of ideal justice, then, severe racial injustices exist in the US today and have existed throughout America's history.

Now turn to 2: whether Rawlsian liberalism supports or expresses skepticism toward dominant legal claims of neutrality, objectivity, color-blindness, and meritocracy. It is often claimed that Rawls's original position problematically supports these dominant claims.[101] After all, the original position is supposed to be a neutral, "color-blind" procedure that does not permit anyone to take their race into account when deliberating about principles of justice.[102] However, we need to be careful here. First, although in *A Theory of Justice* Rawls did present the original position as an "objective" model of justice that may also be understood as an interpretation of Kant's (objective and ahistorical) moral principle "the categorical imperative," Rawls also held that the original position represents our considered judgments *here and now* in the real world.[103] Second, in his later work, Rawls firmly rejected the Kantian/objective grounding of justice as fairness, instead defending it as a political doctrine grounded in an *overlapping consensus*—or shared values—of citizens living under particular historical conditions: specifically, pluralist modern democracies characterized by diversity of thought and values.[104] Rawls then claims that justice as fairness approximates such a consensus *reasonably well*, providing a conception of justice "for a constitutional democracy" that "will seem reasonable and useful, even if not fully convincing, to a wide range of thoughtful political opinions … [that] express an essential part of the common core of the democratic

[100] Shelby, "Race and Social Justice," esp. 1700. See also Arvan, "Educational Justice and School Boosting," 3–11, and "Nonideal Justice as Nonideal Fairness," 211; and Matthew, "Rawlsian Affirmative Action" and "Rawls and Racial Justice."

[101] Foster, "Rawls, Race, and Reason"; Oktay, *Color-Blindness in Rawls's Theory of Justice*. See also Barndt, "The God Trick."

[102] Rawls, *A Theory of Justice*, sec. 4.

[103] Rawls, *A Theory of Justice*, sec. 40, sec. 78, pp. 18–19, 42–45, 507–8.

[104] Rawls, *Political Liberalism*, 3–4, 13–15, and secs. 1–3.

tradition."[105] Finally, intersectional feminism and CRT actually appear to share the values that the original position models. Intersectional feminism and CRT both standardly understand justice as requiring *equity*—that is, the dismantling of unfair privileges.[106] Yet Rawls's principles clearly entail that White privilege, heterosexual cis-male privilege, ableism, and so on, *are* unjust privileges, just as intersectional feminism and CRT hold. We can begin to see how by carefully examining what Rawls's original position represents and what its output principles require.

Consider 6: whether justice as fairness works toward eliminating racial oppression as part of a broader goal of ending all forms of oppression. Rawls's original position is supposed to represent the common convictions of people who are committed to fairness: specifically, the convictions that a fully just society would be one in which no one is unfairly privileged based on social identity.[107] Yet this is what feminism and CRT seek: an *end* to White privilege, cis-male privilege, and so on. Second, the original position's output principles require society to be *equitable*, as they hold that members of all races, genders, and so on should enjoy the fair value of basic political rights, fair equality of opportunity, and a fair distribution of wealth, such that again no one is unfairly privileged. Yet equity as such is precisely what CRT espouses.[108] Third, as we will see in section 3, Rawls's just society would not plausibly contain any of Young's "five faces of oppression"—and so would realize sociopolitical conditions where domination and oppression no longer exist. Fourth, these are merely Rawlsian liberalism's implications within ideal theory. Recent extensions of justice as fairness to nonideal theory—that is, to the unjust world in which we live—reveal that rather than supporting "neutrality" or "color-blindness," Rawlsian liberalism supports *compensatory* forms of equity, including remedial legal, political, and economic goods such as special legal rights and programs that prioritize the voices, perspectives, knowledge, and interests of the oppressed, both domestically and globally.[109] Finally, far from supporting "meritocracy," Rawlsian liberalism supports compensatory institutions to ensure equity, such as affirmative action and (potentially) rectification of historical injustices, such

[105] Rawls, *A Theory of Justice*, xi.

[106] Almeder, "Equity Feminism and Academic Feminism"; and Crenshaw et al., *Critical Race Theory*.

[107] Matthew, "Rawls and Racial Justice" and "Rawls's Ideal Theory."

[108] Ford and Airhihenbuwa, "Critical Race Theory, Race Equity, and Public Health"; and Racial Equity Tools, "Fundamentals."

[109] Arvan, "A Non-Ideal Theory of Justice," chs. 3–5, "Educational Justice and School Boosting," "First Steps toward a Nonideal Theory of Justice" 108–15, and "Nonideal Justice as Nonideal Fairness," 220–26.

as slavery.[110] Rawlsian liberalism, then, does not reify oppressive conceptions of "neutrality, objectivity, color blindness, and meritocracy." It holds that under unjust conditions, justice requires *non*-neutrality: legal, political, social, and economic goods that prioritize the oppressed over the privileged.

Now turn to 3: whether Rawlsian liberalism challenges ahistoricism and insists upon a contextual/historical analysis of the law. As Mills points out, *A Theory of Justice* does not contain a single reference to American slavery (though it does condemn historical slavery in the abstract).[111] Although abstracting away from American slavery may seem problematic, we should recall Rawls's purpose in providing an "ideal theory" of justice. The purpose is to provide a measure of *injustice*, including an explanation of why historical injustices *are* injustices (e.g., American slavery was unjust because it denied people equal basic liberties).[112] Similarly, it is evident that Rawls's liberal conception of international justice would identify colonialism as a grave historical injustice. First, using an international original position, Rawls derives the principle that "peoples are free and independent, and their freedom and independence is to be respected by other peoples."[113] Second, Rawls defends a minimal list of human rights—including a right against forced occupation—precisely to prevent paternalistic interference in "decent" illiberal societies.[114] Third, although Rawls holds that outsiders do have duties to assist "burdened societies"—particularly societies that cannot satisfy the basic human right of subsistence (i.e., non-starvation) or violate the human rights of women—he is explicit that his "Law of Peoples" does *not* justify outsiders attempting to develop "pastoral" societies economically and that "advice" rather than force or occupation is to be used so as to avoid "improperly undermining a society's religion and

[110] See Adams, "Nonideal Justice, Fairness, and Affirmative Action"; Matthew, "Rawlsian Affirmative Action"; and Meshelski, "Procedural Justice and Affirmative Action." Espindola and Vaca argue that Rawls's theory does not necessarily require historical rectification and propose an amendment to the theory to better account for this moral requirement ("The Problem of Historical Rectification for Rawlsian Theory"). In contrast, I argue that when justice as fairness is extended to nonideal theory properly, it can be seen to require empowering members of oppressed groups to *collectively and equitably decide* whether and to what extent historical rectification should be pursued, given the costs and alternatives available (Arvan, "First Steps toward a Nonideal Theory of Justice," 103, 107–14). This is, in effect, to afford the oppressed a kind of collective right to determine how to balance backward- and forward-looking aspects of justice. See also Arvan, "A Non-Ideal Theory of Justice."

[111] Mills, *The Racial Contract*, 77; and Rawls, *A Theory of Justice*, 217, 248.

[112] Rawls, *A Theory of Justice*, 216, 247

[113] Rawls, *The Law of Peoples*, 37 and sec. 3.

[114] Rawls, *The Law of Peoples*, 65.

culture."[115] Finally, Rawlsian liberalism again requires extending Rawls's original position to nonideal theory—that is, to the conditions we actually live in, given the history and present of oppression, including racial oppression. And here, Rawlsian theory has been argued to support empowering marginalized groups to *collectively and equitably decide* whether and to what extent historical injustices should be rectified (such as reparations) as well as (globally) a higher-order human right to collectively and equitably decide the costs that they should have to face for the promotion of their first-order human rights.[116] Insofar as Rawlsian ideal theory thus identifies colonialism as a grave injustice, and Rawlsian nonideal theory supports equitable grassroots activism to address its legacy, Rawlsian liberalism plausibly supports the general goals of decolonial feminism and CRT.[117] So, Rawlsian liberalism is not problematically ahistorical. It provides a framework for *recognizing and rectifying* historical and present-day oppression.

Now turn to 4. Rawlsian liberalism as developed in nonideal theory wholeheartedly supports CRT's insistence upon the "recognition of the experiential knowledge of people of color and our communities of origin in analyzing law and society." As detailed in section 1, under unjust conditions, Rawlsian liberalism requires inclusively centering the voices, experiences, knowledge, and interests of the oppressed in grassroots deliberation precisely because, as a matter of equity, unjustly oppressed groups are owed compensation, and as a matter of epistemology, oppressed groups directly experience "nonideal costs" that other groups do not.[118]

Now turn to 5. Rawlsian liberalism clearly supports interdisciplinary approaches to examining and dismantling oppression. In ideal theory, Rawls holds that the parties to the original position should be aware of "general facts about human society," including "political affairs … principles of economic theory … [and] the basis of social organization and the laws of human psychology."[119] Rawls holds that these interdisciplinary forms of knowledge are vital for evaluating a theory of justice, writing: "General facts of human psychology and principles of moral learning are relevant.… If a conception of justice is unlikely to generate its own support, or lacks stability, then this fact must not be

[115] Rawls, *The Law of Peoples*, 109–11, 117.

[116] Arvan, "A Non-Ideal Theory of Justice," ch. 3, "First Steps toward a Nonideal Theory of Justice," and "Nonideal Justice as Nonideal Fairness." Cf. Espindola and Vaca, "The Problem of Historical Rectification for Rawlsian Theory."

[117] Lugones, "Toward a Decolonial Feminism."

[118] Arvan, "First Steps toward a Nonideal Theory of Justice," 108–15 and "Nonideal Justice as Nonideal Fairness," 220–23.

[119] Rawls, *A Theory of Justice*, 119.

overlooked."[120] Further, under unjust conditions, Rawlsian liberalism has again been shown to require developing and disseminating all-purpose skills and information related to effective and equitable social organizing, constructing remedial legal, social, political, and economic institutions to combat oppression, rationally understanding the costs and benefits of different policies and tactics, and distributing those costs equitably.[121]

Now turn to 7: the thesis of interest convergence. Here, Rawlsian nonideal theory recognizes a deep tension within what justice requires under unjust conditions. On the one hand, individuals in a nonideal original position have grounds to prioritize the perspectives, voices, and interests of the oppressed, seeking to augment marginalized groups' bargaining power to compensate for oppression.[122] On the other hand, the parties also must take seriously the existence of dominant majorities and the fact that members of those majorities may be strongly inclined to prefer social reform only to the extent that they see reform to be consistent with what they take their "legitimate interests" to be.[123] This suggests that social reform is more likely to occur via *overlapping consensus* between oppressed populations and sympathetic majorities—that is, by interest convergence.[124] Rawlsian nonideal theory thus recognizes not only interest convergence but also the general idea (recognized by CRT) that this is a theoretical and practical problem—namely, how to square the *fact* of interest convergence with the idea that justice requires *the opposite*: prioritizing the oppressed. Further, although this is an area of ongoing research, as we see above, Rawlsian theory supports an answer to this quandary that coheres with the contemporary practice of CRT activism: namely, centering the voices and experiences of (multiply) marginalized groups and utilizing formal and informal "levers of power" to augment their social, economic, and legal bargaining power.[125]

Rawlsian liberalism also supports 8: replacing comfortable historical narratives with ones that reflect marginalized minorities' experiences. First, as established earlier, Rawlsian ideal theory recognizes slavery, racism, sexism, and so on *as* injustices—as serious, unjust deviations from what a fully just society would be. Second, as we have just seen, Rawlsian nonideal theory requires the

[120] Rawls, *A Theory of Justice,* 125.

[121] Arvan, "First Steps toward a Nonideal Theory of Justice," 111–12, and "Nonideal Justice as Nonideal Fairness," 223–25.

[122] Arvan, "Nonideal Justice as Nonideal Fairness," 221.

[123] Arvan, "First Steps toward a Nonideal Theory of Justice," 111, 114.

[124] Arvan, "Nonideal Justice as Nonideal Fairness," 221, 223.

[125] Arvan, "Nonideal Justice as Nonideal Fairness," 223–24, and "First Steps toward a Nonideal Theory of Justice," 106–7.

distribution of all-purpose skills and information for *understanding and combating* injustice and *amplifying the voices and perspectives* of the oppressed due to their direct experience with oppression and "nonideal costs" thereof. Insofar as replacing false historical narratives with narratives that reflect the true history and marginalized experiences of oppression promises to do just this, Rawlsian liberalism supports the practice.

Finally, the same is true of 9. Insofar as Rawlsian nonideal theory supports the pursuit and dissemination of all-purpose knowledge related to understanding injustice and "nonideal costs," Rawlsian liberalism supports *understanding* structural determinism: features of society that justly or unjustly determine social outcomes, such as rights, opportunities, income and wealth, mass incarceration, policing, and so on.

## 3. How Rawlsian Liberalism Supports Diverse Feminist and Critical Race Theory Work to Dismantle All Forms of Oppression

Our examination thus far has been programmatic, showing at a high level of abstraction how Rawlsian liberalism supports central commitments of intersectional feminism and CRT. However, what about the great diversity of work in these fields? Does Rawlsian liberalism support the diverse projects of actual intersectional feminists and critical race theorists? We will now see that it does.

Iris Marion Young's "Five Faces of Oppression" has been widely utilized in feminism and CRT to understand and contest many varieties of oppression. Young argues that "instead of focusing on distribution, a conception of justice should begin with the concepts of domination and oppression."[126] Young then defines five types of domination and oppression:

> *Exploitation*: "This oppression occurs through a steady process of the transfer of the results of labor from one social group to benefit another."[127]

> *Marginalization*: "Marginals are people the system of labor cannot or will not use.… A whole category of people is expelled from useful participation in social life."[128]

> *Powerlessness*: "The powerless are … those over whom power is exercised without their exercising it; the powerless are situated so that they must take orders and rarely have the right to give them."[129]

[126] Young, *Justice and the Politics of Difference*, 3.

[127] Young, *Justice and the Politics of Difference*, 49.

[128] Young, *Justice and the Politics of Difference*, 53.

[129] Young, *Justice and the Politics of Difference*, 56.

> *Cultural Imperialism*: "To experience cultural imperialism means to experience how the dominant meanings of a society render the particular perspective of one's group invisible at the same time as they stereotype one's group and mark it out as the Other."[130]
>
> *Violence*: "Members of some groups live with the knowledge that they must fear random, unprovoked attacks on their persons or property, which have no motive but to damage, humiliate, or destroy the person."[131]

The influence of Young's framework on intersectional feminism and CRT can hardly be overstated. Among other things, it has been used to theorize ableism; ageism; anti-Arab and anti-Black racism; antiracist education; anti-oppressive citizenship education; biphobia; child protection reform; Christian privilege; colonialism; cultural appropriation; data justice; decolonial philosophical writing; educational injustice; fatphobia; food justice; hate speech; interspecies oppression; LGBTQIA+ oppression; medical oppression; anti-oppressive, intersectional, and decolonial pedagogy; oppression resistance through the lens of carceral status; the politics of school violence; representation justice; and vegan ecofeminism.[132]

[130] Young, *Justice and the Politics of Difference*, 58–59.

[131] Young, *Justice and the Politics of Difference*, 61.

[132] Acevedo-Zapata, "Writing as a Decolonial Feminist Praxis for Philosophical Writing"; Albornoz, "Pedagogies and Strategies for an Anti-Oppression Classroom"; Albornoz, Reilly, and Flores, "Community-Based Data Justice"; Barnes and Mercer, *Disability*; Blumenfeld, "Christian Privilege and the Promotion of 'Secular' and Not-So 'Secular' Mainline Christianity in Public Schooling and in the Larger Society"; Braithwhite, "Institutional Oppression that Silences Child Protection Reform"; DeJong and Love, "Youth Oppression and Elder Oppression"; Dubeau, "Species-Being for Whom?"; Fejős and Zentai, *Anti-Gender Hate Speech in Populist Right-Wing Social Media Communication*, 10; Gaard, "Vegetarian Ecofeminism"; Gaynor, "Social Construction and the Criminalization of Identity"; Haeffner at al., "Representation Justice as a Research Agenda for Socio-Hydrology and Water Governance"; Heerten-Rodriguez, "Fat Peoples' Experiences of and Responses to Sexualized Oppression: A Multi-Method Qualitative Study"; Henderson, "The Context of the State of Nature," 30; Higgs and Gilleard, "Is Ageism an Oppression?"; Liao and Carbonell, "Materialized Oppression in Medical Tools and Technologies"; Matthes, "Cultural Appropriation and Oppression"; Mayes and Byrd, "An Antiracist Framework for Evidence-Informed School Counseling Practice," 2; Michaelis, "From Indifference to Injustice"; Northway, "Disability and Oppression"; Obradors-Campos, "Deconstructing Biphobia"; Olding, "Racism and English Language Learning"; Owens, "Excavating Oppression in the Wake of Ferguson, Baltimore, and Municipal Everywhere"; Price, Cruz-Garcia, and Narchi, "Foods of Oppression"; Shlasko, "Using the Five Faces of Oppression to Teach about Interlocking Systems of Oppression"; Vinson, "Oppression, Anti-Oppression, and Citizenship Education"; Wingfield, "Arab Americans"; Woodall and Shannon, "Carceral Citizens Rising"; and Writer, "Unmasking, Exposing, and Confronting."

Young claims that no single form of oppression can be assigned causal or moral primacy and that it is not possible to define any single set of criteria that unify different forms of oppression.[133] Indeed, she challenges the "logic of identity," or attempts to provide "totalizing systems in which ... unifying categories are themselves unified under principles, where the ideal is to reduce everything to one first principle."[134] However, is Young right? Is there nothing that unifies her five faces of oppression? This seems false on its face: exploitation, marginalization, powerlessness, cultural imperialism, and arbitrary group-directed violence are all *unfair*—indeed, profoundly so. But, of course, Rawls contends that justice *is* fairness. So, the question arises: Can we explain Young's five faces of oppression—and justify the many anti-oppressive feminist and CRT projects enumerated above—by reference to liberal demands of fairness, holding that each form of oppression is a violation of this deeper value?

Indeed, we can. For let us ask: Would *any* of Young's five forms of oppression exist in Rawls's just society—that is, in a society in which all citizens have equal basic liberties (including the fair value of political liberties), fair equality of opportunity, and fair economic conditions? The answer is no. To see how, begin with exploitation. Would anyone be exploited in Rawls's just society, where exploitation involves the "steady process of the transfer of the results of labor from one social group to benefit another"? Surely not. After all, Rawls's general conception of justice—which his two principles of ideal theory are an instance of—holds:

> All social values—liberty and opportunity, income and wealth, and the social bases of self-respect—are to be distributed equally unless an unequal distribution of any, or all, of these values *is to everyone's advantage*.[135]

The point of Rawls's principles of ideal justice—the equal basic liberties principle, FEO principle, and difference principle—is to describe conditions that accomplish this, defining a society in which *no one* is exploited and *everyone* benefits fairly from social cooperation.[136] As Rawls puts it, the difference principle does not involve the steady transfer of results of labor from one social group to the benefit of another (which Young takes to comprise exploitation).

[133] Young, *Justice and the Politics of Difference*, 40, 42

[134] Young, *Justice and the Politics of Difference*, 98.

[135] Rawls, *A Theory of Justice*, 54 (emphasis added).

[136] Rawls, *A Theory of Justice*, 12–13.

Rather, the difference principle holds instead that "inequalities of wealth and authority … are just only if they result in *compensating benefits for everyone*."[137]

Now turn to marginalization. Would anyone be marginalized in Rawls's just society? It is hard to see how. Rawls's first principle holds that everyone must enjoy the fair value of political liberties, where this means "fair opportunity to take part in and to influence the political process" such that "those similarly endowed and motivated should have roughly the same chance of attaining positions of political authority irrespective of their economic and social class."[138] Consequently, no one would be *politically* marginalized in Rawls's just society: everyone, regardless of race, gender, sexuality, socioeconomic class, and so on, would have roughly the same chance of influencing political decisions and rising to positions of political authority. Next, Rawls's FEO principle holds that the basic structure of a just society must ensure the fair equality of economic opportunities: that everyone, regardless of race, gender, and so on, should have roughly the same chance of obtaining similar jobs, levels of employment, advancement in employment, and so on. So, no one would be *economically* marginalized in Rawls's just society, either. But this is just to say, on Young's own definition of marginalization, that no one would be marginalized in Rawls's just society *tout court*. For, Young writes: "Marginals are people the system of labor cannot or will not use. … A whole category of people is expelled from useful participation in social life." Clearly, for reasons just described, *no one* would satisfy this definition in Rawls's just society—as everyone in Rawls's just society would have fair access to participation in society's system of labor, and no one would be expelled from effective participation in sociopolitical life.[139]

What about powerlessness? Would any group be powerless in Rawls's just society? No. First, Rawls's first principle requires everyone to enjoy the fair value of political liberties, such that everyone would have roughly equal chances to influence political processes, be elected to political office, and so on. Second, Rawls's FEO principle holds that people of all backgrounds should have roughly equal chances of rising to positions of power and authority in the economic sphere. Finally, Rawls holds that the FEO and difference principles support *property-owning democracy*, a socioeconomic system "ensuring

[137] Rawls, *A Theory of Justice*, 13 (emphasis added).

[138] Rawls, *A Theory of Justice*, 197.

[139] One potential exception to this might be dependents such as children and individuals with disabilities that require dependence on others. As Kittay argues, Rawls makes no clear provision for these anywhere in his theory of justice (*Love's Labor* and "Love's Labor Revisited"). Yet as Bhandary argues, Rawls's theory can be plausibly supplemented to accommodate dependency ("Dependency in Justice").

the widespread ownership of productive assets and human capital (educated abilities and trained skills)."[140] Rawls adds:

> In a property-owning democracy … basic institutions must from the outset put in the hands of citizens generally, *and not only of a few*, the productive means to be fully cooperating members of a society. The emphasis falls on the *steady dispersal over time of the ownership of capital* and resources by the laws of inheritance and bequest, on fair equality of opportunity secured by provisions for education and training.[141]

A property-owning democracy is thus a society in which Young's definition of powerlessness applies to *no one*. In such a society, capital would be widely dispersed so that no one has to work at (say) Amazon or Walmart, taking orders but never giving them. Instead, virtually every citizen in Rawls's just society could feasibly start a small, sustainable business, have a fair chance to influence political processes and be elected to political office, and so on. So, no one would be powerless: everyone would have *fair access* to socio-political-economic power.

What about cultural imperialism? Would anyone in Rawls's just society "experience how the dominant meanings of a society render the particular perspective of one's group invisible at the same time as they stereotype one's group and mark it out as the Other"? Surely not. For, if we take Rawls's equal basic liberties (and the fair value of political liberties) and FEO principles seriously, there would *be* no "dominant group(s)" in Rawls's just society. All groups would be similarly situated with respect to political power and influence and to positions of socioeconomic power and privilege.

Finally, what about arbitrary, group-directed violence? Would a society whose basic structure satisfied Rawls's principles result in arbitrary group-directed violence? It is hard to see how, as Rawls argues that all segments of society would see a society governed by Rawls's principles as a *fair deal*—one that would thereby cultivate a sense of justice and reciprocity among them rather than envy or spite (things that plausibly motivate group-directed violence).[142]

But now if this is right—if Rawls's ideal theory describes conditions in which none of Young's five forms of oppression would exist—then Rawlsian liberalism accomplishes what Young denies is possible: it provides a *unified explanation* that grounds domination and oppression in distributive unfairness.[143]

[140] Rawls, *A Theory of Justice*, xv.

[141] Rawls, *A Theory of Justice*, xv (emphases added).

[142] Rawls, *A Theory of Justice*, 156–57, and ch. 9.

[143] Young, *Justice and the Politics of Difference*, ch. 1.

According to Rawlsian liberalism, the diverse forms of oppression that feminists and CRT contest—ranging from ableism to colonialism, racism, sexism, and beyond—are all unjust because they involve unfair inequalities of basic liberties, opportunities, wealth, and income. Finally, in nonideal theory, Rawlsian liberalism aims to *dismantle* these oppressions, which range from the White supremacist "racial contract" to the patriarchal "sexual contract" and beyond.[144] For in nonideal theory, we again find deep affinities between Rawlsian liberalism, Young's five faces of oppression, and the many diverse projects pursued in feminism and CRT using her framework. First, Young defends "an enabling conception of justice," which holds that in addition to redistributing wealth and power, justice requires a dialogic, communicative ethics that empowers marginalized groups to bring particularities of their experiences of domination and oppression to challenge structural domination and oppression.[145] Second, Young argues that justice thus requires democratizing public life in a way that satisfies a principle of representation that centers marginalized voices and perspectives.[146] Third, Young thus defends "a dual system of rights: a general system of rights which are the same for all, and a more specific system of group-conscious policies and rights."[147] Yet as we have seen, Rawlsian nonideal theory supports all of these conclusions.[148] So, Rawlsian liberalism provides another basis for critiquing precisely what feminism and CRT challenge across a diverse range of projects: the modern-day welfare state founded on histories of ableism, colonialism, sexism, racism, LGBTQIA+-phobia, and so on. Finally, the Rawlsian *value basis* for critiquing and dismantling these oppressions again converges with feminism and CRT: the relevant value being *fairness/equity*.[149]

Critically, none of this implies that Rawlsian liberalism should displace Young's framework or the diverse feminist and CRT work utilizing it. First, different approaches to political philosophy approach justice from fruitfully different starting-points.[150] Whereas Rawls's method of reflective equilibrium aims to bring our considered judgments about justice into greater coherence in pursuit of overlapping consensus, feminism takes women, gender,

[144] Mills, *The Racial Contract*; Pateman, *The Sexual Contract*; and Pateman and Mills, *Contract and Domination.*

[145] Young, *Justice and the Politics of Difference*, 39, 106–9.

[146] Young, *Justice and the Politics of Difference*, 184

[147] Young, *Justice and the Politics of Difference*, 174.

[148] Arvan, "First Steps toward a Nonideal Theory of Justice" and "Nonideal Justice as Nonideal Fairness."

[149] Rawls, *Political Liberalism*, 28, and *A Theory of Justice*, 16.

[150] Floyd, "Political Philosophy's Methodological Moment and the Rise of Public Political Philosophy."

consciousness-raising, and advocacy as foci of analyses, and critical theory examines how laws and other features of society uphold an unjust (e.g., racist) *status quo*.[151] Second, if Rawlsian liberalism is correct, then all of Young's five forms of oppression really *are* injustices: they are different forms of sociopolitical unfairness. Third, insofar as many decades of feminist and CRT projects and activism have analyzed, deconstructed, and developed useful discourses and strategies for dismantling various oppressions—something that, to be clear, Rawlsian liberalism has mostly not done (as Mills is correct that most Rawlsian work has been in ideal theory)—this paper's argument shows that Rawlsian liberalism, intersectional feminism, and CRT *complement* each other. On the one hand, Rawlsian liberalism has much to offer feminism and CRT: a distinctly liberal analysis of oppression as unfairness and liberal justification for diverse feminist and CRT projects. On the other, feminism and CRT have much to offer Rawlsian liberals: decades of painstaking, ongoing theoretical and activist work identifying, deconstructing, and dismantling unfair oppressions.

Finally, there are empirical reasons to think that allying feminism, CRT, and liberalism as such may be of great practical importance. First, empirical psychological findings indicate that violations of what people *perceive* to be requirements of fairness motivate people to engage in punishment and retaliation.[152] Conversely, procedural fairness is known to foster cooperativeness.[153] Third, these findings appear to generalize to other primates.[154] This suggests that to effectively dismantle injustice, we should do so in ways perceived to be fair. Yet opponents of feminism and CRT appear to be increasingly successful in casting them as illiberal and *unfair*.[155] Opposition to CRT appears to have been instrumental to the Republican candidate winning the 2021 election for governor of Virginia and to have roughly tripled local school board recalls.[156] While I am not so naive to suggest that using Rawlsian liberalism to support feminism and

[151] Ansell, "Critical Race Theory"; Cook and Fonow, "Knowledge and Women's Interests"; Geuss, *The Idea of a Critical Theory*; Horkheimer, *Critical Theory*; Rawls, *A Theory of Justice*, 18–19, 507–8, and *Political Liberalism*, lecture 5; and Young, *Justice and the Politics of Difference*, 5–8

[152] FeldmanHall et al, "Fairness Violations Elicit Greater Punishment on Behalf of Another than for Oneself." Cf. Blancero, Johnson, and Lakshman, "Psychological Contracts and Fairness"; and Mendoza, Lane, and Amodio, "For Members Only."

[153] De Cremer and Van Knippenberg, "How Do Leaders Promote Cooperation"; and De Cremer and Tyler, "The Effects of Trust in Authority and Procedural Fairness on Cooperation."

[154] Yamamoto and Takimoto, "Empathy and Fairness."

[155] *Economist*, "The Threat from the Illiberal Left"; Kapoor, "Feminism Is Illiberal"; and Powers, "Illiberal Feminism Is Running Amok."

[156] *Ballotpedia*, "School Board Recalls"; and Smith, "How Did Republicans Turn Critical Race Theory into a Winning Electoral Issue?"

CRT would eliminate these counterreactive forces, there are several grounds to think that it may help. First, liberalism is clearly a dominant ideology in Western culture. Second, as just noted, perceived fairness motivates cooperation and perceived unfairness provokes resistance. Third, although many feminist and critical race theorists explicitly invoke the language and resources of postmodernism and Marxism, critics of feminism and CRT routinely (and seemingly effectively) depict feminism and CRT as illiberal and "un-American" on these grounds.[157] Fourth, persistent divisions on the political left—often derisively referred to as a "circular firing squad"—plausibly undermine broad-based solidarity necessary for more effectively dismantling oppression and combating right-wing resistance.[158] Consequently, there are empirical grounds to believe that if we want to realize a more just world—rather than perpetuate counterproductive division and retaliation—the best way to do so may be to show how feminism and CRT are genuinely liberal, seeking a *fair and equitable* world in a *fair and equitable way*.

## 4. REPLIES TO FIVE OBJECTIONS

I foresee at least five related objections. First, my argument might seem problematically *post hoc*, at most showing how Rawlsian liberalism can "support" feminism and CRT long after the many insights of these fields have been developed by marginalized thinkers and activists. Second, this article might be thought to engage in epistemic appropriation, unjustly detaching epistemic resources developed by marginalized knowers in ways that benefit the powerful—in this case, Rawlsian liberals.[159] Third, my argument might be said to constitute a failure of allyship, as allies to marginalized groups have duties to "decenter" their own voices and perspectives, using their positional privilege instead to amplify marginalized voices.[160] Fourth, this paper might be said to constitute an unjust form of "speaking for" marginalized individuals and groups.[161] Finally, my argument might be claimed to be yet another example of CRT's gentrification, where

[157] Crenshaw, "Beyond Racism and Misogyny," 114; Juan, "From Race to Class Struggle"; Leeb, "Marx and the Gendered Structure of Capitalism"; Leonardo, "The Race for Class"; Young, *Justice and the Politics of Difference*, 8, 20–21, 35–36, 48–53, 98–99, 124–30, 219. Cf. Carter, "Expert Says CRT Designed to Undermine American Values"; and Seymour, "How Postmodernism Became the Universal Scapegoat of the Era."

[158] Pengelly, "Barack Obama Warns Progressives to Avoid 'Circular Firing Squad'"; and Scholz, *Political Solidarity*.

[159] Davis, "On Epistemic Appropriation."

[160] Edwards, "Aspiring Social Justice Ally Identity Development."

[161] Trebilcot, "Dyke Methods." See also Minh-ha, *Woman, Native, Other.*

CRT is "watered down" by the "readiness of white liberals to tout themselves and their scholarship as 'off-label' uses of CRT methodology."[162]

These are all very serious concerns—and if I have erred in any (or all) of these ways, then I accept the responsibility thereunto. However, any work in moral and political philosophy takes moral risks, and I have chosen to hazard these risks because I sincerely believe that it may be of real practical importance—indeed, important to realizing justice—to understand the extent to which feminism, CRT, and Rawlsian liberalism converge.[163] As Alcoff argues:

> The source of a claim or discursive practice in suspect motives or maneuvers or in privileged social locations … though it is always relevant, cannot be sufficient to repudiate it. We must ask further questions about its effects, questions which amount to the following: *will it enable the empowerment of oppressed peoples?*[164]

I have written this paper not merely because I believe its argument is sound but because I believe that greater solidarity between feminism, CRT, and liberalism may be necessary for better empowering oppressed peoples and combating injustice. First, there is ample empirical evidence that when in-group or out-group members are thought to violate a particular group's norms—such as feminists and critical race theorists denying liberal norms or liberals denying feminist and CRT norms—it tends to activate the fight-flight-freeze system, generating anger, confrontation, and exclusion.[165] Second, while righteous anger plausibly has legitimate purposes in justice activism, there are also grounds to think that anger *toward* feminism and CRT can significantly set back their causes.[166] For again, one common type of rhetoric used to vilify contemporary feminism and CRT is that they are "illiberal."[167] This rhetoric plausibly affects how many citizens view feminism and CRT, as well as how they vote—it appears to have swung recent elections in favor of Republicans.[168] These phenomena thus

[162] Curry, "Racism and the Equality Delusion."

[163] Brennan and Freiman, "Moral Philosophy's Moral Risk."

[164] Alcoff, "The Problem of Speaking for Others," 29 (emphasis added). Cf. Kendi, *How to Be an Antiracist*. Kendi argues that antiracist activism demands an "outcome advocacy" that puts "equitable outcomes before our guilt and anguish" (210).

[165] Ditrich et al., "You Gotta Fight!"

[166] Cherry, *The Case for Rage*.

[167] Kapoor, "Feminism Is Illiberal"; Powers, "Illiberal Feminism Is Running Amok"; Sullivan, "Removing the Bedrock of Liberalism."

[168] Smith, "How Did Republicans Turn Critical Race Theory into a Winning Electoral Issue?" Cf. Conway, Repke, and Houck, "Donald Trump as a Cultural Revolt against Perceived Communication Restriction"; and O'Hagan, "The 'Anti-Woke' Backlash Is No Joke."

plausibly stand in the way of feminist and CRT goals: eliminating all forms of oppression. Consequently, if our concern is to realize justice, we should combat these counterreactionary forces effectively rather than poorly. The question then is: What *is* the most effective way to combat reactionary right-wing politics and advance the goals of intersectional feminism and CRT? Tommy J. Curry suggests the "militant and revolutionary strategies of Black radicals" and the "praxis of struggle against systems of racist and neo-colonial oppression."[169] Yet some recent findings suggest that militant methods may have the unintended consequence of driving more people to favor White right-wing nationalism.[170] Further, Derrick Bell Jr. (whose work Curry rightly demands serious engagement with) writes that because racial progress only tends to occur when it advances the interests of the White majority,

> the *harsh and perhaps unsettling truths* in those historically enlightened lessons should become essential elements in racial remediation plans and policies for they reveal clearly: … [among several other things Bell lists] the necessity of remediation strategies that are *pragmatic and flexible. Undue commitment to ideology*, whether integration or separation, direct action or emigration, serve better individual actors rather than those for whom they claim to act.[171]

None of this is to say that revolutionary Black radicalism should be dismissed or denigrated. On the contrary, those of us concerned with justice should presumably utilize every potentially useful tool in our arsenal. My argument is merely that there are reasons to think that the central goals and commitments of intersectional feminism, CRT, and Rawlsian liberalism largely converge and that expounding upon this may be of real practical importance in advancing justice. For if, as I have argued, we can make a convincing case that *liberals* should support central insights from feminism and CRT, then we may have a better chance of overcoming harmful (and incorrect) narratives opposing feminism and CRT, which could gain more self-professed liberals as allies. While we should take seriously the concern that "broadening the progressive tent" in this way could amount to a kind of gentrification, we should also be open to the possibility that it might be a particularly effective way to advance the cause of justice—especially given the empirical findings on human motivation discussed above.

[169] Curry, "Racism and the Equality Delusion."

[170] Simpson, Willer, and Feinberg, "Does Violent Protest Backfire?"

[171] Bell, "Racial Remediation," 28 (emphases added).

Finally, I am optimistic that, so understood, this project has not engaged in harmful or unjust forms of epistemic appropriation, failure of allyship, or speaking for others. First, this article has supported the insights of marginalized scholars and activists, which is different than appropriating them. Emmalon Davis defines unjust epistemic appropriation as occurring when:

1. Epistemic resources developed within the margins are *overtly detached* from the marginalized knowers responsible for their production; and
2. Utilized in dominant discourses in ways that *disproportionately benefit the powerful*.[172]

This paper has done neither. First, I have presented insights from intersectional feminism and CRT to *be* the achievements of those fields. Until recently, Rawlsian scholarship focused primarily on ideal theory, neglecting injustice. These were real failures of Rawlsian liberalism, and feminism and CRT played critical roles in revealing them to be serious failures. Second, however, these critiques appear to be precisely what led Rawlsian liberalism to focus more on nonideal theory: that is, on the realities of oppression. Rawlsian liberals have thus listened to and learned from feminism and CRT—which is a good thing: a sign of progress. Third, Rawlsian nonideal theorists have theorized in ways that aim to *benefit the marginalized*, not the powerful (e.g., by supporting feminist and CRT insights in theory and activism). Fourth, many Rawlsian scholars who have engaged in these projects are themselves marginalized knowers arguing that Rawlsian liberalism has much of value to offer in the pursuit of racial and gender justice.[173]

Similar considerations, I believe, relate to questions of allyship and "speaking for." Although this paper has in one obvious sense inserted "dominant" voices and perspectives into the picture (e.g., Rawlsian liberalism), it has aimed to use this position of power and privilege to *advance the insights and voices* of the historically and presently marginalized, which is what proper allyship is generally argued to involve. As Alcoff argues, sound allyship cannot plausibly involve staying silent or abandoning one's position of privilege (the latter of which is impossible in a world with structural injustice). Instead, power and

[172] Davis, "On Epistemic Appropriation," 702 (emphases added).

[173] Espindola and Vaca, "The Problem of Historical Rectification for Rawlsian Theory"; Krishnamurthy, "Completing Rawls's Arguments for Equal Political Liberty and Its Fair Value" and "Reconceiving Rawls's Arguments for Equal Political Liberty and Its Fair Value"; Matthew, "Rawlsian Affirmative Action," "Rawls and Racial Justice," and "Rawls's Ideal Theory"; Mills, *Black Rights/White Wrongs*, ch. 9, epilogue, and "Occupy Liberalism!"; and Shelby, *Dark Ghettos*, "Justice, Deviance, and the Dark Ghetto," and "Race and Social Justice."

privilege (including dominant ideologies, such as liberalism) can be powerful tools for advancing the cause of justice, at least if used in the right way. For example, Audre Lorde is rightly lauded for affirming that "the master's tools will never dismantle the master's house"—and there is doubtless an important insight here: namely, that a master's tools alone will never dismantle a master's house.[174] To fully dismantle a master's house, their slaves *must be liberated*. Still, as Mills argues, we should be careful not to take Lorde's point further than its weight can bear:

> Imagine we're a group of escaped slaves who have begun by dismantling the master (presumably using our own tools) and now wish to move on to his house. Hunting around the plantation, we come across a toolshed of hammers, pickaxes, saws, barrels of gunpowder, and so forth. Cannot we take these tools and—hammering, digging, sawing in half, blowing up—demolish the master's house? Of course we can—you just watch.[175]

Indeed, depending on the other tools that are available, it may well be a mistake not to appropriate at least some of the master's tools. This has been this paper's aim. If I am correct, Rawlsian liberals should support feminism and CRT as genuine allies in pursuit of justice: not by supplanting marginalized voices, perspectives, knowledge, or theories but by providing them distinctly *liberal* support in pursuit of a common, just cause: eliminating all forms of oppression. Used in this way, Rawlsian liberalism can be a good tool indeed.[176]

*University of Tampa*
*marvan@ut.edu*

## REFERENCES

Acevedo-Zapata, Diana María. "Writing as a Decolonial Feminist Praxis for Philosophical Writing." *Hypatia* 35, no. 3 (Summer 2020): 410–23.

Adams, Matthew. "Nonideal Justice, Fairness, and Affirmative Action." *Journal of Ethics and Social Philosophy* 20, no. 3 (November 2021): 310–42.

[174] Lorde, "The Master's Tools Will Never Dismantle the Master's House."

[175] Mills, "Rousseau, the Master's Tools, and Anti-Contractarian Contractarianism," 93.

[176] I thank Laura Wildemann Kane, the editors of the *Journal of Ethics and Social Philosophy*, several anonymous reviewers, and audiences at the Association for Social and Political Philosophy "Crises of Liberalism?" workshop and After Justice: John Rawls' Legacy in the 21st Century conference.

Albornoz, Carolina. “Pedagogies and Strategies for an Anti-Oppression Classroom.” Master’s thesis, University of Alberta, 2022. https://era.library.ualberta.ca/items/354d723a-11ec-47cc-9615-84f8841cbd5c.

Albornoz, Denisse, Katherine Reilly, and Marieliv Flores. “Community-Based Data Justice: A Model for Data Collection in Informal Urban Settlements.” Development Informatics Working Paper Series, no. 82, Global Development Institute, Manchester, October 2019. https://hummedia.manchester.ac.uk/institutes/gdi/publications/workingpapers/di/di_wp82.pdf.

Alcoff, Linda. “The Problem of Speaking for Others.” *Cultural Critique* 20 (Winter 1991–1992): 5–32.

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

Almeder, Robert F. “Equity Feminism and Academic Feminism.” In *Scrutinizing Feminist Epistemology: An Examination of Gender in Science*, edited by Cassandra L. Pinnick, Noretta Koertge, and Robert F. Almeder, 183–201. New Brunswick, NJ: Rutgers University Press, 2003.

Anderson, Elizabeth. *The Imperative of Integration*. Princeton, NJ: Princeton University Press, 2010.

Ansell, Amy E. “Critical Race Theory.” In *Encyclopedia of Race, Ethnicity, and Society*, vol. 1, edited by Richard T. Schaefer, 344–46. Los Angeles: SAGE Publications, 2008.

Arvan, Marcus. “Educational Justice and School Boosting.” *Social Theory and Practice*, forthcoming. Available at https://philpapers.org/rec/ARVEJA.

———. “First Steps Toward a Nonideal Theory of Justice.” *Ethics and Global Politics* 7, no. 3 (September 2014): 95–117.

———. “Nonideal Justice as Nonideal Fairness.” *Journal of the American Philosophical Association* 5, no. 2 (Summer 2019): 208–28.

———. “A Non-Ideal Theory of Justice.” PhD diss., University of Arizona, 2008. ProQuest (3325309).

Ballotpedia. “School Board Recalls.” Accessed September 29, 2022. https://ballotpedia.org/School_board_recalls#2021.

Barndt, Susan McWilliams. “The God Trick: ‘In the Shadow of Justice.’” *Commonweal*, May 11, 2020. https://www.commonwealmagazine.org/god-trick.

Barnes Colin, and Geof Mercer. *Disability.* Cambridge: Polity Press, 2003.

Bell, Derrick A., Jr. “Racial Remediation: An Historical Perspective on Current Conditions.” *Notre Dame Law Review* 52, no. 1 (October 1976): 5–29.

Bentele, Keith G., and Erin E. O’Brien. “Jim Crow 2.0? Why States Consider and Adopt Restrictive Voter Access Policies.” *Perspectives on Politics* 11, no. 4 (December 2013): 1088–116.

Bhandary, Asha. “Dependency in Justice: Can Rawlsian Liberalism

Accommodate Kittay's Dependency Critique?" *Hypatia* 25, no. 1 (Winter 2010): 140–56.

Blancero, Donna, Scott A. Johnson, and C. Lakshman. "Psychological Contracts and Fairness: The Effect of Violations on Customer Service Behavior." *Journal of Market-Focused Management* 1, no. 1 (March 1996): 49–63.

Blake, Matthew, and Patrick Taylor Smith. "International Distributive Justice." In *Stanford Encyclopedia of Philosophy* (Summer 2022). https://plato.stanford.edu/archives/sum2022/entries/international-justice/.

Blumenfeld, Warren J. "Christian Privilege and the Promotion of 'Secular' and Not-So 'Secular' Mainline Christianity in Public Schooling and in the Larger Society." *Equity and Excellence in Education* 39, no. 3 (November 2006): 195–210.

Braithwaite, Valerie. "Institutional Oppression That Silences Child Protection Reform." *International Journal on Child Maltreatment: Research, Policy and Practice* 4 (April 2021): 49–72.

Brennan, Jason, and Christopher Freiman. "Moral Philosophy's Moral Risk." *Ratio* 33, no. 3 (September 2020): 191–201.

Bright, Liam Kofi, Daniel Malinsky, and Morgan Thompson. "Causally Interpreting Intersectionality Theory." *Philosophy of Science* 83, no. 1 (January 2016): 60–81.

Britton-Purdy, Jedediah. "What John Rawls Missed: Are His Principles of a Just Society Enough Today?" *The New Republic*, October 29, 2019. https://newrepublic.com/article/155294/john-rawls-missed-create-just-society.

Byskov, Morten Fibieger. "What Makes Epistemic Injustice an 'Injustice'?" *Journal of Social Philosophy* 52, no. 1 (Spring 2021): 114–31.

Carter, Ray. "Expert Says CRT Designed to Undermine American Values." *OCPA*, September 28, 2021. https://www.ocpathink.org/post/expert-says-crt-designed-to-undermine-american-values.

Cherry, Myisha. *The Case for Rage: Why Anger Is Essential to Anti-Racist Struggle*. New York: Oxford University Press, 2021.

Coleman, Arica L. "What's Intersectionality? Let These Scholars Explain the Theory and Its History." *Time*, March 29, 2019. https://time.com/5560575/intersectionality-theory/.

Collins, Patricia Hill. "Learning from the Outsider Within: The Sociological Significance of Black Feminist Thought." *Social Problems* 33, no. 6 (December 1986): s14–s32.

Collins, Patricia, and Sirma Bilge. *Intersectionality*. 2nd ed. Cambridge: Polity Press, 2016.

Conway, Lucian Gideon, Meredith A. Repke, and Shannon C. Houck. "Donald Trump as a Cultural Revolt against Perceived Communication Restriction:

Priming Political Correctness Norms Causes More Trump Support." *Journal of Social and Political Psychology* 5, no. 1 (May 2017): 244–59.

Cook, Judith A., and Mary Margaret Fonow. "Knowledge and Women's Interests: Issues of Epistemology and Methodology in Feminist Sociological Research." *Sociological Inquiry* 56, no. 4 (January 1986): 2–29.

Coulter, Robert W. S., Christina Mair, Elizabeth Miller, John R. Blosnich, Derrick D. Matthews, and Heather L. McCauley. "Prevalence of Past-Year Sexual Assault Victimization among Undergraduate Students: Exploring Differences by and Intersections of Gender Identity, Sexual Identity, and Race/Ethnicity." *Prevention Science* 18, no. 6 (August 2017): 726–36.

Courtland, Shane D., Gerald Gaus, and David Schmidtz. "Liberalism." In *Stanford Encyclopedia of Philosophy* (Spring 2022). https://plato.stanford.edu/archives/spr2022/entries/liberalism/.

Crenshaw, Kimberlé Williams. "Beyond Racism and Misogyny: Black Feminism and 2 Live Crew." In *Words that Wound: Critical Race Theory, Assaultive Speech, and the First Amendment*, edited by Mari J. Matsuda, Charles R. Lawrence III, Richard Delgado, and Kimberlé Williams Crenshaw, 111–32. Boulder, CO: Westview Press, 1993.

———. "Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics." *University of Chicago Legal Forum* 1 (1989): 139–67.

Crenshaw, Kimberlé Williams, Neil Gotanda, Gary Peller, and Kendall Thomas, eds. *Critical Race Theory: The Key Writings That Formed the Movement*. New York: The New Press, 1995.

Crosthwaite, Jan. "Feminist Criticism of Liberalism." *Political Science* 39, no. 2 (December 1987): 172–84.

Curry, Tommy J. *The Man-Not: Race, Class, Genre, and the Dilemmas of Black Manhood*. Philadelphia: Temple University Press, 2017.

———. "Racism and the Equality Delusion." *IAI News*, July 16, 2021. https://iai.tv/articles/racism-and-the-equality-delusion-auid-1836.

Davis, Emmalon. "On Epistemic Appropriation." *Ethics* 128, no. 4 (July 2018): 702–27.

Davis, Kathy. "Intersectionality as Buzzword: A Sociology of Science Perspective on What Makes a Feminist Theory Successful." *Feminist Theory* 9, no. 1 (April 2008): 67–85.

De Cremer, David, and Tom R. Tyler. "The Effects of Trust in Authority and Procedural Fairness on Cooperation." *Journal of Applied Psychology* 92, no. 3 (May 2007): 639–49.

De Cremer, David, and Daan Van Knippenberg. "How Do Leaders Promote Cooperation? The Effects of Charisma and Procedural Fairness." *Journal of*

*Applied Psychology* 87, no. 5 (October 2022): 858–66.

DeJong, Keri, and Barbara Love. "Youth Oppression and Elder Oppression." In *Teaching for Diversity and Social Justice*, 3rd, ed., edited by Maurianne Adams and Lee Anne Bell with Diane J. Goodman and Khyati Y. Joshi, 339–68. New York : Routledge, 2016.

Delgado, Richard, and Jean Stefancic. *Critical Race Theory: An Introduction*. 2nd ed. New York: New York University Press, 2012.

Devadson, Lousia. "Allyship: A Guide." *One Woman Project*, May 13, 2020. https://www.onewomanproject.org/allyship/allyship-a-guide.

Dinno, Alexis. "Homicide Rates of Transgender Individuals in the United States: 2010–2014." *American Journal of Public Health* 107, no. 9 (September 2018): 1441–47.

Disch, Lisa, and Mary Hawkesworth, eds. *The Oxford Handbook of Feminist Theory*. Oxford: Oxford University Press, 2018.

Ditrich, Lara, Adrian Lüders, Eva Jonas, and Kai Sassenberg. "You Gotta Fight!—Why Norm-Violations and Outgroup Criticism Lead to Confrontational Reactions." *Cognition and Emotion* 36, no. 2 (November 2021): 254–72.

Ditum, Sarah. "Trans Rights Should Not Come at the Cost of Women's Fragile Gains." *The Economist*, July 5, 2018. https://www.economist.com/open-future/2018/07/05/trans-rights-should-not-come-at-the-cost-of-womens-fragile-gains.

Dotson, Kristie, and Ayanna De'Vante Spencer. "Another Letter Long Delayed: On Unsound Epistemological Practices and Reductive Inclusion." *Philosophical Topics* 46, no. 2 (Fall 2018): 51–69.

Dubeau, Mathieu. "Species-Being for Whom? The Five Faces of Interspecies Oppression." *Contemporary Political Theory* 19, no. 4 (December 2020): 596–620.

Dubrow, Joshua. "Why Should We Account for Intersectionality in Quantitative Analysis of Survey Data?" In *Intersectionality und Kritik: Neue Perspektiven für alte Fragen*, edited by Vera Kallenberg, Jennifer Meyer, and Johanna M. Müller, 161–77. Wiesbaden: Springer, 2013.

*Economist*. "The Threat from the Illiberal Left," September 4, 2021. https://www.economist.com/leaders/2021/09/04/the-threat-from-the-illiberal-left.

Edwards, Keith E. "Aspiring Social Justice Ally Identity Development: A Conceptual Model." NASPA *Journal* 43, no. 4 (January 2007): 39–60.

Enslin, Penny, and Mary Tjiattas. "Educating for a Just World without Gender." *Theory and Research in Education* 4, no. 1 (March 2006): 41–68.

Erman, Eva, and Niklas Möller. "Is Ideal Theory Useless for Non-Ideal Theory?" *Journal of Politics* 84, no. 1 (January 2022): 525–40.

———. "Three Failed Charges against Ideal Theory." *Social Theory and Practice* 39, no. 1 (January 2013): 19–44.

Espindola, Juan, and Moises Vaca. "The Problem of Historical Rectification for Rawlsian Theory." *Res Publica* 20, no. 3 (August 2014): 227–43.

Evans, Elizabeth, and Éléonore Lépinard. "Confronting Privileges in Feminist and Queer Movements." In *Intersectionality in Feminist and Queer Movements: Confronting Privileges*, edited by Elizabeth Evans and Éléonore Lépinard, 1–26. Oxford: Routledge, 2020.

Farrelly, Colin. "Justice in Ideal Theory: A Refutation." *Political Studies* 55, no. 4 (December 2007): 844–64.

Federici, Silvia. *Caliban and the Witch: Women, the Body and Primitive Accumulation*. New York: Autonomedia, 2003.

Fejős, Anna, and Violetta Zentai, eds. "Anti-Gender Hate Speech in Populist Right-Wing Social Media Communication." Barcelona: GENHA, 2021. http://genha.eu/publications-reports.

FeldmanHall, Oriel, Peter Sokol-Hessner, Jay J. Van Bavel, and Elizabeth A. Phelps. "Fairness Violations Elicit Greater Punishment on Behalf of Another than for Oneself." *Nature Communications* 5, no. 1 (October 2014): 1–6.

Floyd, Jonathan. "Political Philosophy's Methodological Moment and the Rise of Public Political Philosophy." *Society* 59, no. 2 (April 2022): 129–39.

Ford, Chandra L., and Collins O. Airhihenbuwa. "Critical Race Theory, Race Equity, and Public Health: Toward Antiracism Praxis." *American Journal of Public Health* 100, no. S1 (April 2010): S30–S35.

Forrester, Katrina. *In the Shadow of Justice: Postwar Liberalism and the Remaking of Political Philosophy.* Princeton, NJ: Princeton University Press, 2019.

Foster, Sheila. "Rawls, Race, and Reason." *Fordham Law Review* 72, no. 5 (April 2004): 1715–9.

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

Gaard, Greta. "Vegetarian Ecofeminism: A Review Essay." *Frontiers: A Journal of Women Studies* 23, no. 3 (January 2002): 117–46.

Gasdaglis, Katherine, and Alex Madva. "Intersectionality as a Regulative Ideal." *Ergo: An Open Access Journal of Philosophy* 6, no. 44 (2019–2020): 1287–330.

Gaynor, Tia Sherèe. "Social Construction and the Criminalization of Identity: State-Sanctioned Oppression and an Unethical Administration." *Public Integrity* 20, no. 4 (February 2018): 358–69.

Geuss, Raymond. *The Idea of a Critical Theory*. Cambridge: Cambridge University Press, 1981.

Ghabra, Haneen, and Bernadette Marie Calafell. "From Failure and Allyship to

Feminist Solidarities: Negotiating Our Privileges and Oppressions across Borders." *Text and Performance Quarterly* 38, nos. 1–2 (April 2018): 38–54.

Girvetz, Harry K., Richard Dagger, and Kenneth Minogue "Liberalism." In *Britannica*. Last updated May 13, 2022. https://www.britannica.com/topic/liberalism.

Goodhart, Michael. *Injustice: Political Theory for the Real World*. New York: Oxford University Press, 2018.

Goodkind, Jessica R., and Zermarie Deacon. "Methodological Issues in Conducting Research with Refugee Women: Principles for Recognizing and Re-centering the Multiply Marginalized." *Journal of Community Psychology* 32, no. 6 (November 2004): 721–39.

Grasswick, Heidi. "Feminist Social Epistemology." In *Stanford Encyclopedia of Philosophy* (Spring 2018). https://plato.stanford.edu/archives/fall2018/entries/feminist-social-epistemology/.

Haeffner, Melissa, Dana Hellman, Alida Cantor, Idowu Ajibade, Vinka Oyanedel-Craver, Maura Kelly, Laura Schifman, and Lisa Weasel. "Representation Justice as a Research Agenda for Socio-Hydrology and Water Governance." *Hydrological Sciences Journal* 65, no. 11 (August 2021): 1611–24.

Hartley, Christie, and Lori Watson. "Is a Feminist Political Liberalism Possible?" *Journal of Ethics and Social Philosophy* 5, no. 1 (October 2010): 1–21.

Hay, Carol. *Kantianism, Liberalism, and Feminism: Resisting Oppression*. New York: Palgrave-Macmillan, 2013.

Heerten-Rodriguez, Liam. "Fat Peoples' Experiences of and Responses to Sexualized Oppression: A Multi-Method Qualitative Study." PhD diss., University of Nebraska-Lincoln, 2019.

Henderson, James (Sákéj) Youngblood Henderson. "The Context of the State of Nature." In *Reclaiming Indigenous Voice and Vision*, edited by Marie Battiste, 11–38. Vancouver: UBC Press, 2000.

Higgins, Peter. "Three Hypotheses for Explaining the So-Called Oppression of Men." *Feminist Philosophy Quarterly* 5, no. 2 (August 2019): 1–19.

Higgs, Paul, and Chris Gilleard. "Is Ageism an Oppression?" *Journal of Aging Studies*, 62, art. 101051 (September 2022): 1–6.

hooks, bell. *Feminist Theory: From Margin to Center*. London: Pluto Press, 2004.

Horkheimer, Max. *Critical Theory: Selected Essays*. New York: Continuum Publishing, 1982.

Joyce, Helen. *Trans: When Ideology Meets Reality*. London: Oneworld Publications, 2021.

Juan, E. San Juan., Jr. "From Race to Class Struggle: Marxism and Critical Race Theory." *Nature, Society, and Thought* 18, no. 3 (July 2005): 333–56.

Kaba, Mariame. *We Do This 'Til We Free Us: Abolitionist Organizing and*

*Transforming Justice*. Chicago: Haymarket Books, 2021.

Kang, Hye Ryoung. "Can Rawls's Nonideal Theory Save His Ideal Theory?" *Social Theory and Practice* 42, no. 1 (January 2016): 32–56.

Kapoor, Ravi Shanker. "Feminism Is Illiberal." *Sunday Guardian*, May 29, 2018. https://www.sundayguardianlive.com/opinion/feminism-is-illiberal.

Kendi, Ibrahim X. *How to Be an Antiracist*. New York: One World, 2019.

Khader, Serene J. *Decolonizing Universalism: A Transnational Feminist Ethic*. New York: Oxford University Press, 2018.

Kittay, Eva Feder. *Love's Labor: Essays on Women, Equality and Dependence*. London: Routledge, 1999.

———. "Love's Labor Revisited." *Hypatia* 17, no. 3 (Summer 2002): 237–51.

Koyama, Emi. "The Transfeminist Manifesto." In *Catching a Wave: Reclaiming Feminism for the Twenty-First Century*, edited by Rory Dicker and Alison Piepmeier, 244–62. Lebanon, NH: Northeastern University Press, 2003.

Krishnamurthy, Meena. "Completing Rawls's Arguments for Equal Political Liberty and Its Fair Value: The Argument from Self-Respect." *Canadian Journal of Philosophy* 43, no. 2 (January 2020): 179–205.

———. "Reconceiving Rawls's Arguments for Equal Political Liberty and Its Fair Value." *Social Theory and Practice* 38, no. 2 (April 2012): 258–78.

Lawrence, Charles R., III, Mari J. Matsuda, Richard Delgado, and Kimberlé Crenshaw. "Introduction." In *Words That Wound: Critical Race Theory, Assaultive Speech, and the First Amendment*, edited by Charles R. Lawrence III, Mari J. Matsuda, Richard Delgado, and Kimberlé Crenshaw. Boulder, CO: Westview Press, 1993.

Leeb, Claudia. "Marx and the Gendered Structure of Capitalism." *Philosophy and Social Criticism* 33, no. 7 (November 2007): 833–59.

Leonardo, Zeus. "The Race for Class: Reflections on a Critical Raceclass Theory of Education." *Educational Studies: A Journal of the American Educational Studies Association* 48, no. 5 (September 2012): 427–49.

Lépinard, Éléonore. *Feminist Trouble: Intersectional Politics in Post-Secular Times*. New York: Oxford University Press, 2020.

———. "Impossible Intersectionality? French Feminists and the Struggle for Inclusion." *Politics and Gender* 10, no. 1 (March 2014): 124–30.

Liao, Shen-yi, and Vanessa Carbonell. "Materialized Oppression in Medical Tools and Technologies." *The American Journal of Bioethics* 23, no. 4 (March 2022): 1–15.

Lorde, Audre. "The Master's Tools Will Never Dismantle the Master's House." In *Sister Outsider: Essays and Speeches*. Berkeley, CA: Crossing Press, 1984.

Lugones, Marìa. "Toward a Decolonial Feminism." *Hypatia* 25, no. 4 (Summer 2010): 742–59

Manuel, Jennifer I. "Racial/Ethnic and Gender Disparities in Health Care Use and Access." *Health Services Research* 53, no. 3 (June 2018): 1407–29.

Matthes, Erich Hatala. "Cultural Appropriation and Oppression." *Philosophical Studies* 176, no. 4 (April 2019): 1003–13.

Matthew, D. C. "Rawlsian Affirmative Action." *Critical Philosophy of Race* 3, no. 2 (July 2015): 324–43.

———. "Rawls and Racial Justice." *Politics, Philosophy & Economics* 16, no. 3 (August 2017): 235–58.

———. "Rawls's Ideal Theory: A Clarification and Defense." *Res Publica* 25, no. 4 (November 2019): 553–70.

Mayes, Renae D., and Janice A. Byrd. "An Antiracist Framework for Evidence-Informed School Counseling Practice." *Professional School Counseling* 26, no. 1a (March 2022): 1–10.

McIntosh, Peggy. "Reflections and Future Directions for Privilege Studies." *Journal of Social Issues* 68, no. 1 (March 2012): 194–206.

McKinnon, Rachel. "Trans*formative Experiences." *Res Philosophica* 92, no. 2 Mendoza, Saiid A., Sean P. Lane, and David M. Amodio. "For Members Only: Ingroup Punishment of Fairness Norm Violations in the Ultimatum Game." *Social Psychological and Personality Science* 5, no. 6 (August 2014): 662–70.

Meshelski, Kristina. "Procedural Justice and Affirmative Action." *Ethical Theory and Moral Practice* 19, no. 2 (April 2016): 425–43.

(April 2015): 419–40.

Michaelis, Karen L. "From Indifference to Injustice: The Politics of School Violence." *Journal of School Leadership* 14, no. 1 (January 2004): 32–61.

Michener, Jamila, and Margaret Teresa Brower. "What's Policy Got to Do with it? Race, Gender and Economic Inequality in the United States." *Daedalus* 149, no. 1 (Winter 2020): 100–18.

Mills, Charles W. *Black Rights/White Wrongs: The Critique of Racial Liberalism*. New York: Oxford University Press, 2017.

———. "'Ideal Theory' as Ideology." *Hypatia* 20, no. 3 (Summer 2005): 165–84.

———. "Occupy Liberalism! Or Ten Reasons Why Liberalism Cannot Be Retrieved for Radicalism (And Why They Are All Wrong)." *Radical Philosophy Review* 15, no. 2 (October 2012): 305–23.

———. *The Racial Contract*. Ithaca, NY: Cornell University Press, 1997.

———. "Rawls on Race/Race in Rawls." *The Southern Journal of Philosophy*, 47, no. S1 (March 2009): 161–84.

———. "Retrieving Rawls for Racial Justice? A Critique of Tommie Shelby." *Critical Philosophy of Race* 1, no. 1 (April 2013): 1–27.

———. "Rousseau, the Master's Tools, and Anti-Contractarian Contractarianism." *The* CLR *James Journal* 15, no. 1 (Spring 2009): 92–112.

———. "The Whiteness of John Rawls." YouTube video. Uploaded on May 21, 2015. https://www.youtube.com/watch?v=DVhLoTeR-lQ.

Minh-ha, Trinh T. *Woman, Native, Other: Writing Postcoloniality and Feminism*. Bloomington, IN: Indiana University Press, 1989.

Nash, Jennifer C. "Re-thinking Intersectionality." *Feminist Review* 89, no. 1 (June 2008): 1–15.

Nesbitt, Shawna, and Rigo Estevan Palomarez. "Increasing Awareness and Education on Health Disparities for Health Care Providers." *Ethnicity and Disease* 26, no. 2 (Spring 2016): 181–90.

Northway, Ruth. "Disability and Oppression: Some Implications for Nurses and Nursing." *Journal of Advanced Nursing* 26, no. 4 (October 1997): 736–43.

Nussbaum, Martha C. *The Feminist Critique of Liberalism*. Lawrence, KS: University of Kansas, 1997.

Obradors-Campos, Miguel. "Deconstructing Biphobia." *Journal of Bisexuality* 11, no. 2–3 (June 2011): 207–26.

O'Hagan, Ellie Mae. "The 'Anti-Woke' Backlash Is No Joke—And Progressives Are Going to Lose If They Don't Wise Up." *The Guardian*, January 30, 2020. https://www.theguardian.com/commentisfree/2020/jan/30/anti-woke-backlash-liberalism-laurence-fox.

Okin, Susan Moller. "Justice and Gender." *Philosophy and Public Affairs* 16, no. 1 (Winter 1987): 42–72.

———. *Justice, Gender, and the Family*. New York: Basic Books, 1989.

Oktay, Emine Naz. "Color-Blindness in Rawls's Theory of Justice." PhD diss., Bilkent Universitesi, 2019.

Olding, Lisa. "Racism and English Language Learning: Employing an Anti-Racist Approach to English as an Additional Language Education." *SFU Educational Review* 9 (May 2017): 1–12.

O'Neill, Onora. *Bounds of Justice*. Cambridge: Cambridge University Press, 2000.

Orgera, Kendal, and Samantha Artiga. *Disparities in Health and Health Care: Five Key Questions and Answers*. San Francisco: The Henry J. Kaiser Family Foundation, 2018.

Owens, Michael Leo. "Excavating Oppression in the Wake of Ferguson, Baltimore, and Municipal Everywhere." Opening statement at the Urban Affairs Association colloquy "Re-Thinking Justice in the Wake of Ferguson and Baltimore," San Diego, March 17, 2016.

Pateman, Carole. *The Sexual Contract*. Stanford, CA: Stanford University Press, 1988.

Pateman, Carole, and Charles W. Mills. *Contract and Domination*. Cambridge:

Polity Press, 2007.

Pengelly, Martin. "Barack Obama Warns Progressives to Avoid 'Circular Firing Squad.'" *The Guardian*, April 6, 2019. https://www.theguardian.com/us-news/2019/apr/06/barack-obama-progressives-circular-firing-squad-democrats.

Petri, Alexandra. "Sorry I Can't Comment on the President's Actions, I Just Remembered I'm Turning into a Bird." *The Washington Post*, June 3, 2020. https://www.washingtonpost.com/opinions/2020/06/03/sorry-i-cant-comment-presidents-actions-i-just-remembered-im-turning-into-bird/.

Phillips, Michael. "Reflections on the Transition from Ideal to Non-Ideal Theory." *Noûs* 19, no. 4 (December 1985): 551–70.

Piketty, Thomas, and Emmanuel Saez. "The Evolution of Top Incomes: A Historical and International Perspective." *American Economic Review* 96, no. 2 (May 2006): 200–205.

Powers, Kirsten. "Illiberal Feminism Is Running Amok." *The Daily Beast*, July 12, 2017. https://www.thedailybeast.com/illiberal-feminism-is-running-amok.

Price, Lisa L., Gisella S. Cruz-Garcia, and Nemer E. Narchi. "Foods of Oppression." *Frontiers in Sustainable Food Systems* 5 (March 2021): 1–7.

Racial Equity Tools. "Fundamentals." Accessed May 1, 2023. https://www.racialequitytools.org/resources/fundamentals.

Rauch, Jonathan. *The Constitution of Knowledge: A Defense of Truth*. Washington, DC: The Brookings Institution Press, 2021.

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

———. *The Law of Peoples, with "The Idea of Public Reason Revisited."* Cambridge, MA: Harvard University Press, 1999.

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

———. *A Theory of Justice*. Rev. ed. Cambridge, MA: The Belknap Press of Harvard University Press, 1999.

Roth, Silke. "Intersectionality and Coalitions in Social Movement Research—A Survey and Outlook." *Sociology Compass* 15, no. 7 (July 2021): 1–16.

Rothman, Emily F., Deinera Exner, and Allyson L. Baughman. "The Prevalence of Sexual Assault against People Who Identify as Gay, Lesbian, or Bisexual in the United States: A Systematic Review." *Trauma, Violence, and Abuse* 12, no. 2 (April 2011): 55–66.

Ruíz, Elena. "Framing Intersectionality." In *The Routledge Companion to the Philosophy of Race*, edited by Paul C. Taylor, Linda Martín Alcoff, and Luvell Anderson, 335–48. New York: Oxford University Press, 2018.

Salzman, Philip Carl. "Toxic Feminism." *Frontier Center for Public Policy*, July 11, 2018. https://fcpp.org/2018/07/11/toxic-feminism/.

Schmitt, Glenn R., Louis Reedt, and Kevin Blackwell. *Demographic Differences in Sentencing: An Update to the 2012 Booker Report*. The United States Sentencing Commission. November 14, 2017. https://www.ussc.gov/research/research-reports/demographic-differences-sentencing.

Schneider, Christopher J. "Integrating Critical Race Theory and Postmodernism Implications of Race, Class, and Gender." *Critical Criminology* 12, no. 1 (January 2004): 87–103.

Schoen, John W., and Yelena Dzhanova. "These Two Charts Show the Lack of Diversity in the House and Senate." CNBC, June 2, 2020. https://www.cnbc.com/2020/06/02/these-two-graphics-show-the-lack-of-diversity-in-the-house-and-senate.html.

Scholz, Sally J. *Political Solidarity*. University Park, PA: The Pennsylvania State University Press, 2008.

Schwartzman, Lisa H. *Challenging Liberalism: Feminism as Political Critique*. University Park, PA: Pennsylvania State University Press, 2006.

Seymour, Richard. "How Postmodernism Became the Universal Scapegoat of the Era." *The New Statesman*, June 24, 2021. https://www.newstatesman.com/politics/2021/06/how-postmodernism-became-universal-scapegoat-era.

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

———. "Justice, Deviance, and the Dark Ghetto." *Philosophy and Public Affairs* 35, no. 2 (Spring 2007): 126–60.

———. "Race and Social Justice: Rawlsian Considerations." *Fordham Law Review* 72, no. 5 (April 2004): 1697–714.

———. "Racial Realities and Corrective Justice: A Reply to Charles Mills." *Critical Philosophy of Race* 1, no. 2 (July 2013): 145–62.

Shlasko, Davey. "Using the Five Faces of Oppression to Teach about Interlocking Systems of Oppression." *Equity and Excellence in Education* 48, no. 3 (August 2015): 349–60.

Simmons, A. John. "Ideal and Nonideal Theory." *Philosophy and Public Affairs* 38, no. 1 (Winter 2010): 5–36.

Simpson, Brent, Robb Willer, and Matthew Feinberg. "Does Violent Protest Backfire? Testing a Theory of Public Reactions to Activist Violence." *Socius: Sociological Research for a Dynamic World* 4 (January–December 2018): 1–14.

Smith, David. "How Did Republicans Turn Critical Race Theory into a Winning Electoral Issue?" *The Guardian*, November 3, 2021. https://www.theguardian.com/us-news/2021/nov/03/republicans-critical-race-theory-winning-electoral-issue.

Stefano, Christine Di. "Marxist Feminism." In *The Encyclopedia of Political*

*Thought*, edited by Michael T. Gibbons. Wiley Online Library, 2014. https://doi.org/10.1002/9781118474396.wbept0653.

Stevens, Sydney. “Race and Wrongful Convictions.” Infographic. The National Registry of Exonerations. Accessed September 29, 2022. http://www.law.umich.edu/special/exoneration/Pages/Race-and-Wrongful-Convictions.aspx.

Sullivan, Andrew. “Removing The Bedrock of Liberalism.” *The Weekly Dish*, May 28, 2021. https://andrewsullivan.substack.com/p/removing-the-bedrock-of-liberalism-826.

Táíwò, Olúfémi O. “Being-in-the-Room Privilege: Elite Capture and Epistemic Deference.” *The Philosopher*, 2020. https://www.thephilosopher1923.org/post/being-in-the-room-privilege-elite-capture-and-epistemic-deference.

Trebilcot, Joyce. “Dyke Methods, or Principles for the Discovery/Creation of the Withstanding.” *Hypatia* 3, no. 2 (Summer 1988): 1–13.

Tucker, Ronnie B., Sr. “The Color of Mass Incarceration.” *Ethnic Studies Review* 37, no. 1 (January 2017): 135–49.

Tuggle, Zachary Kincaid. “Towards a Moral Conception of Allyship.” Master’s thesis, University of Tennessee, Knoxville, 2018. https://trace.tennessee.edu/cgi/viewcontent.cgi?article=6573&context=utk_gradthes.

Vinson, Kevin D. “Oppression, Anti-Oppression, and Citizenship Education.” In *The Social Studies Curriculum: Purposes, Problems, and Possibilities*, 3rd. ed., edited by E. Wayne Ross, 57–85. Albany, NY: State University of New York Press, 2001.

Wenar, Leif. “John Rawls.” In *Stanford Encyclopedia of Philosophy* (Spring 2017). https://plato.stanford.edu/archives/spr2017/entries/rawls/.

Wingfield, Marvin. “Arab Americans: Into the Multicultural Mainstream.” *Equity and Excellence in Education* 39, no. 3 (November 2006): 253–66.

Wolff, Jonathan. “Equality: The Recent History of an Idea.” *Journal of Moral Philosophy* 4, no. 1 (January 2007): 125–36.

Woodall, Denise Ruth, and Sarah Shannon. “Carceral Citizens Rising: Understanding Oppression Resistance Work through the Lens of Carceral Status.” *Social Service Review* 96, no. 2 (June 2022): 308–52.

Writer, Jeanette Haynes. “Unmasking, Exposing, and Confronting: Critical Race Theory, Tribal Critical Race Theory and Multicultural Education.” *International Journal of Multicultural Education* 10, no. 2 (December 2008): 1–15.

Yamamoto, Shinya, and Ayaka Takimoto. “Empathy and Fairness: Psychological Mechanisms for Eliciting and Maintaining Prosociality and Cooperation in Primates.” *Social Justice Research* 25, no. 3 (August 2012): 233–55.

Young, Donna E. “Post Race Posthaste: Towards an Analytical Convergence of Critical Race Theory and Marxism." *Columbia Journal of Race and Law* 1, no. 3 (July 2011): 499–510.

Young, Iris Marion. *Justice and the Politics of Difference*. Princeton, NJ: Princeton University Press, 1990.

Yuval-Davis, Nira. "Dialogical Epistemology—An Intersectional Resistance to the 'Oppression Olympics.'" *Gender and Society* 26, no. 1 (February 2012): 46–54.

Zanghellini, Aleardo. "Philosophical Problems with the Gender-Critical Feminist Argument against Trans Inclusion." *Sage Open* 10, no. 2 (April–June 2020): 1–14.

Zippia. "President Demographics and Statistics in the US." *Zippia,* September 9, 2022. https://www.zippia.com/president-jobs/demographics/.

*Journal of Ethics and Social Philosophy*
vol. 26, no. 2 · december 2023

*https://doi.org/10.26556/jesp.v26i2.2653*

# COMBATANTS, MASCULINITY, AND JUST WAR THEORY

*Graham Parsons*

Over the last several decades, the ethics of war has grown into a major subfield in philosophy. At least five major handbooks have been published on the subject in recent years.[1] There are journals, professional societies, newsletters, and major annual conferences in the United States, Europe, and Asia devoted to the topic. Over roughly the same period, there has developed a large literature spanning gender studies, political science, international relations, legal studies, and philosophy on the profound and complex relationship between gender and war.[2] This literature not only explores how gender can explain the occurrence and conduct of war. It also explores ways in which gender can *legitimate* the occurrence and conduct of war. This legitimation is sometimes thought to occur by grounding—explicitly or implicitly—the very moral and legal principles used to justify and criticize war itself. Partly because of this recognition of the connection between gender and the normative theory of war, there has also developed a large body of feminist theory calling, to varying degrees, for reconsideration of the ethics and law of war.[3]

Despite the obvious connections between these literatures, the conventional ethics of war literature and the literature on gender and war have remained largely independent. The mainstream ethics of war has not meaningfully

[1] Allhoff, Evans, and Henschke, *Routledge Handbook of Ethics and War*; Johnson and Pattison, *The Ashgate Research Companion to Military Ethics*; Lucas, *Routledge Handbook of Military Ethics*; May, *The Cambridge Handbook of the Just War*; and Frowe and Lazar, *The Oxford Handbook of Ethics of War*.

[2] See Tickner, *Gender in International Relations*; Elshtain, *Women and War*; Gardam, "Women and the Law of Armed Conflict"; Enloe, *Maneuvers*; Goldstein, *War and Gender*; Braudy, *From Chivalry to Terrorism*; Hutchings, "Making Sense of Masculinity and War"; Kinsella, *The Image before the Weapon*; Digby, *Love and War*; Mann, *Sovereign Masculinity*; and Sjoberg, *Gender, War, and Conflict*.

[3] See Ruddick, *Maternal Thinking*; Peach, "An Alternative to Pacifism"; Elshtain, *Women and War*; Gardam, "Women and the Law of Armed Conflict"; Young, "The Logic of Masculinist Protection"; Sjoberg, *Gender, Justice, and the Wars in Iraq*; Held, *Why Terrorism Is Wrong*; and Robinson, *The Ethics of Care*.

engaged with the question of gender and its influence on war. Consider that of the five recent handbooks on the ethics of war cited above, only one, *The Routledge Handbook of Military Ethics*, contains chapters (just two out of thirty-seven) that deal with issues of gender and sexuality by way of discussing the inclusion of women and homosexual or bisexual people in the armed forces.[4] Still, none of these handbooks engages meaningfully with the gender and war literature, and terms such as "feminism," "gender," "masculinity," and "sex" are missing from their indexes (although, notably, the index of the above *Routledge Handbook* has an entry on "sexism"). At the same time, the literature on gender and war has not been deeply engaged with recent ethics of war literature. While the gender and war literature has been occupied with major historical figures in just war theory, the myriad debates that have sprung up in analytical just war theory in the twenty-first century—and which currently constitute the mainstream of the field—have not been of much interest.

This article aims both to contribute to each of these literatures and to show at least one way they interconnect. I will argue that there is an important and underappreciated relationship between the concept of the moral equality of combatants and masculinity. The doctrine of the moral equality of combatants holds that combatants in war have an equal right to attack and kill one another regardless of the justice of their wars from the perspective of *jus ad bellum*. In other words, even if the combatants on one side are fighting an unjust war and the combatants on the other side are fighting a just war, all the combatants are equally morally permitted to fight. I will call the challenge of defending permissive views of attacks on combatants in war such as the moral equality of combatants the *external problem* of the soldier in just war theory.

While it has been widely discussed, I will argue that the basis of the moral equality of combatants in just war theory has been misunderstood in the ethics of war literature. Rather than being based on a peculiar view of the ethics of interpersonal self-defense, the moral equality of combatants is based on a view of the inferior political standing of soldiers vis-à-vis their political authorities. Thus, the moral equality of combatants treats combatants as soldiers who have a subordinate political status. However, at the same time they attribute this political standing to soldiers, the theories of political justice that canonical just war theorists advocate undermine that standing. I call this the *internal problem* of the soldier in just war theory.

To mitigate the internal problem, these just war theorists appeal, sometimes explicitly and sometimes implicitly, to a presupposed gender ontology that prescribes the role of self-sacrificial military servant to men on the basis of

[4] Lucas, *The Routledge Handbook of Military Ethics*.

their sex. In other words, the combatant in just war theory has been conceived of as a man who is bound to fight for his community and family on command because of his sexual nature. This gender ontology solves both the internal and external problems: by conceiving of the combatant as effectively a natural soldier, combatants both are expendable in war by their own governments and have a right to attack their enemies on the battlefield—they can be ordered into battle and kill their opponents not by virtue of their free choices but by virtue of who they are. If this is true, then the ethics of war literature should be much more concerned with the gender and war literature and make common cause with those who have treated the problem of the political standing of soldiers as a philosophical priority.

The idea of the moral equality of combatants has been the subject of intense debate in the field of just war theory over the last several decades. The debate has cleaved much of the ethics of war community into two camps—traditionalists and revisionists.[5] Traditionalists argue that the moral equality of combatants is basically sound. Revisionists, on the other hand, argue that the doctrine is false and, instead, that only combatants fighting in a war that is justified (i.e., meets the standards of *jus ad bellum*) are morally permitted to fight, while combatants fighting in a war that is unjustified have no moral permission to fight. The debate between traditionalists and revisionists has tended to treat the moral equality of combatants as originating in an implausible view of the ethics of interpersonal violence. On my reading of the tradition, this is wrong. In fact, the moral equality of combatants originates in a gender ontology that treats the male members of political communities as bound by nature to engage in self-sacrificial violence in war. Gender is therefore deeply entwined with the moral equality of combatants, the issue that is so central to the mainstream ethics of war literature.

Much literature on the relation between gender and war has already shown that masculinity has been foundational to the construction of the duties and rights of soldiers. This article is deeply indebted to these commentators, especially Jean Bethke Elshtain, Lucinda J. Peach, and Helen Kinsella. My argument contributes to this literature by showing, first, how masculinity relates to a topic of great interest to recent ethics of war scholarship—the moral equality of combatants—and, second, that there are notable appeals to masculinity in the just war tradition to ground the duties and rights of soldiers that have gone unnoticed in the gender and war literature.

This article is divided into three parts. The first section introduces the internal and external problems of the soldier by examining the debate over the moral equality of combatants and showing how, contrary to conventional readings, it

[5] See Lazar, "Just War Theory."

is based on a particular view of the domestic political standing of soldiers. This section also explains the tension between the political standing of soldiers and canonical just war theorists' visions of political justice. The second section aims to uncover the ways in which just war theorists mitigate the problem of justifying the political status of soldiers by appealing both overtly and covertly to gender and argues that this solves both the internal and the external problems. Last, I conclude that an urgent challenge for the ethics of war is to rethink the rights and duties of soldiers at both the international and domestic levels. In particular, we need to grant members of the military basic civil liberties domestically and recognize greater restrictions on attacks against combatants internationally. These goals can only be accomplished once we recognize the pernicious ways in which gender continues to lead us to reduce the people in military service to their office.

## 1. The Soldier as a Philosophical Problem

Conventional military ethics conceives of soldiers as having a relatively diminished moral standing. This is evident in at least two places. One—the external problem—is the permissive attitude taken toward attacks against combatants in war. I call this the external problem because it is a problem regarding the treatment of soldiers representing foreign or external political societies in war. The other—the internal problem—is the subordination of military servicemembers within their domestic armed forces and the denial of their basic civil liberties, especially their right to self-preservation. I call this the internal problem because it is a problem regarding the treatment of soldiers within their own political society. I refer to both as problems because, as I will contend, the prominent arguments for them are quite weak and it is unclear how they can be persuasively defended. In this section, I will examine both of these problems and argue that they are connected in the sense that the political subordination of military servicemembers is the source of the permissive view of killing combatants in war in the just war tradition.

### *1.1. The External Problem*

One of the most important constraints on conduct in war is the prohibition of all deliberate attacks on noncombatants. One may, according to the doctrine of double effect, subject noncombatants to the risk of unintentional harm in certain circumstances. Nevertheless, the immunity of noncombatants to deliberate harm is a bedrock principle of military ethics.[6]

[6] While some revisionist just war theorists have sought to undermine the immunity of noncombatants as such at the moral level, it is unclear to what extent they intend their moral assessments to alter conventional military practice and law. Jeff McMahan, for instance,

The conventional view of the status of combatants in war is much different. It is often claimed that while noncombatants are immune to intentional attack, combatants are generally fair game. Of course, wounded or captured combatants regain immunity as well as other positive rights. There are also prohibitions on the use of certain weapons used against combatants. But other than that, the law and conventions of war offer little or no further restrictions on harming combatants.[7] According to this view, the right to kill a combatant is not limited by the reasons for the combatant's enrollment in the armed forces, the justice of the combatant's cause, or whether they are currently engaged in combat. As Gabriella Blum concludes, "the striking feature of the mainstream literature is its general acceptance (albeit at times with some moral discomfort) of the near-absolute license to kill all combatants."[8]

This is an astonishingly permissive vision of the ethics of killing in war. Killing combatants is unconstrained by the principles of necessity and proportionality or the goal of achieving a peaceful resolution to the conflict. Regardless of their cause, personal motivations, or their present activity, combatants are legitimate targets. If they are not wounded or actively attempting to surrender, it appears combatants can be killed on sight. This view would seem to permit such scenes as the controversial "highway of death" in the 1991 Persian Gulf War.[9]

I have argued elsewhere that this view is expressed not just in law and conventions but also defended by canonical just war theorists such as Michael Walzer.[10] As Walzer puts it, combatants in war "can be attacked and killed at will by their enemies."[11] While he may not consistently assert this view, in multiple works, Walzer describes the permission to attack combatants as a class-based

while highly critical of the idea that noncombatants as a group are not liable to attack, nevertheless concludes that the prohibition of attacks on noncombatants ought to remain in place as a practical matter. See McMahan, "The Morality of War and the Law of War."

[7] See Blum, "The Dispensable Lives of Soldiers"; Ohlin, "Sharp Wars Are Brief"; and Haque, *Law and Morality at War*.

[8] Blum, "The Dispensable Lives of Soldiers," 72.

[9] On February 26–27, 1991, American-led coalition forces attacked large numbers of retreating Iraqi military personnel as they tried to escape Kuwait and enter Iraq. The estimates of how many people were killed vary widely from a few hundred to several thousand. In addition to allegations of indiscriminate attacks on civilians, the event is controversial because the destruction was arguably unnecessary given that the goal of the war was effectively achieved at the time of the attacks. For one account, see Atkinson, *Crusade,* chs. 16 and 17.

[10] See Parsons, "Walzer's Soldiers."

[11] Walzer, *Just and Unjust Wars,* 135–36.

permission tied to the combatant's status in the war, not their present activities, intentions, or strategic significance.[12]

But even if Walzer does apply the principles of necessity and proportionality to combatants, it is undeniable that he accepts the moral equality of combatants and its permission to attack combatants regardless of the justifiability of their cause. This is a strikingly permissive view of attacks on combatants in its own right. As has been pointed out by many revisionist commentators, this view permits violence against combatants that violates conventional restrictions on violence against people in other circumstances.

These revisionist critics have shown that one prominent argument Walzer offers in defense of the moral equality of combatants fails. The argument in question comes from his *Just and Unjust Wars*. He argues that all combatants are liable to attack because, as combatants, they are currently threatening their enemies. Whether they have just cause to threaten their enemies is not relevant. In Walzer's view, it is merely their threatening activity, whatever its cause, that makes combatants liable to be killed. As he says, "simply by fighting, whatever their private hopes and intentions, [combatants] have lost their title to life and liberty, and they have lost it even though, unlike aggressor states, they have committed no crime."[13]

As many revisionists have pointed out, this argument rests on an implausible view of liability to harm.[14] In all other circumstances, it is highly counterintuitive to hold that a person loses their right not to be killed simply by threatening others. A police officer, for example, who resorts to force to stop a person committing assault does not thereby become liable to attack by the assailant. Even though the officer poses an immediate threat to the assailant, it seems obvious that the officer retains his right to not be attacked while the assailant does not. The divergent causes of their threatening behavior seem to explain their divergent entitlements. It is only when one poses an *unjust* threat to others that one can be liable to attack. Those who *justly* threaten others maintain their right to not be harmed.

These critics argue that, by extension, the moral equality of combatants is wrong. Only combatants engaged in unjust wars are liable to attack, whereas

[12] In the preface to the second edition of *Just and Unjust Wars*, Walzer criticizes the treatment of combatants on the highway of death (see note 9 above) seemingly on the ground that the attacks were unnecessary. Despite such statements, there are numerous other passages where Walzer clearly does not apply necessity and proportionality to combatants. Some of these passages will be referenced in the subsequent discussion.

[13] Walzer, *Just and Unjust Wars*, 136.

[14] See McMahan, "Innocence, Self-Defense, and Killing in War"; and Rodin, *War and Self-Defense*.

combatants engaged in just wars are not liable to attack. Hence, combatants do not have an equal right to kill other combatants regardless of their cause.[15]

Moreover, these revisionist theorists have consistently argued that attacks on combatants in war must also be necessary and proportionate.[16] As pointed out above, the legal view of the right to kill combatants is indifferent to the necessity, proportionality, and conduciveness to sustainable peace of attacks against combatants. According to this view, combatants are simply fair game. The above criticism of the moral equality of combatants only takes issue with the traditional view's insensitivity to the moral justifiability of a combatant's overall cause. But Walzer's argument, even if true, gives no reason to think that attacks against combatants must not be necessary, proportionate, and conducive to future peace.

In sum, it seems plausible to conclude that there is more to justifying an attack on persons than their mere participation in hostilities. An attack against a combatant could be unethical because they are doing no wrong or because it is unnecessary, disproportionate, or hinders a sustainable peace.

### *1.2. The Internal Problem*

Critics of the moral equality of combatants have also been puzzled by another feature of traditional just war theory. In addition to the assertion of an equal right to kill between combatants, the traditional theory claims that combatants are not responsible for *jus ad bellum* but are responsible for *jus in bello*. This is sometimes called the *independence thesis* because it implies that *jus ad bellum* is logically independent of *jus in bello* in the sense that a war that violates *jus ad bellum* can nevertheless be fought in accordance with *jus in bello*.

Most commentators have taken the independence thesis to originate in Walzer's above defense of the moral equality of combatants.[17] Because that argument justifies killing any threatening combatant regardless of their cause, it seems to make killing in war independent of the reasons for resorting to war, or *jus ad bellum*.

In fact, the independence thesis has a different origin. Appreciating its actual origin helps us to see the internal problem of the soldier and highlights the relationship between the internal and external problems. Foundational figures in the just war tradition, including Walzer, quite clearly state that soldiers, not simply combatants, are not responsible for *jus ad bellum* because of their

[15] See Rodin, *War and Self-Defense*; McMahan, *Killing in War*; Frowe, *Defensive Killing*; Draper, *War and Individual Rights*; and Tadros, *To Do, to Die, to Reason Why*.

[16] See Lazar, "Necessity in Self-Defense and War"; and McMahan, "Proportionate Defense."

[17] See Rodin, *War and Self-Defense;* and McMahan, *Killing in War*.

prior obligations as occupants of peculiar social roles. Specifically, soldiers are obligated to fight in wars they are ordered to participate in by their legitimate political authority. This duty to obey can obligate soldiers to participate in wars even when those wars are unjust. Responsibility for *jus ad bellum* is thus divided between political authorities and soldiers—the authorities are obligated to abide by *jus ad bellum*, and soldiers are obligated to obey their authorities. Hence, a soldier could serve in an unjust war justly. If they were ordered to serve in a war that turned out to violate *jus ad bellum*, soldiers could still participate in that war and do nothing immoral.

Contrary to Walzer, most canonical just war theorists allow for exceptions to the soldier's responsibility to follow *jus ad bellum* decisions by their political leaders. When it is obvious that a war violates *jus ad bellum*, soldiers are not merely permitted to disobey but are obligated to. However, if they are unsure that a war they are ordered to participate in meets the standards of *jus ad bellum*, their duty to obey their sovereign trumps their duty to avoid participation in an unjust war. To cite just one example, Francisco de Vitoria argues that soldiers ought not to participate in wars that are patently unjust, but when they are unsure about the justice of the war they are "required to carry out the sentence of [their] superior."[18]

This view of the political obligations of soldiers is the ground of the independence thesis in the just war tradition. Because they are not responsible for *jus ad bellum* and are responsible for *jus in bello*, it is possible for soldiers to fight in a war that violates *jus ad bellum* yet fight justly. In this way, a soldier can fight an unjust war justly. For instance, contrary to how many commentators have read him, Walzer quite clearly states this is the ground of the independence thesis.[19] In his discussion of Erwin Rommel's conduct in World War II, he says:

> We draw a line between the war itself, for which soldiers are not responsible, and the conduct of war, for which they are responsible, at least within their own sphere of activity.… *We draw [the line] by recognizing the nature of political obedience.*… By and large we don't blame a soldier, even a general who fights for his own government. He is not the member of a robber band, a willful wrongdoer, but a loyal and obedient subject and citizen.… We allow him to say what an English soldier says in Shakespeare's *Henry V*: "We know enough if we know we are the king's men.

[18] Vitoria, *Political Writings*, 312; see also Suarez, *Selections from Three Works*.

[19] Most commentators have interpreted Walzer as arguing that soldiers who participate in an unjust war are innocent in the sense that they are excused from blame, not that they are justified. See Mapel, "Coerced Moral Agents?"; Primoratz, "Michael Walzer's Just War Theory"; McPherson, "Innocence and Responsibility in War; and McMahan, *Killing in War*.

> Our obedience to the king wipes the crime of it out of us."... [War] is conceived, both in international law and in ordinary moral judgment, as the king's business—a matter of state policy, not of individual volition, except when the individual is the king.[20]

This picture of the division of responsibility for *jus ad bellum* not only produces the independence thesis. It also grounds a version of the moral equality of combatants. If soldiers are generally not responsible for *jus ad bellum* but are obligated to obey their legitimate authority, then it will be common for the soldiers on opposing sides of a conflict to have the same moral status: they will be innocently carrying out their duties. Therefore, even if one side is fighting a war in violation of *jus ad bellum* and the other is fighting a war consistent with *jus ad bellum*, the soldiers on both sides will be equally innocent. As Vitoria puts it, "subjects neither must nor ought to examine the causes of war, but may follow their prince into war, content with the authority of their prince and public council; so that in general, even though the war may be unjust on one side or the other, the soldiers on each side who come to fight in battle or to defend a city are all equally innocent."[21]

But this subordination of soldiers creates a problem for the just war tradition. The problem is that it is hard to see how the obligation to fight in war on command can be justified in the first place. This is the internal problem of the soldier. It has to do with domestic political justice and the limits of political authority.

The idea that a person can be obligated to participate in war on command is the idea that a person can be a violent instrument of another. It entails that a person can justifiably engage in unjust violence because their obligations to obey trump their obligation to not engage in unjust violence. This is why traditional just war theorists thought *jus ad bellum* was independent of *jus in bello*: as instruments of their sovereign, soldiers can be obligated to fight even if the sovereign's war is unjust. Additionally, and crucially for my argument, this instrumentalization of soldiers implies that their lives are expendable. Given that the duty to fight in war on command binds soldiers to fight even under the threat of death—as traditional just war theorists clearly hold—the instrumentalization of soldiers entails that their right to life and self-preservation can be trumped by their duty to serve others. In this way, to conceive of soldiers as instruments in war is to conceive of them as expendable resources of communal defense: their personal interests in health and survival are not legitimate grounds to refuse an order to participate in war, even a war that turns out to be

[20] Walzer, *Just and Unjust Wars*, 38–39 (emphasis added).

[21] Vitoria, *Political Writings*, 321.

unjust. Indeed, most classical just war theorists considered it a capital crime to disobey an order out of fear that following it would cause injury to oneself.

Prior to the seventeenth century, canonical just war theorists operated within the framework of scholastic, quasi-Aristotelian social ontologies and theories of justice.[22] For them, political communities are natural bodies and individuals are their parts, analogous to the limbs of biological bodies. As such, the good of individuals is inextricably tied to their proper contribution to the community. The good of the political community, or the common good, can eclipse the private good of the individual. This vision of political justice makes justifying the obligations of soldiers relatively unproblematic. Vitoria, for instance, can argue that the commonwealth's right to use the soldier in war is analogous to the body's "right" to use its limb in self-defense. As he says,

> Every man has the power and right of self-defense by natural law, since nothing can be more natural than to repel force with force. Therefore the commonwealth, in which "we, being many, are one body, and every one member one of another" as the Apostle says (Rom. 12:5), ought not to lack the power and right which individual men assume or have over their bodies, to command the single limbs for the convenience and use of the whole. Individuals may even risk the loss of a limb if this is necessary to the safety of the rest of the body; and there is no reason why the commonwealth should not have the same power to compel and coerce its members as if they were its limbs for the utility and safety of the common good.[23]

But this method of grounding the subordination of soldiers was threatened by a philosophical revolution initiated by the seventeenth-century just war theorist Hugo Grotius. Grotius was arguably the first social contract theorist.[24] He rejects the natural character of the political community and instead argues that it is a human artifact made voluntarily by men to protect their private natural rights. According to this view, the rights of men are prior to the rights of the state, and the purpose of the state is to protect the rights of its male members. As Grotius says, the state is "a compleat Body of free Persons, associated together to enjoy peaceably their Rights, and for their common Benefit."[25] Grotius makes it clear that the "persons" of the political association are only men.

[22] See Parsons, "What Is the Classical Theory of Just Cause?"

[23] Vitoria, *Political Writings*, 11.

[24] See Tuck, *Natural Rights Theories* and *The Rights of War and Peace*; Haakonssen, "Hugo Grotius and the History of Political Thought"; Schneewind, *The Invention of Autonomy*; and Darwall, "Grotius at the Creation of Modern Moral Philosophy."

[25] Grotius, *The Rights of War and Peace*, 162.

He excludes women as naturally inferior to men and asserts that the patriarchal household is prior to the social contract.[26]

While the social contract is supposed to protect the rights of male participants, Grotius argues that soldiers are instruments of their political communities who are bound to risk their lives for the sake of others.[27] On the surface, the origin of this obligation is the social contract. According to Grotius, the reason the commonwealth offers more protection for the rights of men than the state of nature is precisely its ability to command its members to come to the assistance of other members and their association as an organized military force. For him, "the Design of Society is, that everyone should quietly enjoy his own, with the Help, and by the united Force of the whole Community."[28]

Some version of this contract argument for the right of the state to treat its members as instruments in war is shared by all the canonical early modern just war theorists. Samuel von Pufendorf and Emer Vattel, for instance, offer a similar theoretical framework that includes the explicit restriction of the social contract to men. Pufendorf is not quite as strident as Grotius regarding the natural inferiority of women but nevertheless treats patriarchal marriage as prior to the social contract and political society as an association of men.[29] Unlike Grotius and Pufendorf, Vattel does not develop a theory of marriage and its place in nature. However, he consistently speaks of the state as a "society of men" and, based on the rights and duties of citizens he develops (including the right of nations to "carry off" women in foreign countries), it is clear that he views political society as literally a society of men and not women.[30] Early modern just war theory thus has a gender hierarchy at its very foundation and, in turn, conceives of political membership and military service as roles for men exclusively and reduces women to natural domestic assistants to men.[31]

[26] Grotius, *The Rights of War and Peace*, 709; see also Kinsella, *The Image Before the Weapon*, 71–72.

[27] Grotius, *The Rights of War and Peace*, 386.

[28] Grotius, *The Rights of War and Peace*, 184.

[29] See Pateman, *The Sexual Contract*, 50–51; Sreedhar, "Pufendorf on Patiarchy"; Drakopoulou, "Samuel Pufendorf, Feminism, and the Question of 'Women and Law'"; Parsons, "Contract, Gender, and the Emergence of the Civil-Military Distinction."

[30] Vattel, *The Law of Nations*, 321.

[31] This point has been made about social contract theory generally by many prominent feminist critics of modern liberalism (see, for instance, Pateman, *The Sexual Contract*; and Okin, *Justice, Gender, and the Family*). My argument relies on this interpretation of social contract theory and accepts that women are victims of acute gender oppression in theory and in social practice. The fact that the argument I develop below focuses on the way that gender imposes self-sacrificial norms on men should not be taken to imply that women are not severely oppressed by gender. It would be illuminating to compare and contrast

But even if we put aside the gender hierarchy presupposed by this theory, the contract method of grounding the subordination of soldiers still has serious problems. If the purpose of the social contract is to protect the individual rights of the participants, and the rights of the participants are prior to the rights of the political authority, then the rights of individual men always prevail over the rights of the state. Regardless of which sexes are included in the social contract, the contract method strongly prioritizes the individual over the state such that the individual cannot be treated as a mere instrument of the state. In effect, the contract argument for the subordination of soldiers amounts to arguing that participants to the social contract render themselves instruments of the state to be used and sacrificed in war for the sake of protecting their private rights. This seems simply irrational. G. W. F. Hegel recognizes this problem for the contract tradition when he states that "it is a grave miscalculation if the state, when it requires this sacrifice [service in war], is simply equated with civil society, and if its ultimate end is seen merely as the *security of the life and property* of individuals. For this security cannot be achieved by the sacrifice of what is supposed to be secured—on the contrary."[32]

But the failure of the contract argument for military subordination is not simply a failure of the means-to-ends reasoning of the supposed participants to the social contract. More fundamentally, the problem is simply that to treat soldiers or anyone else as instruments is to violate their status as free and equal persons. Even if there were a social contract theory that could make alienating one's rights rational, we should still object to the treatment of people as instruments. Such subordination of persons is intrinsically wrong in that it patently treats others as mere means. Indeed, this is Immanuel Kant's very objection to standing armies. As he says, "the hiring of men to kill or be killed seems to mean using them as mere machines and instruments in the hands of someone else (the state), which cannot easily be reconciled with the rights of man in one's own person."[33]

Some might respond by asserting that this is only a concern for systems of conscription. If militaries recruit only volunteers, then there is no conflict between the rights of servicemembers and the obligations of military service. Voluntary military service is no different from employment in hierarchical private firms.

This reply misunderstands the severity of military subordination both in theory and in practice. To be under command is to be legally bound to obey

the experiences of women in the household and men in the military. Unfortunately, there is no room for that comparison in this article.

[32] Hegel, *Elements of the Philosophy of Right*, 361; see also MacIntyre, "Is Patriotism a Virtue?"

[33] Kant, "Perpetual Peace," 96.

orders even under the danger of death.[34] Insubordination in the military is a crime, not simply grounds for dismissal. Soldiers exist in a political space separate from civilians. Soldiers have radically diminished civil standing; they are literally second-class citizens. As the enlistment contract of the US Armed Forces puts it, the "enlistment/reenlistment agreement is more than an employment contract. It effects a change in status from civilian to military member of the Armed Forces."[35] Once a civilian becomes a service member, they are legally obligated to obey commands that can impact nearly all aspects of their lives. As one commentator summarizes the difference between civilians and service members:

> Once military status is acquired, military service loses its voluntary character. Once an individual has changed his or her status from civilian to military, that person's duties, assignments, living conditions, privacy, and grooming standards are all governed by military necessity, not personal choice. In a nation that places great value on freedom of expression, freedom of association, freedom of travel, and freedom of employment, the armed forces stand as a stark exception. Military commanders have the authority, as they have throughout our nation's history, to tell servicemembers where to live, where to work, and when they must put their lives at risk.[36]

For this reason, even voluntarily enlisted military service is a violation of the rights of the volunteer.

## 2. Masculinity and the Problem of the Soldier

To reiterate, traditional just war theory treats soldiers as possessing a diminished moral standing in at least two respects. First, soldiers are legitimate targets in war regardless of its cause. Second, soldiers may be subordinated to their political authorities such that they can be used and sacrificed by their states on command. As we have seen, the prominent arguments for these positions in the just war tradition are weak. Nevertheless, both positions are manifest in domestic and international law and are often treated as common sense. What is it that has made these positions seem so defensible over the centuries?

[34] Ned Dobos demonstrates the extent to which this is a departure from the rights workers are granted in other contexts. See Dobos, "Punishing Non-Conscientious Disobedience."

[35] US Department of Defense, Enlistment/Reenlistment Document.

[36] Nunn, "The Fundamental Principles of the Supreme Court's Jurisprudence in Military Cases," 5.

A significant part of the answer to this question is the influence gender has played in just war theory. These problematic aspects of just war theory are grounded in part in the presumption of a natural, gendered division of social labor. We have seen that early modern just war theory presupposes a hierarchical gender ontology that reduces women to the status of domestic instruments for men and grounds the restriction of political society to men only. The rights and duties of political subjects in these theories are the rights and duties of men. But while these theories describe men as free and equal individuals who come together in political society to protect their freedom and equality, they simultaneously rely on masculine virtues to ground the duty of military service. To be specific, the just war tradition has relied on the assumption that it is good for men as men to sacrifice themselves in violent combat to protect their families and communities.

Kinsella has argued persuasively that just war theory's moral distinction between combatants and civilians is constructed on a conception of individuals that reduces them to their sex based upon a presupposed gender ontology.[37] As she argues, we can see how just war theory uses a gender ontology to divide communities into combatants and civilians by observing how its major protagonists construct the category of the civilian. The boundaries of the "civilian" are constructed in part by appeal to the status of women, whom traditional just war theorists have held to be subordinate to men politically and excluded from combat because of their supposedly natural characters and social roles. The view I develop here complements Kinsella's reading of the role of gender in the just war tradition. However, my argument focuses more directly on how just war theorists construct the category of the combatant. As I argue, when major protagonists of just war theory defend the duties and rights of combatants, they appeal to the supposedly natural characters and social roles of men. This reinforces Kinsella's view of the role of gender in the construction of the combatant/civilian distinction and, I argue, explains the emergence of the external and internal problems of the soldier in just war theory.

A substantial body of literature in gender studies has concluded that there is a prominent construction of masculinity embedded in many cultures that links manhood with military service.[38] According to this construction, by virtue of their sex, it is good and honorable for men to provide protective martial labor in defense of their communities and families. Men as men ought to carry out this type of labor and bear the burdens it entails. If a man fails to provide this labor

[37] Kinsella, *The Image Before the Weapon*.

[38] See Elshtain, *Women and War*; Goldstein, *War and Gender*; Braudy, *From Chivalry to Terrorism*; and Digby, *Love and War*.

either by choice or because he lacks the supposedly appropriate character traits (e.g., he fears being killed or maimed, or he is repulsed by violence), then he has failed ethically. A reliable way to affirm one's manhood across cultures is to be adept in the arts of war and to demonstrate the physical and characterological capacity to engage in battle without fear.

A crucial feature of this construction of masculinity is that while it is nevertheless a social construction, it treats the normative content of masculinity as natural to biological sex. In other words, this social construction of masculinity asserts that it is due to the biological nature of men that they ought to have warrior virtues. Failures of masculine virtue are failures to have or achieve the supposed essence proper to the male sex. As General George Patton, who notoriously assaulted soldiers for suffering from apparent cases of shell shock, said in a speech to his troops, "a real man will never let his fear of death overpower his honor, his sense of duty to his country, and his innate manhood."[39]

However, as the bulk of the gender studies literature concludes, there is no evidence of a natural connection between the male sex and propensity for war or warrior characteristics. Human males seem to be just as naturally inclined to the full array of human emotion, connectivity, and forms of social labor as human females. In fact, the naturalistic vision of masculinity is belied by the obvious and pervasive efforts to enforce masculine norms on men and boys. From a very early age, boys begin to experience social pressure to exhibit toughness and joy in violent activity.[40] Physical and mental strength, as well as expertise in the arts of physical domination, are highly praised in men and boys, whereas tenderness, sensitivity, and any disinclination to violence are shamed. Most men are constantly aware that if they fail to display the appropriate masculine standards of toughness, they can be subjected to abuse of the most homophobic and misogynistic kind. As Goldstein says:

> Cultures produce male warriors by toughening up boys from an early age. . . . Although boys on average are more prone to more rough-and-tumble play, they are not innately "tougher" than girls. They do not have fewer emotions or attachments, or feel less pain. It is obvious from the huge effort that most cultures make to mold "tough" boys that this is not an easy or natural task. When we raise boys within contemporary gender norms, especially when we push boys to toughen up, we pass along authorized forms of masculinity suited to the war system.[41]

[39] Hirshson, *General Patton*, 474.

[40] See Way, *Deep Secrets*; and Chu, *When Boys Become Boys*.

[41] Goldstein, *War and Gender*, 287–88.

This construction of masculinity is a problem for all of us. It is a source of various social, psychological, and interpersonal problems affecting all sexes. In addition to this, it is a source of numerous philosophical problems. The external and internal problems of the soldier are the products of the influence of this notion of masculinity on our moral theory. As we have seen, many of the arguments in defense of the permissibility of attacks against combatants and the subordination of soldiers are unpersuasive. However, when they attempt to directly explain their positions regarding the expendability of soldiers, many canonical just war theorists abandon these arguments and appeal directly to masculinity.

### *2.1. Contractarian Arguments*

While not a canonical just war theorist, it is illuminating to begin by considering Thomas Hobbes's defense of the obligation of soldiers to fight in war on command. Hobbes's struggle to justify this obligation is well documented.[42] What is less well documented is that in his discussion of the problem, he clearly presupposes a warrior masculinity. Hobbes excuses women and feminine men from the duty to serve in war on command. He says that "there is allowance to be made for natural timorousness, not only to women (of whom no such dangerous duty is expected), but also to men of feminine courage."[43] Strikingly, the basis for the distinction between men's and women's duties is a presumption about their divergent natural characters. Women and some men are feminine precisely in the sense that they lack the courage to risk their lives in war on command. While this character trait is natural to women, it is unnatural to men and therefore an ethical failing. Feminine men are cowardly. This presumption helps mitigate the weakness of Hobbes's contract argument for the political obligation to fight in war on command. Interestingly, Hobbes also asserts this gender division in defense of his view that the succession of the throne should go to the monarch's male descendants over his female descendants. Male descendants should inherit the throne because "men are naturally fitter than women, for actions of labor and danger."[44] As we can see, Hobbes thinks men, and not women, are naturally suited to military service.

This is roughly the same approach to defending the subordination of soldiers taken by many canonical just war theorists. While Grotius describes the rights of men as prior to political society and political society as designed to protect those rights, he does not defend the obligation to risk one's life in

[42] See Hampton, *Hobbes and the Social Contract Tradition*; Goldie, "The Reception of Hobbes"; and Sreedhar, "In Harm's Way."

[43] Hobbes, *Leviathan*, 142.

[44] Hobbes, *Leviathan*, 126.

battle as based on the social contract. Rather, he grounds the duty to engage in self-sacrificial military service in natural virtue. He says, “Some Acts of Virtue may by a human law be commanded, though under the evident Hazard of Death. As for a Soldier not to quit his Post. . . .”[45] In this case, Grotius appeals to the virtue of charity. In an early work, however, Grotius appeals more clearly to warrior masculinity. He argues that the virtue expressed by risking one’s life in battle for the state is fortitude. He describes fortitude as one of the two virtues “most beneficial [to others], both in private and in public life.”[46] Grotius then directly connects fortitude with masculinity. He quotes approvingly a passage from the poet Tyrtaeus: “It is a glorious and manly thing,/To risk one’s life in battle with the foe,/Defending loved ones, wife and native land.”[47]

Pufendorf, too, describes the rights of men as prior to political society and political society as a contract made by its male members to protect their natural rights. However, Pufendorf also appeals directly to natural virtue to support the duty to engage in self-sacrificial military action on command. He claims that it would be cowardly for a man to refuse to engage in combat out of fear of injury or death. In fact, Pufendorf claims that a good man will praise his commander for ordering him to risk his life. He says that a soldier “is bound to defend the Post his Commander appoints him to, tho’ perhaps he foresees he must in all probability lose his Life in it. . . . And no man of Bravery or Spirit will ever complain that he is commanded upon such a Duty, but will rather commend his General’s Judgment and Conduct in it.”[48] The character of the “man of Bravery or Spirit” lines up neatly with the manly warrior ideal. In this way, masculinity is serving to ground Pufendorf’s view of the subordination of the soldier.

Vattel is another canonical just war theorist who argues that political society is a voluntary association of its male members to protect their equal natural rights. Yet he too appeals to the same gender division in his discussion of the duties of military service. As he says, “every man capable of carrying arms should take them up at the first order of him who has the power of making war. . . . Although there be some women who are equal to men in strength and courage, yet such instances are not usual: and rules must necessarily be general, and derived from the ordinary course of things.”[49] For Vattel, only men have the political obligation to participate in military service because men are by nature inclined to have the moral character suited to “supporting the fatigues

[45] Grotius, *The Rights of War and Peace*, 357.

[46] Grotius, *Commentary on the Law of Prize and Booty*, 440.

[47] Grotius, *Commentary on the Law of Prize and Booty*, 441.

[48] Pufendorf, *The Law of Nature and Nations*, 567.

[49] Vattel, *The Law of Nations*, 474.

of war." In particular, men tend to have the courage to risk their lives in battle on command. The implication is that for a man to fail to have such a character is a moral failure, one much more severe than the same condition in a woman.

### 2.2. *Walzer's Argument*

Walzer describes his theory in *Just and Unjust Wars* as a social contract theory that grounds the rights of states and the political obligations of citizens in the image of an agreement of mutual protection between free and equal individuals.[50] This appears similar to the early modern approach discussed above but is stripped of its gender exclusivity. However, this account of his theory is misleading. The most substantive discussions Walzer offers of the duty of soldiers to engage in self-sacrificial military service occur in works of his other than *Just and Unjust Wars*. In these works, Walzer expresses skepticism about the ability of traditional social contract theory to ground the self-sacrificial duties of soldiers. For instance, in an early essay, "The Obligation to Die for the State," Walzer endorses Hegel's criticism (quoted above) of the social contract method and instead offers a nonliberal theory of the responsibilities of soldiers. According to this argument, after the formation of the state of which one is a member, a person can find their identity transformed from a private individual to a member of a common life that they share with their fellows. This new identity enables the possibility of obligations of self-sacrificial military service for the sake of the state. As Walzer says, "so long as the state survives, something of the citizen lives on, even after the natural man is dead. That state, or rather, the common life of the citizens, generates these 'moral goods' for which … men can in fact be obligated to die."[51] For Walzer, the self-sacrificial duties of the soldier are grounded in a communitarian theory of justice that prioritizes the state over the individual.

In *Spheres of Justice*, the book he published immediately after *Just and Unjust Wars*, Walzer takes this nonliberal approach to justice even further. He argues that justice is grounded not in the equal rights of abstract individuals but in the shared meanings embedded in communities with a common way of life. For him, justice is relative to the shared understandings of the good rooted in the culture, institutions, and language of each community.

Central to Walzer's theory in *Spheres* is his pluralistic account of the just distribution of goods within distinct communities. Justice is not only relative to specific communities. It is also relative to each good and the particular

[50] Walzer, *Just and Unjust Wars*, 54; see also Walzer, *Spheres of Justice*. For an extended discussion of the role of gender in Walzer's view of soldiers, see Parsons, "Walzer's Soldiers."

[51] Walzer, "The Obligation to Die for the State," 92; see also Walzer, "Involuntary Association."

understanding of it within each community. Walzer distinguishes between the goods of national security, personal security, welfare, and recognition, among others. He argues that each community forms a shared understanding of each of these goods and that within that understanding there is contained an understanding of how the good ought to be distributed. In order for justice to be achieved, each good ought to be distributed in accordance with its own internal meaning. Injustice occurs when the standards of distribution peculiar to one good are applied to other goods. We must not let one good come to dominate or invade other goods. For example, we must not let medical care be distributed in accordance with the standards of justice for commodities in markets. Medical care, according to our shared understanding, is not a commodity, and to treat it as if it were is to allow markets to dominate goods that ought to maintain their autonomy.

In this way, Walzer's communitarianism can ground the duties of military service while also offering material to protect individuals from complete subservience to the state. The good of national security and the burdens of military service can be distributed according to standards that can require self-sacrificial labor from members of the community. However, the good of national security cannot dominate other goods such as personal security, welfare, and recognition. This pluralism of goods and their distribution protects members of the community from being reduced to instruments of the state while permitting the self-sacrificial duties of military service.

However, while this theory of the pluralistic autonomy of the various goods protects some members of the community from domination by the good of national security, it does not protect the individuals who are burdened with military service from such domination. Soldiers are expected to fight and risk death on command. Walzer argues that military service requires self-sacrificial labor and that this labor must be forced, at least once the enlistment contract has been completed.[52] Because it asks everything of its participants, those who provide military service are necessarily dominated by it, especially at the moment the service is provided. The personal security, welfare, liberty, and all other goods of the military servicemember are overridden by the good of national security. For the servicemember, it seems impossible to maintain the autonomy of all the spheres of justice.

Like the way the other just war theorists appeal to gender to overcome the weaknesses in their contractarian arguments for the subordination of the soldier, Walzer's discussion also benefits from implicit appeals to gender to overcome the weaknesses of his communitarian argument for military service.

[52] Walzer, *Spheres of Justice*, 169, 180.

Walzer explicitly rejects the idea that negative goods such as military service can be justly distributed to some subgroup only because of their supposed natural suitability for it. Nevertheless, immediately after asserting that such unequal distributions are unjust, Walzer seems to assert that military service should be restricted only to men due to their nature. As he says,

> Soldiering is a special kind of hard work. In many societies, in fact, it is not conceived to be hard work at all. It is the normal occupation of young men, their social function, into which they are not so much drafted as ritually initiated, and where they find the rewards of camaraderie, excitement, and glory.… Young men are energetic, combative, eager to show off, fighting for them is or can be a form of play.… John Ruskin had a wonderfully romantic account of "consensual war," which aristocratic young men fight in much the same spirit as they might play football. Only the risks are greater, the excitement at a higher pitch, the contest more "beautiful."…We might attempt a more down-to-earth romanticism: young men are soldiers in the same way that the French socialist writer Fourier thought children should be garbagemen. In both cases, passion is harnessed to social function. Children like to play in the dirt, Fourier thought, and so they are more ready than anyone else to collect and dispose of garbage.[53]

Walzer criticizes this view of military labor, but not on the grounds that it unjustifiably burdens men only with the responsibility to perform it on the basis of false assumptions about the nature of men and women. Rather, he criticizes the view that soldiering is similar to play. He argues that such a view ignores the burdensome nature of the work. While Walzer is surely right that military service is dissimilar to play, it is remarkable that he does not take issue with the gender-based account of the distribution of responsibility for military service clearly expressed in the above passage. In fact, in that passage Walzer appears to be endorsing this gendered account. In the following paragraph, he embraces conscription as a method "to universalize or randomize the risks of war over a given generation of *young men*."[54] This gender-exclusive distribution of the burdens of military service appears to be based on the account of natural masculinity offered in the preceding passage.

It is tempting to discount these appeals to gender in Walzer's theory as momentary lapses in an otherwise progressive work that attacks gender-exclusive distributions of goods. Still, the fact that it is easy for Walzer to depart from

[53] Walzer, *Spheres of Justice*, 168–69.

[54] Walzer, *Spheres of Justice*, 169 (emphasis added).

his other settled positions and make such a clear appeal to masculine nature shows us how deeply engrained in our thinking about soldiers and war gender is. But more troublingly, the appeal to masculinity to ground the duties of military service helps Walzer's theory. As we have seen, Walzer's communitarian argument for the distribution of military service is challenged by his insistence that the plurality of goods embedded in the shared meanings of communities must maintain their autonomous spheres; one good cannot come to dominate the others. The problem is that for soldiers, who are obligated to put their lives and liberties on the line for their community, the good of national security necessarily dominates all their other personal goods. Walzer's implicit appeal to the natural character of men as the basis of their military obligations would help mitigate this problem. If by making men responsible for military service we were harnessing passion to social function, then we would minimize the tension between the burdens of military service and the personal good of soldiers. For according to this view of masculinity, the performance of military service would be the personal good of the (male) soldier.

### 2.3. *From the Internal to the External Problem*

As we can see, then, the appeal to natural masculinity has served all these canonical just war theorists. Most immediately, it has helped them solve the internal problem of the soldier. In as much as it is unclear how the self-sacrificial duties of soldiers are reconcilable with the rights or basic interests of individuals, the appeal to masculinity helps fill the gap. Soldiers can have their self-sacrificial responsibilities because they are men whose natural virtue is realized by carrying out those responsibilities regardless of whether those responsibilities can be reconciled with the rights of individuals or the purpose of the political society.

But the appeal to gender to solve the internal problem of the soldier also helps solve the external problem as well. As we have seen, traditional just war theorists have struggled to straightforwardly defend the moral equality of combatants. The gender-based justification of the duty to serve in war on command helps solve this problem. If the reason soldiers are bound to serve in war on command is that, as men, it is their natural duty to engage in military service on command and even under the danger of death, then they are being conceived of as expendable instruments of war prior to the initiation of any particular war. A person who does not have the standing to refuse to obey an order that puts their life in danger, even the most acute danger imaginable, is a person who does not have a right to life, at least not a right that is able to offer any meaningful protection of their life in a time of war. Combatants understood this way are not people who begin with a right to life and then voluntarily waive that right by accepting danger to themselves. Rather, they are people whose purpose is to engage in violent combat with others no matter the risk to themselves. From this perspective, we simply do not presume combatants have a right to life that needs to be overridden to justify their engagement in combat.

It is true that being bound to face the danger of death in combat is not equivalent to the right to kill others in combat. It is possible to imagine a normative world where combatants on all sides are bound by nature to face the fire of their opponents, but none have a right to dispense fire on their opponents. How, then, does the masculine gender ontology bridge this gap and justify not only using combatants as instruments of defense in war but also giving them a permission to attack and kill their opponents?[55]

The masculine gender ontology links the duty to face death in war with the right to kill in war by construing the natural purpose of men to be to risk their lives not in any activity that protects their communities or families but *in violent combat*. On this view, manhood is tied not to any self-sacrificial activity for the sake of others. It is tied specifically to self-sacrificial violence. As the poem Grotius relies on puts it, "It is a glorious and manly thing,/To risk one's life *in battle with the foe*,/Defending loved ones, wife and native land."[56] Hence, the performance of violence is a central component of this masculine good. It is not achieved in decidedly nonviolent self-sacrificial activities such as nursing during a pandemic, carrying a child to term with inadequate access to healthcare, or working for the International Committee of the Red Cross in a warzone without a personal security detail. It is through engaging in violent combat that self-sacrificial labor comes to affirm manhood most effectively. Attacking and killing others in combat is a central part of this vision of masculine virtue.

In this way, the gender ontology that just war theorists presuppose to solve the internal problem also solves the external problem. Men as men are bound both to risk their lives for the sake of their communities and families in war and to attack and kill the combatants representing their opponents. Not only is the self-sacrifice good for men; so is the dispensing of violence upon others. Combatants, therefore, are conceived from the beginning as lacking a right to refuse self-sacrificial orders in war and possessing the right to attack and kill other combatants. They are people who are reduced entirely to the status of combatants—those who *fight and die* in war.

That the assumption of the expendability of soldiers is helping to ground the permissive view of killing combatants in war is illustrated in a more recent essay by Walzer. In "Terrorism and Just War," Walzer defends the permission to

[55] I am grateful to an anonymous referee for this journal for posing this question to me.

[56] Grotius, *Commentary on the Law of Prize and Booty*, 441 (emphasis added).

kill combatants in war "at random."[57] Central to his argument is the assertion of a fundamental difference between soldiers and civilians. As Walzer describes it, the singular purpose of soldiers is to fight wars, whereas civilians have varied other purposes. According to Walzer,

> the army is an organized, disciplined, trained, and highly purposeful collective, and all its members contribute to the achievement of its ends. Even soldiers who don't carry weapons have been taught how to use them, and they are tightly connected, by way of the services they provide, to the actual users. It doesn't matter whether they are volunteers or conscripts; their individual moral preferences are not at issue; they have been mobilized for a singular purpose, and what they do advances that purpose. For its sake, they are isolated from the general public, housed in camps and bases, all their needs provided for by the state. In time of war they pose a unified threat.[58]

Civilians, on the other hand, are quite different:

> Civilians have many different purposes; they have been trained in many different pursuits and professions; they participate in a highly differentiated set of organizations and associations, whose internal discipline, compared to that of any army, is commonly very loose. They don't live in barracks but in their own houses and apartments; they don't live with other soldiers but with parents, spouses, and children; they are not all of an age but include the very old and the very young; they are not provided for by the government but provide for themselves and one another. As citizens, they belong to different political parties; they have different views on public issues; many of them take no part at all in political life; and, again, some of them are children. Even a *levée en masse* cannot transform this group of people into anything like an organized military collective.[59]

Based in part on this distinction between soldiers and civilians, Walzer concludes that there is a blanket permission to kill combatants in war whereas noncombatants are immune from attack. But Walzer's account of what makes civilians unlike soldiers is certainly false. Soldiers, like civilians, have multiple social roles and engage in the full spectrum of social activities that civilians do. The person who is a soldier is also many other things. Soldiers too live in

[57] Walzer, "Terrorism and Just War," 264–65.

[58] Walzer, "Terrorism and Just War," 265.

[59] Walzer, "Terrorism and Just War," 265–66.

houses and apartments with parents, spouses, and children. Soldiers too have many different views on public issues or no view at all. Soldiers too participate in a wide variety of private organizations and associations with no connection whatsoever to the military. Soldiers are, in other words, people in just the same way that civilians are. And just like civilians, they cannot be reduced to any particular office.

Walzer's argument for the blanket permission to attack combatants presumes otherwise. For him, combatants in war are fair game for attack because combatants are simply inseparable from their office. That this obviously false position can make its way so explicitly into this argument, I submit, is explained by the assumption of the masculine nature of soldiers. As we have seen, according to this assumption, the labor of combat is to be performed by those who are bound to engage in combat because of their sexual nature. The combatant and the individual in the role of combatant are inseparable; combatants are conceived of as natural combatants. While he does not appeal to masculinity directly, Walzer appears to embrace this reduction of the military servicemember to their office, and he uses this position to defend the right to attack combatants in war. In this way, Walzer uses the resources provided by the masculine gender ontology to solve the external problem of the soldier.

## 3. RETHINKING THE SOLDIER

Revisionist just war theorists have pushed us to reconsider traditional assumptions about the equal right to kill combatants in war and, in so doing, emphasized the responsibilities of soldiers for the wars they fight. Nevertheless, the debate over the moral equality of combatants in contemporary just war theory has not appreciated the full extent of the subordination of soldiers in theory and in practice. Recognizing the origin of the notion of the moral equality of combatants in just war theory reveals the need for a more fundamental reengagement with the political status of soldiers and their relationship to civil society and the state. Importantly, the need for this more fundamental rethinking has been recognized by feminist theorists for some time.

For instance, in a summary of feminist criticisms of just war theory, Peach argues that just war theory relies on an abstract conception of people and that this enables the theory to construe enemy forces as dehumanized "Others" who can readily be killed in war.[60] Moreover, Peach argues that feminists rightly criticize just war theory for subordinating soldiers to their political communities and the commands of political authorities. She argues that this subordination

[60] Peach, "An Alternative to Pacifism?," 159.

contributes to the view of enemies as mere instruments, thereby making it easier to justify killing them.[61] In this way, Peach argues that what I am calling the external and internal problems are connected and that they have been central concerns for many feminist theorists of the ethics of war. According to her, these problems arise because of the tendency of just war theory to rely on an abstract vision of the person.

As Peach concludes,

> a feminist approach to just-war theory would entail reformulated understandings of the proper relationship between the individual and the state. It would consider both the impact of war on individuals as well as the obligations of both men and women to defend the nation. It should provide a formulation with which the merits of a particular military engagement may be assessed by the individual soldiers and civilians involved in it as well by the relevant "authorities."… It would include a reassessment of women's exemption from military combat and draft registration, as well as established laws governing conscientious objection and civil disobedience.[62]

I hope that this article has shown anyone engaged in military ethics that this approach is a worthy one. That said, my argument identifies a source of the problem of the status of combatants that is different from Peach's. According to traditional just war theorists, the unique duties of the military servicemember are identical to the natural duties of the individuals who are supposed to occupy the office. This reduction of the individual to the office is accomplished because of the presupposition that they are men whose duties and virtues are determined by a gender ontology that prescribes for them the role of military servant. In this way, rather than treating the soldier as an abstract person, just war theory has reduced the individual occupying the office of soldier to the office itself. The soldier has been conceived of as a man whose natural obligations bind him to carry out the duties of his office even under the danger of death, thereby making him expendable in war.

Once we consciously reject this picture of the combatant, the permissive view of killing combatants in war, as well as the subordinate political status of soldiers, should also become untenable. Once we stop thinking of people who fill the ranks of militaries as nothing more than soldiers, people who are merely instruments of their country's security, we will need to rethink the status of combatants at all levels. As this article has emphasized, soldiers are not just seen

[61] Peach, "An Alternative to Pacifism?" 161.

[62] Peach, "An Alternative to Pacifism?" 167.

as expendable externally (i.e., when confronting them as enemies in war); they are also seen as expendable internally (i.e., when we create militaries for the purposes of confronting enemies). In fact, it is the internal expendability that has helped ground the external expendability. In our ethics and laws, soldiers are denied civic equality with civilians. Their country's interest in national security overrides soldiers' private interests, even their interest in survival.

Ending this internal subordination will require fundamental changes to the way we conceive of the military. The revisionist literature on the liability of combatants to be killed in war has led to criticism of conscientious objection laws in many countries. It has created a push to embrace selective conscientious objection, that is, the legal option to apply for conscientious objector status for specific wars or campaigns, not simply conscientious objector status for all wars. While this is a step in the right direction, it does not address the full depth of the subordination of the soldier. Selective conscientious objection rights give soldiers more liberty to refuse orders on the grounds that the orders are immoral. However, such entitlements do not recognize the right of soldiers to refuse an order or to leave the profession because it conflicts with their personal interests. Most strikingly, even with conscientious objection rights, soldiers are not permitted to disobey an order to avoid death or injury or to leave the profession at will. What we need to do is recognize the full scope of the rights and interests of the individuals who serve as soldiers and bring them fully into line with the status of civilian employees. Simply put, we need to recognize that the people who happen to be soldiers are more fundamentally *people* with lives and interests that can transcend the interests of the state.

Moreover, this effort is linked with the ongoing battle to end gender exclusions and discrimination within the military. That integrating the military with genders and sexualities other than cisgender heterosexual men has proven harder than it has in other institutions should not be surprising. According to the argument of this article, gender exclusivity and subordination in the armed forces have the same origin. As we reconsider the office of military servicemember and its relation to its officeholders, we need to simultaneously appreciate the imperative of gender inclusivity. In fact, achieving a truly gender-integrated armed force requires abandoning the picture of the servicemember as expendable.[63]

*United States Military Academy*
*graham.parsons@westpoint.edu*

[63] Disclaimer: the views expressed in this article are those of the author and do not represent the views of the United States Military Academy, the United States Army, or the Department of Defense.

## REFERENCES

Allhoff, Fritz, Nicholas Evans, and Adam Henschke, eds. *The Routledge Handbook of Ethics and War*. New York: Routledge, 2013.

Atkinson, Rick. *Crusade: The Untold Story of the Persian Gulf War*. New York: Houghton Mifflin Co., 1993.

Blum, Gabriella. "The Dispensable Lives of Soldiers." *Journal of Legal Analysis* 69, no. 2 (Spring 2010): 69–124.

Braudy, Leo. *From Chivalry to Terrorism: War and the Changing Nature of Masculinity*. New York: Vintage, 2005.

Chu, Judy. *When Boys Become Boys: Development, Relationships, and Masculinity*. New York: NYU Press, 2014.

Darwall, Stephen. "Grotius at the Creation of Modern Moral Philosophy." *Archiv für Geschichte der Philosophie* 94, no. 3 (2012): 296–325.

Digby, Tom. *Love and War: How Militarism Shapes Sexuality and Romance*. New York: Columbia University Press, 2014.

Dobos, Ned. "Punishing Non-Conscientious Disobedience: Is the Military a Rogue Employer?" *Philosophical Forum* 46, no. 1 (Spring 2015): 105–19.

Drakopoulou, Maria. "Samuel Pufendorf, Feminism, and the Question of 'Women and Law.'" In *Feminist Encounters with Legal Philosophy*, edited by Maria Drakopoulou, 66–91. New York: Routledge, 2015.

Draper, Kai. *War and Individual Rights: The Foundations of Just War Theory*. New York: Oxford University Press, 2015.

Elshtain, Jean Bethke. *Women and War*. Chicago: University of Chicago Press, 1995.

Enloe, Cynthia. *Maneuvers: The International Politics of Militarizing Women's Lives*. Berkeley: University of California Press, 2000.

Frowe, Helen. *Defensive Killing*. New York: Oxford University Press, 2014.

Frowe, Helen, and Seth Lazar, eds. *The Oxford Handbook of Ethics of War*. New York: Oxford University Press, 2018.

Gardam, Judith. "Women and the Law of Armed Conflict: Why the Silence?" *International and Comparative Law Quarterly* 46, no. 1 (January 1997): 55–80.

Goldie, Mark. "The Reception of Hobbes." In *The Cambridge History of Political Thought: 1450–1700*, edited by J. H. Burns, 589–615. New York: Cambridge University Press, 1991.

Goldstein, Joshua. *War and Gender*. New York: Cambridge University Press, 2001.

Grotius, Hugo. *Commentary on the Law of Prize and Booty*. Edited by Martine Julia van Ittersum. Translated by Gwladys L. Williams. Indianapolis: Liberty Fund, 2006.

———. *The Rights of War and Peace,* edited by Richard Tuck. Indianapolis: Liberty Fund, 2005.

Haakonssen, Knud. "Hugo Grotius and the History of Political Thought." *Political Theory* 13, no. 2 (May 1985): 239–65.

Hampton, Jean. *Hobbes and the Social Contract Tradition*. New York: Cambridge University Press, 1988.

Haque, Adil. *Law and Morality at War*. New York: Oxford University Press, 2017.

Hegel, G. W. F. *Elements of the Philosophy of Right*. Edited by Allen W. Wood. Translated by H. B. Nisbet. New York: Cambridge University Press, 1991.

Held, Virginia. *How Terrorism is Wrong: Morality and Political Violence*. New York: Oxford University Press, 2008.

Hirshson, Stanley. *General Patton: A Soldier's Life*. New York: HarperCollins, 2002.

Hobbes, Thomas. *Leviathan*. Edited Edwin Curley. Indianapolis: Hackett Publishing Company, 1994.

Hutchings, Kimberly. "Making Sense of Masculinity and War." *Men and Masculinities* 10, no. 4 (June 2008): 389–404.

Johnson, James Turner, and James Pattison, eds. *The Ashgate Research Companion to Military Ethics*. Burlington: Ashgate, 2015.

Kant, Immanuel. "Perpetual Peace: A Philosophical Sketch." In *Kant: Political Writings,* edited by H. S. Reiss, 93–130. New York: Cambridge University Press, 1991.

Kinsella, Helen. *The Image Before the Weapon: A Critical History of the Distinction between Combatants and Civilians*. Ithaca, NY: Cornell University Press, 2011.

Lazar, Seth. "Just War Theory: Revisionists versus Traditionalists." *Annual Review of Political Science* 20 (2017): 37–54.

———. "Necessity in Self-Defense and War." *Philosophy and Public Affairs* 40, no. 1 (Winter 2012): 3–44.

Lucas, George, ed. *The Routledge Handbook of Military Ethics*. New York: Routledge, 2015.

MacIntyre, Alistair. "Is Patriotism a Virtue?" In *Global Ethics: Seminal Essays,* edited by T. Pogge and K. Horton, 119–38. St. Paul, MN: Paragon House, 2008.

Mann, Bonnie. *Sovereign Masculinity: Gender Lessons from the War on Terror*. New York: Oxford University Press, 2014.

Mapel, David. "Coerced Moral Agents? Individual Responsibility for Military Service." *The Journal of Political Philosophy* 6, no. 2 (June 1998): 171–89.

May, Larry, ed. *The Cambridge Handbook of the Just War*. New York: Cambridge University Press, 2018.

McMahan, Jeff. "Innocence, Self-Defense, and Killing in War." *Journal of Political*

*Philosophy* 2, no. 3 (September 1994): 193–221.

———. *Killing in War*. New York: Oxford University Press, 2009.

———. “The Morality of War and the Law of War.” In *Just and Unjust Warriors,* edited by David Rodin and Henry Shue, 19–43. New York: Oxford University Press, 2008.

———. “Proportionate Defense.” In *Weighing Lives in War,* edited by J. D. Ohlin, L. May, and C. Finkelstein, 131–54. New York: Oxford University Press, 2017.

McPherson, Lionel. “Innocence and Responsibility in War.” *Canadian Journal of Philosophy* 34, no. 4 (December 2004): 485–506.

Nunn, Sam. “The Fundamental Principles of the Supreme Court’s Jurisprudence in Military Cases.” In *Evolving Military Justice,* edited by Eugene R. Fidell and Dwight H. Sullivan, 3–14. Annapolis, MD: Naval Institute Press, 2002.

Ohlin, Jens David. “Sharp Wars Are Brief.” In *Weighing Lives in War,* edited by J. D. Ohlin, L. May, and C. Finkelstein, 58–76. New York: Oxford University Press, 2017.

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

Parsons, Graham. “Contract, Gender, and the Emergence of the Civil-Military Distinction.” *The Review of Politics* 82, no. 3 (Summer 2020): 416–37.

———. “Walzer’s Soldiers: Gender and the Rights of Combatants.” In *Walzer and War: Reading “Just and Unjust Wars” Today,* edited by Graham Parsons and Mark Wilson, 241–67. New York: Palgrave, 2020.

———. “What Is the Classical Theory of Just Cause? A Response to Gregory Reichberg.” *Journal of Military Ethics* 12, no. 4 (2013): 357–69.

Pateman, Carole. *The Sexual Contract*. Stanford, CA: Stanford University Press, 1988.

Peach, Lucinda J. “An Alternative to Pacifism? Feminism and Just War Theory.” *Hypatia* 9, no. 2 (May 1994): 152–72.

Primoratz, Igor. “Michael Walzer’s Just War Theory: Some Issues of Responsibility.” *Ethical Theory and Moral Practice* 5, no. 2 (June 2002): 221–43.

Pufendorf, Samuel von. *The Law of Nature and Nations*. Translated by B. Kennett. Clark, NJ: The Lawbook Exchange, 2005.

Robinson, Fiona. *The Ethics of Care: A Feminist Approach to Human Security*. Philadelphia: Temple University Press, 2011.

Rodin, David. *War and Self-Defense*. New York: Oxford University Press, 2002.

Ruddick, Sara. *Maternal Thinking: Toward a Politics of Peace*. Boston: Beacon Press, 1989.

Schneewind, J. B. *The Invention of Autonomy*. New York: Cambridge University Press, 1998.

Sjoberg, Laura. *Gender, Justice, and the Wars in Iraq: A Feminist Reformulation of Just War Theory*. Lanham, MD: Lexington Books, 2006.

———. *Gender, War, and Conflict*. Malden MA: Polity Press, 2014.

Sreedhar, Susan. "In Harm's Way: Hobbes on the Duty to Fight for One's Country." In *Hobbes Today: Insights for the 21st Century*, edited S. A. Lloyd, 209–28. New York: Cambridge University Press, 2012.

———. "Pufendorf on Patriarchy." *History of Philosophy Quarterly* 31, no. 3 (July 2019): 209–27.

Suarez, Francisco de. *Selections from Three Works*. Edited by Thomas Pink. Indianapolis: Liberty Fund, 2013.

Tadros, Victor. *To Do, To Die, To Reason Why: Individual Ethics in War*. New York: Oxford University Press, 2020.

Tickner, J. Ann. *Gender in International Relations: Feminist Perspectives on Achieving Global Security*. New York: Columbia University Press, 1992.

Tuck, Richard. *Natural Rights Theories*. New York: Cambridge University Press, 1979.

———. *The Rights of War and Peace*. New York: Oxford University Press, 1999.

US Department of Defense. Enlistment/Reenlistment Document—Armed Forces of the United States. January 2007.

Vattel, Emer de. *The Law of Nations*. Edited by Béla Kapossy and Richard Whatmore. Indianapolis: Liberty Fund, 2008.

Vitoria, Francisco de. *Political Writings*. Edited by Anthony Pagden and Jeremy Lawrance. New York: Cambridge University Press, 1991.

Walzer, Michael. "Involuntary Association." In *Freedom of Association*, edited by Amy Gutmann, 64–74. Princeton, NJ: Princeton University Press, 1998.

———. *Just and Unjust Wars: A Moral Argument with Historical Illustrations*. 5th ed. New York: Basic Books, 2015.

———. "The Obligation to Die for the State." In *Obligations: Essays on Disobedience, War, and Citizenship*, 77–98. Cambridge, MA: Harvard University Press, 1970.

———. *Spheres of Justice: A Defense of Pluralism and Equality*. New York: Basic Books, 1983.

———. "Terrorism and Just War." In *Thinking Politically: Essays in Political Theory*, edited by David Miller, 264–77. New Haven, CT: Yale University Press, 2007.

Way, Niobe. *Deep Secrets: Boys' Friendships and the Crisis of Connection*. Cambridge, MA: Harvard University Press, 2013.

Young, Iris Marion. "The Logic of Masculinist Protection: Reflections on the Current Security State." *Signs* 29, no. 1 (Autumn 2003): 1–25.

*Journal of Ethics and Social Philosophy*
vol. 26, no. 2 · december 2023

*https://doi.org/10.26556/jesp.v26i2.2733*

# CONTEXTUALIZING, CLARIFYING, AND DEFENDING THE DOCTRINE OF DOUBLE EFFECT

*Melissa Moschella*

According to the doctrine of double effect (DDE), bad effects that it is not morally permissible to intend may be permissibly accepted as side effects if the good effects one intends are proportionate to the unintended bad side effects. The DDE is often invoked to explain, for instance, why (as is commonly supposed) it may be morally permissible to engage in tactical bombing of a legitimate military target in a just war, despite foreseeing that nearby civilians will be killed, but not permissible to engage in terror bombing, that is, directly targeting civilians to terrorize the population and undermine morale. Many, however, have attacked the DDE, arguing that it is incoherent, lacks an underlying rationale, or leads to apparently absurd conclusions in certain cases. Stephen Kershnar and Robert Kelly, for instance, argue in a recent paper that the principle of double effect is incoherent, because the principle presumes that people have stringent moral rights and (in their view) violations of rights should be defined without reference to the intent of the agent.[1] Uwe Steinhoff also criticizes the DDE in two recent papers. In one, he surveys various purported rationales for the DDE and argues that none are convincing.[2] In another, Steinhoff argues that the intuitive appeal of the DDE is based on biased and methodologically flawed presentations of contrasting cases, such as tactical bombing versus terror bombing. In such cases, Steinhoff believes that differences other than the agent's intent are actually the source of our contrasting judgments and that the typical case comparisons obscure this by failing to make the contrasting cases the same in every respect apart from the agent's intent. When cases are presented in which everything except for agent intent is held constant, Steinhoff sees no difference in intuitive moral permissibility.[3]

[1] Kershnar and Kelly, "The Right-Based Criticism of the Doctrine of Double Effect."

[2] Steinhoff, "Wild Goose Chase."

[3] Steinhoff, "The Secret to the Success of the Doctrine of Double Effect (and Related Principles)."

T. M. Scanlon similarly argues that the DDE's appeal is "illusory" and that agent intent cannot in itself determine moral permissibility.[4]

These and many other criticisms of the DDE rest on a failure to understand the principle's broader theoretical context and presuppositions. It should not be surprising that if one isolates a principle from the broader moral theory of which it is a part (and within which the principle was first formulated), that principle will end up seeming arbitrary and difficult to defend. Analogously, Newton's second law of motion (force equals mass times acceleration) would be puzzling to someone who stumbled upon the formula but did not understand the broader theory of physics of which it forms a part.[5]

In this paper, I aim to clarify and advance the debate surrounding the DDE in three stages. First, I outline a contemporary version of the broader normative theory (i.e., the Aristotelian-Thomistic natural law tradition) within which the DDE finds its proper context and explain how this theory provides a rationale for the DDE. Second, I clarify the DDE's proportionality condition to avoid common misinterpretations. Third, I show how recent criticisms of the DDE fail insofar as they attack a straw man quite different from the "real" DDE (properly formulated and understood within the appropriate theoretical context).

The reader may be tempted to dismiss my argument out of hand because it relies on a moral theory that seems "out of fashion" among academic philosophers, but this would be a mistake for several reasons. First, the theory just might turn out to be true or at least better than the alternatives, but one will never know unless one seriously considers it with an open mind. Second, if (as I believe to be the case) the DDE is inextricably bound up with this broader theory, understanding that is crucial for debates about the DDE. Finally, the DDE's persistent intuitive appeal, in spite of the many supposedly decisive refutations of it, may itself provide a reason to reconsider the only tradition of moral thought that can actually provide a coherent rationale for it.[6]

## 1. THE DDE'S THEORETICAL CONTEXT

Historically, the DDE has typically been traced to Thomas Aquinas's discussion of self-defense in the *Summa Theologiae*. In this passage, Aquinas claims that while

[4] Scanlon, *Moral Dimensions*, ch. 1.

[5] This comparison is inspired by Alasdair MacIntyre's argument in chapter 1 of *After Virtue*.

[6] I take as evidence of this persistent appeal the fact that many ethicists frequently refer to the DDE and/or rely upon it and that even its critics refer to its intuitive appeal and think it necessary to provide lengthy arguments against it. Further, empirical studies indicate that it has a strong intuitive appeal for many people. See, e.g., Cushman, "The Psychological Origins of the Doctrine of Double Effect."

intentional killing is always wrong (at least for private citizens), self-defense (even by lethal means) can be morally acceptable insofar as the attacker's death is an unintended side effect of an act aimed at saving one's own life.[7] The term "double effect" comes from Aquinas's claim that "nothing hinders one act from having two effects, one of which is intended, while the other is beside the intention." The reason why this is morally relevant, in Aquinas's view, is that "moral acts take their species according to what is intended, and not according to what is beside the intention."[8] In other words, Aquinas believes that the agent's intent is crucial for determining the nature of an action in the morally relevant sense, and determining the nature (or species) of an action with precision is important because Aquinas believes that certain types of actions are intrinsically evil and therefore always prohibited regardless of the circumstances. (Note that this does *not* imply that foreseen side effects are morally irrelevant, as I will explain below.)

While Aquinas himself does not offer an explicit defense of this claim about the importance of intent in defining the moral act, so-called "new natural law" theorists have filled this gap in the argument by explaining its place within a broader moral theory that recognizes the existence of absolute moral prohibitions on certain types of actions.[9] According to the new natural law theory (NNLT) developed by Germain Grisez, Joseph Boyle, and John Finnis, morality is about acting in accordance with the requirements of practical reason. The first principles of practical reason direct us to preserve and promote basic human goods, understood as constitutive aspects of human flourishing that are intrinsically valuable and provide ultimate reasons for action. These goods are recognized as such by an immediate (nondeductive) insight of practical reason once they are experienced.[10] They include life and health, knowledge,

[7] While Aquinas states in this passage that public officials may engage in intentional killing for the sake of the public good (as would be the case with capital punishment or a soldier killing in war), Germain Grisez argues that Aquinas's views on capital punishment and killing in war are not really coherent with his broader moral theory but are perhaps due to the influence of the dominant opinions of his day. Grisez argues that a coherent Thomistic view would only allow for nonintentional killing. Thus, capital punishment and killing in war could only be justified as forms of community self-defense in which the killing is not strictly intended (Grisez, "Toward a Consistent Natural-Law Ethics of Killing," 91).

[8] Aquinas, *Summa Theologiae,* II-II, q. 64, a. 7, corpus.

[9] The label was given by the theory's critics, who coined it to imply (mistakenly, in my view) that the theory marked a substantive (and erroneous) departure from the Thomistic natural law tradition. However, the label is now widely known and used by both friends and critics of the theory alike, so I use it here for the sake of convenience.

[10] Finnis further explains how basic human goods are identified:

> There are many objects of human interest, but many of them make sense only as instrumental to, or parasitic on, the realization of other, more basic purposes

the appreciation of beauty, excellence in work and play, friendship, marriage, integrity, and religion.[11] Any action that directly or indirectly aims at a basic good is *rational*, for the identification of basic goods just is the identification of intelligible motivations for action.[12] However, an act aimed at a basic good may not be *fully reasonable*, because it may be following *one* of the directives of practical reason (i.e., aiming at one basic good) in a way that fails to respect or take into consideration the *other* directives of practical reason (identifying other goods as also worthy of protection and promotion). For instance, a medical researcher might seek to gain knowledge of a deadly pathogen by deliberately infecting people with it and observing the course of the disease. The researcher's action is rational insofar as it is aimed at the basic good of knowledge, but it is not fully reasonable insofar as it completely fails to respect the health and life of the research subjects (which are also basic goods that practical reason directs us to protect and promote). To act in a fully reasonable way—that is, to act morally—is, in this view, to act in accordance with the *integral directiveness* of practical reason, which means choosing and acting in a way that is compatible with a will toward integral human fulfillment (the fulfillment of all human beings with respect to all of the basic goods). This is one way to formulate the master moral principle of NNLT, from which flow all of the theory's more specific moral norms.[13]

> and benefits. By reflectively analyzing human volitions—one's own and other people's—with their intelligible objects, one can uncover a number of basic purposes, basic benefits of human action, basic human goods. Each of these is an irreducible aspect of the fulfillment of human persons and is instantiated in inexhaustibly many ways in the lives of human persons. (Finnis, *Moral Absolutes*, 42)

[11] Pleasure is not a basic good. A full explanation of why this is the case is beyond the scope of this paper, but the basic reason is the Aristotelian insight that pleasure supervenes on activity and that the goodness of pleasure depends on the goodness of the activity that produces it. Thus, pleasure's value is extrinsic, while basic goods have intrinsic value. See Lee and George, *Body-Self Dualism in Contemporary Ethics and Politics*, ch. 3. See also Finnis, *Natural Law and Natural Rights*, 95–97 and corresponding notes in the postscript.

[12] Because each of the basic goods offers a distinct intelligible benefit, none of the goods is reducible to any of the others or to any more fundamental category of goodness, which means that goods are incommensurable in value. This point will be discussed in greater detail in the next section, as it is crucial for a proper understanding of the proportionality condition.

[13] Another way to formulate this principle is to say that acting morally is acting with unfettered reason. This formulation recognizes that various forms of emotional motivation can be out of line with rational motivation, thus "impairing the rational guidance of action, fettering one's reason, limiting its directiveness, and harnessing it as feeling's ingenious servant." It complements the formulation stated above, for subrational motivations out of line with the directives of practical reason fetter reason precisely "by deflecting one

Among the more specific moral norms that flow from this master moral principle is the norm forbidding intentional damage to or destruction of basic human goods. A choice to damage or destroy a basic human good (either as a means to some further end or as an end in itself) sets one's will against that good and thus is clearly incompatible with a will toward integral human fulfillment. This is not the only moral norm that flows from NNLT's master moral principle. One example is the requirement of fairness, which prohibits arbitrarily prioritizing the good(s) of one person or group over another. Another is reasonable fidelity to commitments, which is required for any deep and significant participation in human goods (where such fidelity excludes fanaticism, which would unreasonably discount the importance of competing goods). However, the norm forbidding intentional damage to basic human goods is especially relevant for understanding the DDE, for it is this norm that identifies certain types of actions (specified by the agent's intent) as always immoral, regardless of the circumstances.

At this point, a clarification of what is meant by "intent" is in order. In simple terms, to intend something is to choose it as an end or as a means to the achievement of one's end. What this means can be clarified with reference to the idea of a proposal for action. In choosing to do something, one adopts a proposal for action that includes the end one is seeking to bring about and all of the means one takes to be necessary in one's plan for the achievement of that end. This account of intention is narrow, including *only* what falls within one's proposal for action so described. Any other effects of one's action fall outside of the intention, even if they are causally inseparable from what one intends and even if they are foreseen with certainty.[14] For instance, a soldier who jumps on a grenade seeking to save his comrades by muffling the impact of the blast with his body adopts a proposal for action in which saving the lives of his comrades is the end and muffling the impact of the blast by absorbing the shrapnel with his body is the chosen means necessary in his plan for the achievement of that end. That the absorption of the shrapnel with his body is almost certainly causally inseparable from his own death does not mean that the soldier intends his death. For it is not his death as such but the muffling of the blast and absorption of the shrapnel that he chooses as necessary means

to objectives not in line with *integral human fulfillment*, with the good of all persons and communities" (Finnis, *Moral Absolutes*, 44–45).

[14] For further explanation and defense of this narrow account of intentions, see Finnis, Grisez, and Boyle, "Direct and Indirect," 1–44; Tollefsen, "A First Person Account of Human Action," 441–60; and Lee, "Distinguishing between What Is Intended and Foreseen Side Effects," 1–21.

to achieve his end. Only the muffling and absorption, therefore, fall within the proposal for action he adopts.

Having clarified the concept of intention, we can now consider the crucial question: Why does NNLT prohibit only *intentional* damage or destruction to basic human goods, despite recognizing (as will be explained further below) that we nonetheless do have moral responsibility for the foreseeable but unintended side effects of our actions? Several reasons can be given for this. First, the theory holds that the capacity for free choices—choices in which nothing but the choice itself determines what will be chosen—is what makes us moral agents. What we *choose* to do determines our moral character in a way that what we merely accept as a side effect does not. For to choose (i.e., intend) something is to set one's will on that thing, to integrate one's will with it, and to treat it as good, such that if one does not achieve it, one will have failed. As Finnis explains, "In choice, then, the human person integrates himself around, and in a certain sense synthesizes himself with, the object of his choice."[15] The agent's self-determination toward the object of choice and resulting self-constitution around that object persist unless and until a contrary choice is made. But to accept something as a side effect does not implicate one's self-constitution in this way.

Grisez elaborates on this point in explaining the crucial moral difference between intending to kill and accepting someone's death as a side effect of one's action:

> A difference of intention can relate identical behavior in quite different ways to our moral attitude, and to the self being created through our moral attitude. If one intends to kill another, he accepts the identity of killer as an aspect of his moral self. If he is to be a killer through his own self-determination, he must regard himself in any situation as the lord of life and of death. The good of life must be rated as a measurable value, not as an immeasurable dignity. Others' natural attitudes toward their own lives must be regarded as an irrational fact, not as a starting point for reasonable community. However, if one intends not the death of another but only the safety of his own life, then one need not identify himself as a killer. One's attitude toward human life itself and toward everything related to it can remain that of a person unwilling to take human life.[16]

[15] Finnis, "Human Acts," 149.

[16] Grisez, "Toward a Consistent Natural Law Ethics of Killing," 76. I should emphasize here that my claims about the nature of intention and its impact on self-determination are not empirical but conceptual and metaphysical, based on the nature of the will and of the act of choosing. Further explanation and defense of these presuppositions is beyond the scope

Note that this integration of the will with the object of one's choice occurs not only when one is choosing something as an end in itself but also when one is choosing something only as a means to some further end, even when that choice of means is made with deep emotional regret (as a spy might regretfully choose to kill his girlfriend as a means to keeping his true identity a secret if she accidentally discovers who he really is).

Second, and related, the prohibitions on what we intend are stricter than prohibitions on what we accept as a side effect, because what we choose to do—what we intend or set our will on in choosing—is by definition completely within our control. By contrast, the transitive effects of our actions on the world are not entirely within our control, for they depend upon a variety of external causal forces as well as upon the free choices of other agents. Thus, it makes sense that the most stringent moral rules should be rules that govern what we choose/intend, by contrast with the unchosen (even if foreseen) effects of our actions.

Third, avoiding unintended damage to or destruction of basic human goods is literally impossible, for free choice presupposes that we are choosing among a variety of basic goods, each of which offers a distinct benefit that is not reducible to any of the others or to any more basic category of goodness.[17] (If all goods were reducible to some common measure, then there would always be one option that offered all of the benefits that the other options offer, plus more. And if that were the case, the other options would lose all rational appeal and fall away, leaving no room for choice.) Anytime we choose, therefore, we are accepting as a side effect the failure to pursue the other goods that we might have pursued if we had chosen otherwise. If, for instance, I decline an invitation to dine with friends in order to spend the evening working on this paper, I accept that I have missed an opportunity to build up our friendship and that my friends may feel snubbed or neglected by my failure to make time for them. I would thus indirectly (and unintentionally) be damaging the good of friendship as a side effect of my choice to pursue the goods of knowledge and professional excellence. Scarcity of time and resources means that this sort of indirect damage to the goods that could have been but were not chosen is a pervasive and unavoidable feature of life. Any moral theory that did not limit absolute moral

of this paper, but they can be found in, e.g., Lee and George, *Body-Self Dualism in Contemporary Ethics and Politics*, 59–65; Finnis, "Human Acts" and "Intention and Side-Effects"; and Finnis, Boyle, and Grisez, *Nuclear Deterrence, Morality and Realism*, 288–91.

[17] As Finnis explains, "There is a free choice (in the sense that matters morally) only when one is rationally motivated towards incompatible alternative possible purposes (*X* and *Y*, or *X* and not-*X*) which one considers desirable by reason of the intelligible goods (instrumental and basic) which they offer—and when nothing but one's choosing itself *settles* which alternative is chosen" (Finnis, "Intention and Side Effects," 194).

prohibitions to actions defined by intention would therefore be impossible to follow and utterly inadequate as a guide to practical deliberation and choice.[18]

Obviously, a defense of NNLT and its claims about the existence and nature of moral absolutes is beyond the scope of this paper. My goal in this section has only been to show that natural law theory offers a theoretical context that explains the moral relevance of the DDE's distinction between intention and side effects by offering an account of moral absolutes and why moral absolutes apply only to what we intend.

## 2. CLARIFYING THE DDE'S PROPORTIONALITY CONDITION

The DDE has been formulated in a variety of ways.[19] At the core of these various formulations is the claim that certain harms that it is not permissible to intend may be permissible to accept as side effects if the acceptance of these side effects is proportionate. Most of the debates about the DDE have focused either on why intent should make such a difference to the moral permissibility of knowingly bringing about certain effects or on how to distinguish between intention and side effects, particularly in cases where the supposed side effects are causally close to or inseparable from what is intended. In this section, I set these debates aside in order to focus on the proportionality condition, because it has been both misunderstood and neglected. The proportionality condition is often interpreted as requiring that the good effects one intends "outweigh" the unintended bad side effects. It is often interpreted, in other words, as some form of utilitarian calculus.

The hypothetical trolley problems that pervade the double-effect literature give the illusion that such a calculus is possible, because the competing goods at stake are presented in a way that makes them commensurable. If a runaway trolley is careening toward five unknown people, and you could save the five by diverting the trolley onto another track on which there is only one unknown person, it is clear that the choice to save five people is proportionate to the unintended side effect of killing one person. Yet once more details about the people are provided, numbers alone are insufficient to make the proportionality determination. What if the one person who will be killed if you divert the trolley away from the five is your eight-year-old daughter? And what if the five are escaped convicts, perhaps convicted rapists or murderers?

[18] As Finnis notes, "moral norms exclude irrationality over which we have some control; they do not exclude accepting the inevitable limits we face as rational agents" (Finnis, "Intention and Side Effects," 195). See also Boyle, "Who Is Entitled to Double Effect?," 486–87.

[19] For an account of the various traditional formulations and a clarification of their meaning, see Boyle, "Toward Understanding the Principle of Double Effect," 527–38.

From the NNLT perspective, proportionality cannot be determined by the application of a utilitarian calculus, because such a calculus is usually both impossible and incoherent. A utilitarian calculus would only be possible and meaningful if human goods were commensurable in value, as is the case when you are comparing more or less of the same good—for example, "human life" in general—with no knowledge of the particulars that make those goods (e.g., lives) incommensurable.[20] In such cases, such as the standard trolley problem, however, there really is no choice to be made, because in the absence of any specific knowledge about who the people are, by saving five lives, you get all the value of one life, plus more. If one is thinking clearheadedly about the scenario, the intelligible appeal of the alternative (letting the track continue on its course to kill the five) simply drops away. The only intelligible option in such a scenario is to divert the trolley to save the five. Most real-life cases, however, are not like this.

NNLT holds that there are multiple basic human goods, each of which has distinct, intrinsic value and cannot be reduced to any other more basic unit of value. What follows from this has sometimes been called the "incommensurability thesis," according to which "basic values and their particular instantiations as they figure in options for choice cannot be weighed and measured in accordance with an objective standard of comparison."[21] In other words, when there are a variety of basic goods (or a variety of instantiations of the same good) that one might pursue, one is faced with multiple intelligible options, each of which has distinct and incommensurable value, and therefore a genuine choice has to be made. This does not mean, however, that choices among incommensurable options are arbitrary or that one can never have a conclusive reason to choose one option over another. For while basic goods

[20] Finnis lists a number of other cases in which value comparisons are possible:

> States of affairs considered in abstraction from their origins, context and consequences (for example, their relation to a choosing human will) can often be compared in value; a happy village five minutes before and five minutes after a devastating hurricane. And many other comparisons of value are possible. Moral good can be ranked higher than nonmoral, intelligible good higher than merely sensible good, basic good than instrumental, divine and human than animal, heavenly than earthly. More of the same can be compared with less of the same; a genocidal society is worse than a murderous individual; killing is worse than wounding.

Thus, explains Finnis, the "incommensurability which makes proportionalism [or utilitarianism] irrational is incommensurability of goods involved in *options*"—that is, in situations that require a genuine choice, because more than one option has intelligible appeal (Finnis, *Moral Absolutes*, 53).

[21] George, "Does the Incommensurability Thesis Imperil Common Sense Moral Judgments?," 187.

provide first-order reasons for action, these reasons may be defeated (but not destroyed) by moral norms, which are second-order reasons for action. For instance, if I am playing golf and see a child drowning in the water trap, the Golden Rule is a moral norm that provides a conclusive second-order reason to interrupt my golf game in order to save the child. This second-order reason defeats my first-order reason (grounded in the basic value of play) to continue my golf game, but it does not destroy that reason—that is, it would still be intelligible, though not fully reasonable and therefore morally wrong, to continue the game and ignore the child.[22]

Therefore, apart from cases in which there is no real choice to be made, because only one option has intelligible appeal, proportionality cannot be a matter of applying a utilitarian calculus to weigh the premoral goods at stake against each other. Instead, proportionality is determined by considering whether or not there are proportionate *moral* reasons for one's choice despite the foreseen bad side effects. In other words, determining proportionality requires considering whether or not any *other* moral norms (apart from the norm forbidding intentional damage or destruction of basic human goods) are violated by one's choice. For morality, according to NNLT, requires acting in line with the *integral directiveness* of practical reason, with due regard for all of the basic human goods in all persons, which means choosing such that one's will is compatible with a will toward integral human fulfillment. As already explained, intentional damage to or destruction of basic human goods is one obvious way of violating that fundamental moral requirement, but it is certainly not the only way.

Since practical reason grasps human goods as worthy of pursuit not just for me and those I care about but for all who can be fulfilled by them—that is, all human persons—it is unreasonable (and therefore immoral) to arbitrarily prioritize the good of one person or group over another. Thus, there is a requirement to act with fairness, often understood as an application of the Golden Rule. In double-effect cases that involve intended benefits and unintended harms that fall on different people or groups of people, the Golden Rule is usually the appropriate test for determining whether the action meets the proportionality condition. For example, bombing a tiny munitions factory with no particular strategic importance in the course of a just war, even though the factory is

[22] For a more detailed discussion of this case along with a full explanation and defense of the incommensurability thesis, see George, "Does the Incommensurability Thesis Imperil Common Sense Moral Judgments?" For further clarification and defense of the incommensurability thesis, as well as an account of its importance for the possibility of free choice, see Boyle, "Free Choice, Incomparably Valuable Options, and Incommensurable Categories of Good," 123–41.

located in the midst of a densely populated city, would be morally wrong even if the civilian deaths were unintended side effects, because the acceptance of those side effects would be unfair (a violation of the Golden Rule), as it would exhibit a wanton disregard for the value of those civilians' lives.

The reason why I invoke the Golden Rule here, rather than simply talking about fairness, is that it is a helpful heuristic device for determining whether or not we have unfairly disregarded or deprioritized the good of others by acting despite foreseeing that they will be harmed. The Golden Rule does this by asking us to imaginatively place ourselves in the other's shoes and consider whether or not we would think we were being treated unfairly if the tables were turned. In the bombing case described above, for instance, we might ask if we would still be willing to bomb the military target if it were right next to a prisoner-of-war camp with many of our own country's soldiers or the people living in that city were citizens of our own country rather than citizens of the country with which we are at war. If the answer is no, then the action violates the Golden Rule and is therefore unfair.[23]

Application of the Golden Rule or fairness principle can also explain many philosophers' intuitions regarding the "fat man" variation of the trolley case, in which the only way to stop the runaway trolley from killing five people is to push a fat man over the side of a bridge onto the trolley tracks so that his body will obstruct the course of the train and stop it.[24] Some see this variation on the trolley case as involving the intentional killing of the fat man, but for the reasons mentioned earlier when discussing intention, I believe that this interpretation of the case is mistaken.[25] It is not the fat man's death in itself that is the means to your end of saving the five (any more than it is the soldier's death that is the means of saving his comrades from the grenade blast in the grenade case previously discussed). Rather, the means to one's end in the fat man case is stopping the train by obstructing its course with his body. Pushing the fat man onto the tracks is wrong not because it is intentional killing but because it is unfair (a violation of the Golden Rule) to force him to do this against his will. If, on the other hand, you are the fat man and you choose to throw yourself in front of the

[23] Not all prioritization of one person or group over another is arbitrary. Some degree of self-preference, for instance, is reasonable given that we have more direct control over and responsibility for our own good than we do over and for anyone else's good. Similar arguments can be made for prioritization (up to a point) of the good of family members or others to whom we have a morally relevant connection that gives rise to special obligations. For more on the connection between personal relationships and special obligations, see Moschella, *To Whom Do Children Belong?*, 29–34.

[24] See, e.g., Edmonds, *Would You Kill the Fat Man?*, ch. 5.

[25] Edmonds, *Would You Kill the Fat Man?*, ch. 5.

train to stop it from hitting the five, you have not done anything wrong (unless accepting your death in this situation would be unfair to others, such as your dependents)—on the contrary, you have heroically accepted your death as a side effect in order to save the people about to be killed by the runaway trolley.

Another moral norm that is particularly relevant to double-effect cases in which one must choose between competing goods is what, following Christopher Tollefsen, I will call the vocation principle.[26] According to this principle, it is only from the perspective of one's overall vocational commitments and obligations, broadly understood, that one can establish a reasonable order of priorities among competing basic goods. The reason for this principle is that, while all the goods are in themselves equally worthy of pursuit, none can be pursued meaningfully without a considerable dedication of time and effort that takes away from one's ability to pursue other goods (or other instantiations of the same category of good). A commitment to dedicate oneself in a particular way to a certain good, so as to be able to make a deep and meaningful contribution to human flourishing with respect to that good, therefore provides a reason to prioritize pursuit of that good and enables one to do so without discounting the value of other goods. [27] For example, this principle is relevant to determining whether or not my hypothetical choice to forgo having dinner with friends in order to work on my paper meets the DDE's proportionality condition. (Fairness is also at play in this case insofar as the good of my friends is also at stake, but that

[26] Curlin and Tollefsen, *The Way of Medicine*, 45–50. Finnis describes this principle as a requirement to make and follow a rational life plan, but I believe Tollefsen is right that the language of vocation, understood in a broad, not specifically religious sense, better captures our common moral experience in this regard. For many if not most people do see at least some of the overarching commitments in their lives (commitments to a particular profession, to a specific sort of service within the community, to marriage, to the raising of children, to long-term care for sick, disabled, or elderly relatives, etc.) as a response to some sort of calling, rather than purely as a matter of choice or preference.

[27] A full account of the moral relevance of commitments is beyond the scope of this paper. George offers a brief explanation:

> Basic values and their particular instantiations can sometimes be brought into a certain form of rational commensurability with respect to future choices by a choice or commitment (embodied in a choice) which one reasonably makes here and now. In light of a reasonable personal (e.g., vocational, relational, educational) commitment I have made, it may be perfectly reasonable for me to treat, and, indeed, it may be patently unreasonable for me to fail to treat, certain basic values or certain possible instantiations of a single basic value as superior to others in their directive force (for me). Choosing in harmony with one's past reasonable commitments, and, thus, establishing or maintaining one's personal integrity (in the nonmoral as well as the moral sense), constitutes an important moral reason which often guides our choices between rationally grounded options. (George, "Does the Incommensurability Thesis Imperil Common Sense Moral Judgments?," 189)

is not my focus here.) Proportionality in this case would depend on whether my choice reflects reasonable fidelity to my professional commitments or instead flows from a fanatical obsession with professional excellence that reflects a failure to respond adequately to the value of other goods, such as friendship.

In sum, I have argued in this section that the correct way to understand the proportionality condition of the DDE is to think of it not as asking us to determine whether the premoral benefits one is seeking outweigh the foreseen premoral harms one is accepting as a side effect but rather as requiring that there be proportionate *moral* reasons for one's choice, which means determining that accepting the bad side effect(s) does not in itself violate a moral norm. Though I do not pretend to offer an exhaustive list, the moral norms that are most likely to be relevant to this determination are the Golden Rule (or fairness principle) and the vocation principle. Thus, I propose that the DDE be formulated as follows:

> *Doctrine of Double Effect*: Harms that it is never permissible to intend (i.e., harms to basic human goods) may be permissible to accept as side effects if and only if the acceptance of these side effects does not itself violate any moral norm—for example, does not unfairly prioritize one person's or group's good over another's and does not arbitrarily prioritize one good (or instantiation of a good) over another.

While this formulation may seem to lack a genuine proportionality condition, that is not the case, for the second clause is equivalent to: "if and only if there are proportionate moral reasons." I prefer the above formulation, however, because it more clearly indicates what it means for there to be "proportionate moral reasons" to act despite the foreseen harmful side effects.

## 3. RESPONSE TO RECENT CRITICISMS OF THE DDE

Having situated the DDE within its proper theoretical context and reformulated the DDE's proportionality condition in a way that is coherent with that context, I now seek to defend the DDE, thus contextualized and reformulated, from recent criticisms.[28]

### 3.1. *Kershnar and Kelly's Right-Based Critique of the* DDE

Kershnar and Kelly argue that the DDE is incoherent or trivial, because it presumes that people have stringent moral rights, but "if people have stringent

[28] Of course, since space limitations make it impossible to defend the broader natural law theory that makes sense of the DDE, the best I can do here is to offer a dialectical defense conditional on the truth of that broader theory.

moral rights, then the doctrine of double effect is either false or unimportant."[29] The authors consider both strong and weak versions of the DDE and conclude that the strong version is false and the weak version is trivial. I agree that the weak version—"other things being equal, it is deontically worse to intentionally infringe a norm than to foreseeably do so"—is indeed trivial, so I will focus here on their critique of the strong version.

They formulate the strong version as follows:

> An act is morally right if and only if the agent does not intentionally infringe a moral norm and the act brings about a desirable result (perhaps the best state of affairs to the agent or a promotion of the common good).[30]

Both clauses in this formulation are problematic, but here I will focus only on the first clause, around which the bulk of the authors' argument revolves.[31] The DDE is presented as prohibiting intentional infringement of *moral norms* rather than prohibiting intentional *harm*.[32] As explained in section 1, however, the true absolute moral prohibition that is presumed by the DDE is an absolute prohibition on intentional harm to basic human goods.

Kershnar and Kelly offer several reasons for formulating the DDE as they do. They argue, first, that "if the best theory of harm is the counterfactual comparative account, . . . too many things are harmful and would then mistakenly

[29] Kershnar and Kelly, "The Right-Based Criticism of the Doctrine of Double Effect," 215.

[30] Kershnar and Kelly, "The Right-Based Criticism of the Doctrine of Double Effect," 215.

[31] While it is less important to the overall argument, Kershnar and Kelly's formulation of the DDE is also problematic because it lacks a genuine proportionality condition. They recognize that this is a concern, and note that "a desirable result might, and perhaps should, include a proportionality condition." However, they avoid this because they presume that a proportionality condition must be utilitarian in nature. As they state, "the problem is to fill that condition without it becoming an optimality condition" ("The Right-Based Criticism of the Doctrine of Double Effect," 216). As explained above, however, proportionality need not be understood in this way. Kershnar and Kelly do mention the possibility of a "moderate version" of the doctrine of double effect that "allows for the satisfaction of other considerations, such as desert, fairness, or rights, to be necessary for an act to be morally right" (216). This is closer to the interpretation of proportionality I have proposed here.

[32] While this point is not central to my argument, Kershnar and Kelly's framing of the DDE as prohibiting intentional infringement of moral norms seems to be a category mistake, for what we intend, strictly speaking, is what we take to be good, what we take to offer some intelligible benefit, either in itself, or as a means to the benefit we seek. Thus, while people knowingly and intentionally act in ways that violate moral norms, they do not *intend* the violation of moral norms just as such (except perhaps in unrealistic hypothetical cases). Rather they intend to do something (in pursuit of some perceived benefit) which violates a moral norm.

trigger a doctrine-of-double-effect analysis."[33] However, on my account, the only harms that would require a double-effect analysis are harms to basic human goods, since those are the only harms that are (if intentional) absolutely morally prohibited according to NNLT.[34] Second, Kershnar and Kelly do not think intentional harms are wrong unless they "set back someone's legitimate interests," and they offer the example of a woman hitting a robber to defend herself as a case of intentional harm that is morally permissible for this reason.[35] In cases of self-defense, however, the tradition of double-effect reasoning beginning with Aquinas himself has held that harming (or even killing) others in self-defense is justified only if the harm is a side effect of an act that seeks to preserve one's own life (or some other important good) by rendering one's attacker harmless. Harm to the attacker's life or health is not part of the proposal one adopts in such cases and is therefore not intended, for if a blow temporarily stuns the attacker without actually damaging his health, one will still have achieved one's goal.[36] Third, Kershnar and Kelly claim that "intentional harm is permissible if the person who is harmed validly consents to it."[37] Yet this is not true in the natural law tradition, according to which intentional harm to basic human goods is always wrong, regardless of consent.

These problems with Kershnar and Kelly's formulation of the DDE are crucial for understanding why their argument—while perhaps successful as a critique of other versions of the DDE—fails as a critique of the DDE as presented in this paper and as generally understood in the natural law tradition. Kershnar

[33] Kershnar and Kelly, "The Right-Based Criticism of the Doctrine of Double Effect," 217.

[34] Further, given (as explained above) that any choice to pursue one good rather than another involves damage to the good not chosen, I believe that double-effect reasoning is in fact pervasive (though often implicit) in our moral deliberations. The reason why double-effect analysis is implicit in most cases is that the harm to the good not chosen is usually quite obviously a side effect, and thus the moral analysis is really about proportionality, which comes down to the application of the relevant moral norm—e.g., the Golden Rule or the vocation principle, as described above.

[35] Kershnar and Kelly, "The Right-Based Criticism of the Doctrine of Double Effect," 217.

[36] Many of the other cases Kershnar and Kelly discuss in their article also incorrectly interpret the harm as intentional when it would actually be a side effect on the account of intention explained in section 1. For instance, in the "Sophie" case a woman in a Nazi prison asks the guard to kill her instead of her daughter. Kershnar and Kelly believe that Sophie is intentionally trying to kill herself, but this is false ("The Right-Based Criticism of the Doctrine of Double Effect," 223). Sophie is intending to save her daughter by taking her daughter's place. Her death is a side effect. If the guard is moved by her offer and decides not to kill either of them, or if they are unexpectedly rescued before the execution takes place, she will still have achieved her goal, which means that her death was not necessarily part of her proposal for action.

[37] Kershnar and Kelly, "The Right-Based Criticism of the Doctrine of Double Effect," 223.

and Kelly's argument against the strong version of the DDE rests on the claim that the DDE presumes that people have rights and that the relevant norms that the DDE forbids one from intentionally infringing are rights. They go on to conclude that the DDE is incoherent because infringing a right is wrong regardless of whether or not one does so intentionally or as a side effect.

The problem with this argument, however, is that it interprets the DDE in a way that is not coherent with the DDE's theoretical context. In the natural law tradition, absolute rights are the flip side of absolute moral prohibitions, and absolute moral prohibitions (as explained above) are prohibitions on *intentional* harms to basic human goods. One can also speak of rights more generally as articulating the requirements of justice from the perspective of the beneficiary. On this broader account of rights, one can also violate others' rights by, for instance, unfairly accepting harm to their basic goods as a side effect.[38] This, I believe, is the correct explanation for why many—including Kershnar and Kelly—think that pushing the fat man over the bridge in the fat man version of the trolley problem would violate the fat man's rights. Thus, my account preserves common intuitions about the fat man case without adopting Kershnar and Kelly's view of absolute rights as defined independent of the agent's intention.

Kershnar and Kelly do consider the objection that their argument begs the question by failing to recognize that "intentions are at the core of rights."[39] They believe, however, that defining rights violations as intentional infringements of a norm, combined with their formulation of the DDE as forbidding intentional norm-infringement and their definition of norms as rights, leads to an infinite regress. This problem is resolved, however, by correctly formulating the DDE in light of its proper theoretical context within the natural law tradition. If we recognize that the DDE is premised upon the existence of moral absolutes that prohibit certain acts as specified by their intent (i.e., the intent to damage basic human goods), then the idea that intentions are at the core of absolute rights is completely coherent with the DDE and presents no problem of regress.

[38] Property rights (which are not absolute rights according to the natural law tradition) would fall under this category of intention-independent rights. Property rights flow from the Golden Rule, not from the norm prohibiting intentional damage to basic goods, and the justification for private property in the natural law tradition is indirect, based on the instrumental value of private ownership for both the individual and common good. It is also worth noting that the Thomistic natural law tradition does not view one's body as property, for the tradition holds that the body is an intrinsic aspect of the person and that persons cannot be owned. For a natural law account of property rights (and their limits), see Boyle, "Fairness in Holdings."

[39] Kershnar and Kelly, "The Right-Based Criticism of the Doctrine of Double Effect," 226.

Given that Kershnar and Kelly's formulation of the DDE, their concept of rights, and their concept of intention are all at odds with the natural law tradition within which the DDE finds its proper theoretical context, it is no surprise that they find the DDE to be incoherent. If I presented an analysis of Newton's second law of motion but did so by giving an account of force, mass, and acceleration at odds with the basic principles of Newtonian physics, I would likewise be likely to conclude that the second law of motion is incoherent. Kershnar and Kelly's *version* of the DDE is indeed incoherent with *their* broader theoretical presuppositions, but their argument leaves the real DDE—accurately understood within the context of the natural law tradition—completely untouched.

### *3.2. Steinhoff on the Lack of a Rationale for the DDE*

Steinhoff examines a number of purported rationales for the DDE and claims to find none of them persuasive. Here I will only consider Steinhoff's discussion of the rationale that is in line with the account I have provided in section 1, along with his presentation of several cases that might be considered a challenge to that rationale.

Steinhoff first considers Thomas Nagel's argument that the rationale for the DDE is based on the special wrongness of aiming at evil. Steinhoff quotes the core of Nagel's argument: "to aim at evil *even as a means*, is to have one's action *guided* by evil.… But the *essence* of evil is that it should *repel* us.… So, when we aim at evil we are swimming head-on against the normative current."[40] Steinhoff believes that Nagel's analysis of the phenomenological importance of aiming at evil has already been decisively refuted by Bennett and others. Bennett argues that Nagel's argument is persuasive only in cases where the evil is intended *as an end*, not cases (such as the standard terror bomber case) where the evil is sought only as a means to some further good.[41] This argument, however, fails to appreciate that intending as a means is still *intending* and thus still involves setting oneself in the direction of the evil. For ends and means are relative. Every end can also be intended as a means to some further end, and every means is itself a proximate end.

Bennett's failure to grasp this fundamental point can be seen in his argument that the only difference between the tactical bomber (who accepts the civilian deaths as a side effect of a raid on a military target) and the terror bomber (who seeks civilian deaths to terrorize the population and end the war) is simply a difference in "how intensely or thoroughly hostile they are" to the civilians. Each bomber, says Bennett, will "manoeuvre towards" the civilians' death in a

[40] Steinhoff, "Wild Goose Chase," 3; and Nagel, "The Limits of Objectivity," 132.

[41] Steinhoff, "Wild Goose Chase," 3; and Bennett, *The Act Itself*, 224.

variety of ways. The tactical bomber, just like the terror bomber, "relentlessly and ingeniously pursues, *for as long as he has any reason to*, a path that inevitably leads to" the death of the civilians.[42] Yet this account of the intentions of each bomber is imprecise. The terror bomber maneuvers toward the civilians' death, but the tactical bomber maneuvers toward the military target and avoids (insofar as is possible) killing the civilians. For instance, if the bombing can be done at a time when fewer civilians are likely to be around, he will do that. If he has access to a type of bomb that is more accurate and less likely to cause collateral damage, he will use that one. Bennett might respond that here I am bringing in differences in probable outcomes, which he does think can make a difference morally, but that is not my point. Rather, my point is that the tactical bomber is not aiming at the civilian deaths *in any way* or showing any hostility at all toward the civilians. This absence of hostility is shown in his deliberations about the precise way in which he will carry out the raid.[43] By contrast, the terror bomber really is aiming at the deaths (even if only as a means) and will maneuver and deliberate with that end in mind. The end of killing the civilians may be a proximate one, sought merely for its usefulness in achieving a further goal, but it is nonetheless still an end. Thus, Bennett's supposed refutation of Nagel's argument that intentions matter (at least phenomenologically) because to aim at evil is to have one's action be "guided by evil" is not as decisive as Steinhoff takes it to be.

Steinhoff is nonetheless ultimately correct to say that Nagel's view is insufficient to provide a rationale for the DDE, but not because Nagel's phenomenological account of the difference between intending evil and aiming at evil is faulty. Rather, Nagel's argument, while suggestive, is insufficient because it is simply meant to explain why "the fact that you must try to produce evil" is "phenomenologically important."[44] When considering whether or not this fact is "morally important," Nagel states that there is no "decisive answer."[45] The NNLT view is compatible with Nagel's phenomenological analysis but also provides the missing rationale for the moral significance of this analysis. NNLT does this by recognizing, first, that practical reason directs us to preserve and promote the good in all of its fundamental dimensions; second, that to intend something (even as a means) sets one's will (and with it one's whole self) in the direction

[42] Bennett, *The Act Itself*, 223.

[43] Technically, the terror bomber need not be "hostile" to the civilians either, if by "hostile" one means that he seeks their deaths as an end in itself.

[44] Nagel, *The View from Nowhere*, 183. Note that while the passage Steinhoff quotes comes from an earlier essay prior to the book's publication, that passage is identical to the one in the book.

[45] Nagel, *The View from Nowhere*, 183.

of what one intends; and thus, third, that to intend damage to a basic good (even as a means) is always inherently contrary to the integral directiveness of practical reason, while acceptance of such damage as a side effect is not.[46]

Steinhoff also relies on the argument of Dana Kay Nelkin and Samuel C. Rickless, who deny that it is always wrong to intend great harm. They believe, for instance, that it is not wrong to intend harm to an unjust attacker.[47] This claim, however, directly contradicts the natural law account. As already noted, at the very origin of the DDE is a case of killing in self-defense, and the reason why the DDE is invoked to justify such killing is precisely that intentional killing is considered wrong even in response to an unjust attack.[48] Once again, it is not surprising that the DDE would seem incoherent or unnecessary if the core moral premise that gave rise to it—that is, that some intentional harms are never morally permissible, regardless of the circumstances—is simply denied. The rationale for the DDE provided by the NNLT is therefore not refuted by the arguments of Bennett or Nelkin and Rickless on which Steinhoff relies.

### 3.3. *Steinhoff on the Lack of Intuitive Support for the DDE*

The rest of the arguments that Steinhoff criticizes in his article differ so significantly from the natural law account that they are not worth discussing here.[49] In the course of his discussion of one of these arguments, however, Steinhoff considers Alexander Sarch's comparison of two cases of arson—a pair of cases that Steinhoff believes are superior to those usually used to test our intuitions because the arson cases actually keep everything equal except for the difference in intent.[50] Steinhoff believes that such properly constructed examples actually do not elicit the intuitions that would support the DDE. Here are the two arson cases as described by Steinhoff:

[46] For more on the insufficiency of Nagel's rationale for the DDE and of other attempts to provide a rationale for the DDE apart from the natural law tradition, see Boyle, "Who Is Entitled to Double Effect?," 475–94.

[47] Nelkin and Rickless, "So Close , Yet So Far," 403, cited by Steinhoff, "Wild Goose Chase," 4.

[48] When Aquinas discusses self-defense, he does say that public authorities (unlike private citizens) may sometimes intend to kill or maim. Grisez and other NNL theorists, however, believe that Aquinas's view on this matter is inconsistent with his broader moral theory and make the case for an absolute prohibition on intentional killing of human beings, regardless of whether they are guilty or innocent, and regardless of whether the killer is a private citizen or a public official. Grisez, "Toward a Consistent Natural Law Ethics of Killing."

[49] Warren Quinn's proposed rationale for the DDE, which Steinhoff discusses and critiques at length, is also criticized by Boyle from an NNLT perspective in "Who Is Entitled to Double Effect?" (Quinn, "Actions, Intentions, and Consequences").

[50] Sarch, "Double Effect and the Criminal Law," 462.

> In the first case, Alan is paid to burn down a building, and he indeed burns it down to get the money, foreseeing with certainty the death of a victim who happens to be in the house. He regrets the victim's death, but the money is more important to him. In the second case, Bobby is paid to kill the victim (he would not get the money if the victim survived) by burning down the house, and in order to get the money he indeed kills the victim by burning down the house. He regrets the victim's death, but the money is more important to him.[51]

I agree with Steinhoff that (*contra* Sarch's claim) Alan and Bobby are equally culpable, despite the fact that Alan accepted the victim's death as a side effect while Bobby intended the victim's death. Yet the DDE, properly understood, does not presume that intending harm is the only or worst way to act immorally. The DDE as I have explained it, and as the natural law tradition has long understood it, only says that some harms that it is never permissible to intend *may* be permissible to accept as a side effect if and only if the acceptance of the side effect is proportionate (i.e., does not violate any other moral norm). Callous disregard for the harms that will result as a side effect of one's actions may be just as bad as or worse than the intentional infliction of harm. The arson cases therefore do not show that the DDE, properly understood, yields counterintuitive results.

What about other cases—carefully formulated to keep everything equal except for the case of intent—that Steinhoff presents in a different article and uses to argue that our intuitions about them do not support the DDE? I do not believe that any of the cases presented, if analyzed in light of the account of the DDE I have offered above, are any more problematic than the arson cases. Here I look at only one of Steinhoff's pair of cases by way of example:

> *Pedestrian I*: A private person (pursuing a just cause) throws his hand grenade on a mark *X* in front of a pedestrian, intending that the pedestrian will die.
>
> *Pedestrian II*: A private person (pursuing a just cause) throws his hand grenade on a mark *X* in front of a pedestrian, foreseeing that the pedestrian will die.[52]

Steinhoff says that he discerns no difference in the justifiability of the two cases. I disagree. But for those who share Steinhoff's intuition, I would argue that the reason why it is hard to see the difference in the cases is that we are not told

[51] Steinhoff, "Wild Goose Case," 13–14.

[52] Steinhoff, "The Secret to the Success of the Doctrine of Double Effect (and Related Principles)," 247.

what the "just cause" is or why throwing the grenade onto the *X* is necessary for the cause. One might thus be inclined to think that throwing the grenade is wrong in both cases, because we have not been presented with proportionate moral reasons for throwing the grenade despite the foreseen bad side effect. And I believe that we also find it hard to imagine why one could not warn the pedestrian and wait for her to move before throwing the grenade. Thus, the case is a very bad test of intuitions.[53]

Allow me to propose a better set of cases—a revised version of the original trolley cases—that meet Steinhoff's desiderata of keeping everything equal except the intent:

> *Trolley and Politician I*: A runaway trolley is careening toward five people who are stuck on the track. The people will be killed if the trolley hits them. You could divert the trolley onto a side track by flipping the switch. There is one person stuck on the sidetrack who will be killed if you do this. That person happens to be a local politician whose policies you believe are harmful to the community. You choose to divert the trolley onto the sidetrack intending to save the five and foreseeing (but not intending) the death of the public official.
>
> *Trolley and Politician II*: A runaway trolley is careening toward five people who are stuck on the track. The people will be killed if the trolley hits them. You could divert the trolley onto a side track by flipping the switch. There is one person stuck on the sidetrack who will be killed if you do this. That person happens to be a local politician whose policies you believe are harmful to the community. You choose to divert the trolley onto the sidetrack intending to save the five and also intending the death of the public official (as a means to ending the harmful community policies).

I think that the moral difference in these cases is clear and that while the action in Trolley and Politician II is morally wrong, the action in Trolley and Politician I is morally right. If your intuitions do not support this conclusion, consider the difference it would make to the deceased politician's spouse if you gave

[53] While there is a significant body of empirical literature regarding the DDE, I do not engage with this literature because my critique of Steinhoff is not empirical. Rather, my appeal here (like Steinhoff's appeal in presenting the Pedestrian cases) is an appeal to the judgment *of the reader*, not a prediction of how most people would respond to the case. Further, most of the empirical literature on this issue is based on assumptions that are incompatible with natural law theory—e.g., the assumption that our intuitive judgments about morality are primarily the result of psychological causes or "structural feature(s) of the mind," rather than, for instance, the result of education and habituation through prior choices, as discussed in note 54 below. Cushman, "The Psychological Origins of the Doctrine of Double Effect," 765.

him an honest account of your actions in each case. In Trolley and Politician I, you could honestly say: "I am very sorry about your loss. I know that I have publicly criticized your wife's policies, but I swear to you that the only reason I diverted the trolley was to save the five. My reasons for acting truly had nothing to do with my political disagreements with your wife." In Trolley and Politician II, however, you could not honestly say this. While you could honestly say that you acted to save the five, you could not honestly say that saving the five was your *only* reason or deny that your political disagreements (and accompanying desire to remove his wife from the scene by killing her) were part of the reason for your choice. To the mourning widower, I believe that this difference would indeed be morally significant, and considering this perspective may help sensitize us to the objective moral difference between the two cases.

Steinhoff has thus not provided any clear evidence that the DDE runs contrary to common intuitions. At any rate, my own defense of the DDE is not intuition-based but rather based on the moral theory outlined above, which I believe to be true for independent reasons.[54]

### 3.4. *Scanlon on Critical and Deliberative Uses of Principles*

Like Steinhoff, Scanlon argues that the DDE's intuitive appeal is illusory. Since many of Scanlon's critiques of the DDE overlap with those I have already addressed in response to Steinhoff, I focus here on an aspect of Scanlon's critique that has not already been implicitly addressed in the preceding sections. Scanlon believes that the appeal of the DDE in many cases is due to a failure to distinguish between what he calls the "critical" and "deliberative" uses of moral principles. In their use as a guide to deliberation, moral principles identify "the considerations that make it permissible or impermissible to do *X* under the circumstances in question."[55] Critical use of moral principles, however, goes

[54] Indeed, I am generally skeptical of attempts to base moral claims on intuitions, for I believe that morality is a matter of practical reasonableness, while intuitions (at least as the word is commonly used) are based on feelings, and feelings are shaped by beliefs (which may or may not be true), as well as by bias and numerous other subrational factors. Nonetheless, those whose feelings have been shaped by true moral beliefs and who habitually seek to avoid the distortions of bias or other subrational factors in their judgments will tend to have moral intuitions that correspond to moral truth. Further, natural law theory holds that practical reason's first principles (i.e., the identification of basic human goods) are self-evident to those who understand their terms and that the most basic moral requirements of natural law are easily accessible to reason. Thus, we would expect that those basic requirements—such as the Golden Rule—would be widely agreed upon at least implicitly by most people who seek to lead morally upright lives. For these reasons, showing that the implications of one's moral theory accord with at least some people's moral intuitions—or at least do not seriously contradict them—does have some value.

[55] Scanlon, *Moral Dimensions*, 22.

beyond identifying "which considerations are relevant to the permissibility of such an action and how they should be taken into account," because it requires also "asking whether the agent in question in fact took those considerations into account in the proper way."[56] Scanlon illustrates this distinction with the following example: "I have promised to sell you my house, and … under the circumstances this counts as a decisive reason for doing so. In particular, the fact that I could get more money by breaking my promise and selling the house to someone else is not a sufficient reason to do that. But suppose I do break the promise in order to get this benefit."[57] Scanlon says we might be tempted to say that his action was wrong because he "acted for a bad (selfish) reason," but in fact the reason why the action was wrong is that he "had promised to sell you the house."[58] The former, argues Scanlon, "is a criticism of the way [he] went about deciding, not an explanation of why [his] action would be wrong."[59] He believes, therefore, that we are mistaken when we make judgments of permissibility by using moral principles critically rather than deliberatively.

This contention is relevant to Scanlon's critique of the DDE because Scanlon believes that appealing to intent to explain why, for instance, terror bombing is impermissible and tactical bombing is permissible is to make the mistake of determining permissibility through the critical rather than deliberative use of moral principles. Scanlon believes that the difference between the two cases flows not from differences in intent but from general moral principles about conduct in war. In particular, "the principle relevant to these cases states a class of exceptions to the general prohibition against the use of deadly force, and specifies the limits to those exceptions."[60] He spells out this principle more concretely as follows:

> In war, one is sometimes permitted to use destructive and potentially deadly force of a kind that would normally be prohibited. But such force is permitted only when its use can be expected to bring some military advantage, such as destroying enemy combatants or war-making materials, and it is permitted only if expected harm to noncombatants is as small as possible, compatible with gaining the relevant military advantage, and only if this harm is "proportional" to the importance of this advantage.[61]

[56] Scanlon, *Moral Dimensions*, 22–23.

[57] Scanlon, *Moral Dimensions*, 23.

[58] Scanlon, *Moral Dimensions*, 24.

[59] Scanlon, *Moral Dimensions*, 24.

[60] Scanlon, *Moral Dimensions*, 28.

[61] Scanlon, *Moral Dimensions*, 28.

He believes that it is this principle—rather than differences in the agents' intent—that accounts for our judgment that tactical bombing, but not terror bombing, is permissible.

Yet Scanlon's argument here simply begs the question of why the use of force in war should only be allowed for the destruction of military targets rather than for the purposes of killing civilians to demoralize the population and bring the war to a swift conclusion. What, in other words, is the moral justification for this principle limiting the exceptions to the general prohibition on the use of deadly force? Indeed, both the exceptions and the limits of those exceptions sound suspiciously *ad hoc,* tailored to explain the difference between tactical and terror bombing. Scanlon does recognize that his discussion simply presupposes the existence of the moral principles upon which he relies (principles that have roots in the natural law tradition) without any attempt to defend them or to justify "the importance they attach to the distinction between combatants and noncombatants." He believes, however, that he does not need to do so, because his point was merely to show that "these principles need not be understood as making permissibility dependent on intent."[62] Unless, however, he can provide some plausible explanation for these principles that is completely independent of intent, he has not actually proven his point.

There is nothing *ad hoc*, by contrast, about the NNL account of the distinction between tactical and terror bombing. For NNLT holds that intentional damage or destruction of basic human goods (such as life) is always wrong, because it is always incompatible with a will toward integral human fulfillment, while accepting death as a side effect is not *necessarily* wrong, because it is not *necessarily* incompatible with a will toward integral human fulfillment. This account also provides an explanation of what Scanlon takes to be the relevant moral principle in this case by spelling out the underlying moral justification of that principle. In the absence of any better account of the moral difference between tactical and terror bombing, therefore, the NNL account (including the DDE) remains the best explanation.

Further, the NNL account outlined in section 1 shows that the distinction between the "critical" and "deliberative" uses of moral principles does not work and begs the question by presuming that agent intent is morally irrelevant (which is precisely what the distinction is trying to prove). In the NNL view, actions can only be identified (in the morally relevant sense) and judged from the first-person perspective. Determining, for example, whether lethally shooting an assailant is an act of homicide or self-defense depends on the proposal that I adopted in choosing to shoot. Was my choice to shoot an adoption of

[62] Scanlon, *Moral Dimensions*, 32.

the proposal "I am going to save my life and rid the community of this dangerous criminal by killing him" or the proposal "I am going to incapacitate this attacker so that I can get away and save my life"? If I adopted the first proposal, my act would be homicide (though the circumstances would likely diminish my culpability); if I adopted the second, it would be self-defense. I have already explained why, from the NNL perspective, intent has this moral significance. Further defense of the view is beyond the scope of this paper and has been provided elsewhere.[63] My point here is simply to note that Scanlon's use of the critical/deliberative distinction is incompatible with the broader moral theory that provides the proper theoretical context for the DDE. Once again, therefore, it is unsurprising that one would find it difficult to defend the DDE if one jettisons the core assumptions of the tradition of moral thought within which the DDE was developed.

## 4. Conclusion

My goal in this paper has been threefold: first, to explain the rationale for the DDE within the broader theoretical context of natural law theory; second, to provide an account of the proportionality criterion that is in line with that theoretical context; and third, to (conditionally) defend the DDE, properly understood, from recent attacks. I have argued that while recent critiques of the DDE may be successful in attacking their own version or interpretation of the DDE, they are largely attacking a straw man. The "real" DDE remains untouched by their critiques. This argument does not necessarily imply that the DDE is justified—only that it stands or falls with the broader natural law theory of which it forms a part and thus cannot be successfully defended or critiqued in isolation from that broader theory. Yet this is precisely what many critics of the DDE (such as those discussed above) attempt to do.[64] I should clarify that my argument is not meant to show that the DDE stands or falls specifically with the "new" natural law approach—though I believe that this approach offers the most consistent and rigorous version of natural law ethics—but rather that it stands or falls with natural law theory broadly construed to include any theory that has an account of absolute moral prohibitions in which the prohibited

[63] See, e.g., Tollefsen, "A First Person Account of Human Action."

[64] As one reviewer has pointed out, these critics would likely also disagree with the broader moral and metaphysical framework within which the DDE is embedded, but their critiques do not directly engage with this framework and often do not even acknowledge its existence. Instead, they treat the DDE as an isolated principle rather than part of a broader natural law theory.

act-types are identified at least in part by agent intent.[65] For those already inclined to reject both the DDE and natural law theory more generally, this paper may not present much of a challenge, though it may nonetheless offer a helpful clarification of what is actually at stake in the debate. On the other hand, for those who believe that the DDE does capture an important moral truth but are skeptical of natural law theory, this paper suggests that coherence requires them to either abandon the DDE or reevaluate their skepticism of natural law.[66]

*Catholic University of America*
*moschella@cua.edu*

## REFERENCES

Aquinas, Thomas. *Summa Theologiae*. Translated by the Fathers of the English Dominican Province. New York: Benziger Brothers, 1948.

Bennett, Jonathan. *The Act Itself*. New York: Oxford University Press, 1998.

Boyle, Joseph. "Fairness in Holdings: A Natural Law Account of Property and Welfare Rights." *Social Philosophy and Policy* 18, no. 1 (Winter 2001): 206–26.

———. "Free Choice, Incomparably Valuable Options, and Incommensurable Categories of Good." *American Journal of Jurisprudence* 47, no. 1 (2002): 123–41.

———. "Toward Understanding the Principle of Double Effect." *Ethics* 90, no. 4 (July 1980): 527–38.

———. "Who Is Entitled to Double Effect?" *The Journal of Medicine and Philosophy* 16, no. 5 (October 1991): 475–94.

Curlin, Farr, and Christopher Tollefsen. *The Way of Medicine*. Notre Dame: University of Notre Dame Press, 2021.

Cushman, Fiery. "The Psychological Origins of the Doctrine of Double Effect." *Criminal Law and Philosophy* 10 (December 2016): 763–76.

Edmonds, David. *Would You Kill the Fat Man?* Princeton: Princeton University Press, 2014.

[65] I believe that to justify such an account of absolute moral prohibitions one would also need (1) an objective account of human well-being (i.e., human goods) knowable through reason and (2) an account of moral norms in which acting morally is fundamentally about having a will or character that is correctly disposed toward human well-being. However, defending this claim is beyond the scope of this paper.

[66] I would like to thank David Hershenov for inviting me to present a draft of this paper at the Romanell Center Workshop in April 2022, and also to thank all of the workshop participants for their helpful comments and criticisms.

Finnis, John. "Human Acts." In *Intention and Identity*, 133–51. Oxford: Oxford University Press, 2011.

———. "Intention and Side Effects." In *Intention and Identity*, 173–97. Oxford: Oxford University Press, 2011.

———. *Moral Absolutes*. Washington, DC: The Catholic University of America Press, 1991.

———. *Natural Law and Natural Rights*. 2nd ed. Oxford: Oxford University Press, 2011.

Finnis, John, Germain Grisez, and Joseph Boyle. "Direct and Indirect: A Reply to Critics of Our Action Theory." *The Thomist* 65, no. 1 (January 2001): 1–44.

———. *Nuclear Deterrence, Morality and Realism*. New York: Oxford University Press, 1987.

George, Robert. "Does the Incommensurability Thesis Imperil Common Sense Moral Judgments?" *American Journal of Jurisprudence* 37, no. 1 (1992): 185–95.

Grisez, Germain. "Toward a Consistent Natural-Law Ethics of Killing." *American Journal of Jurisprudence*, 15, no. 1 (1970): 64–96.

Kershnar, Stephen, and Robert Kelly. "The Right-Based Criticism of the Doctrine of Double Effect." *International Journal of Applied Philosophy* 34, no. 2 (Fall 2020): 215–33.

Lee, Patrick. "Distinguishing between What Is Intended and Foreseen Side Effects." *The American Journal of Jurisprudence* 62, no. 2 (December 2017): 1–21.

Lee, Patrick, and Robert George. *Body-Self Dualism in Contemporary Ethics and Politics*. New York: Cambridge University Press, 2008.

MacIntyre, Alasdair. *After Virtue.* Notre Dame: University of Notre Dame Press, 1981.

Moschella, Melissa. *To Whom Do Children Belong? Parental Rights, Civic Education and Children's Autonomy.* New York: Cambridge University Press, 2016.

Nagel, Thomas. "The Limits of Objectivity." In *The Tanner Lectures on Human Values 1*, edited by Sterling McMurrin, 77–139. Salt Lake City: University of Utah Press, 1980.

———. *The View from Nowhere*. New York: Oxford University Press, 1986.

Nelkin, Dana Kay, and Samuel C. Rickless. "So Close, Yet So Far: Why Solutions to the Closeness Problem for the Doctrine of Double Effect Fall Short." *Noûs* 49, no. 2 (June 2015): 376–409.

Quinn, Warren, "Actions, Intentions, and Consequences: The Doctrine of Double Effect." *Philosophy and Public Affairs* 18, no. 4 (Autumn 1989): 334–51.

Sarch, Alexander. "Double Effect and the Criminal Law." *Criminal Law and Philosophy* 11, no. 3 (September 2017): 453–79.

Scanlon, T. M. *Moral Dimensions*. Cambridge, MA: Harvard University Press, 2008.

Steinhoff, Uwe. "The Secret to the Success of the Doctrine of Double Effect (and Related Principles): Biased Framing, Inadequate Methodology, and Clever Distractions." *Journal of Ethics* 22, nos. 3–4 (December 2018): 235–63.

———. "Wild Goose Chase: Still No Rationales for the Doctrine of Double Effect and Related Principles." *Criminal Law and Philosophy* 13, no. 1 (March 2019): 1–25.

Tollefsen, Christopher. "A First Person Account of Human Action." *Ethical Theory and Moral Practice* 9 (2006): 441–60.

*Journal of Ethics and Social Philosophy*
VOL. 26, NO. 2 · DECEMBER 2023

*https://doi.org/10.26556/jesp.v26i2.2777*

# REPUBLICAN FREEDOM AND LIBERAL NEUTRALITY

*Lars J. K. Moen*

Philip Pettit has in recent decades been the foremost defender of a republican way of conceptualizing political freedom.[1] On his account, you are free insofar as you are not subjected to another agent's power of uncontrolled interference—that is, others cannot interfere with you simply as it pleases them.[2] Pettit uses this conception of freedom as nondomination as the basis for his support for a state that is neutral between the different conceptions of the good that exist in a modern, pluralistic society.[3] The ideal of republican freedom, Pettit says, is "compatible with modern pluralistic forms of society."[4] John Rawls gives his support to this view by finding "no fundamental opposition" between republicanism and his own political liberalism.[5]

Pettit also thinks promoting republican freedom conflicts in important ways with the promotion of freedom as noninterference.[6] In assessing this claim, I shall apply the pure negative conception of freedom as noninterference

[1] Pettit, *Just Freedom*, *On the People's Terms*, and *Republicanism*.

[2] Pettit now prefers the terminology of "controlled" and "uncontrolled" interference to his earlier "nonarbitrary" and "arbitrary" interference, respectively. He gives two reasons for this change. First, he wants to avoid any association with "arbitrary" as it is often used to describe actions not conforming to established rules. Such rules may, after all, conflict with the interests of those subject to them. Second, he wants to avoid the connotation of arbitrary with morally unacceptable and nonarbitrary with morally acceptable (*On the People's Terms*, 58). Against Pettit, I show how his distinction between arbitrary and nonarbitrary, or controlled and uncontrolled, is moralized (Moen, "Republicanism and Moralised Freedom").

[3] A person's "conception of the good," John Rawls explains, consists of the ends and purposes the person considers worthy of her or his pursuit over a complete life (*Political Liberalism*, 104).

[4] Pettit, *Republicanism*, 8.

[5] Rawls, *Justice as Fairness*, 144, and *Political Liberalism*, 205. Rawls associates republicanism with Quentin Skinner and Niccolò Machiavelli; he does not mention Pettit. But as both Pettit and Skinner have made clear, they subscribe to the same republican tradition.

[6] Pettit, "Freedom and Probability," *On the People's Terms*, ch. 1, *Republicanism*, ch. 2, "Republican Freedom," and "The Instability of Freedom as Noninterference." Several other republicans also take this view. See, for example, Ingham and Lovett, "Republican Freedom,

associated with contemporary theorists like Hillel Steiner, Ian Carter, and Matthew Kramer.[7] On this view, you are free to perform an action, *x*, as long as no one prevents you from doing *x*. You are therefore not made unfree merely by someone possessing the power to interfere with you in a way you do not control, as on the republican account. And on this pure negative view, any act of prevention makes you unfree, not just uncontrolled interference. Pettit has criticized pure negative freedom in particular.[8] In this paper, however, I show why, despite these differences, Pettit cannot defend liberal neutrality between conceptions of the good without also promoting pure negative freedom.[9] Maintaining his view that promoting republican freedom conflicts with promoting pure negative freedom requires a conception of republican freedom that conflicts with neutrality.

I develop this argument by first, in section 2, explaining the two freedom concepts and showing how they differ. While the two are no doubt different, it is far less clear that their differences matter when we think about promoting freedom in our society. Republican freedom requires that certain institutions be in place to protect citizens against uncontrolled interference. Pure negative freedom, on the other hand, requires only that individuals not restrict each other's ability to act, and it is therefore achievable in the absence of particular institutions. However, as I show in section 3, the institutions required for republican freedom might still be essential for promoting pure negative freedom. While institutions themselves make individuals unfree, in a purely negative sense, to perform certain actions, they can nonetheless enhance citizens' overall pure negative freedom by enabling them to pursue courses of action that would otherwise be unavailable to them. If this is the purpose of the controlled interference compatible with republican freedom, such interference also promotes pure negative freedom.

We shall see in section 4 that making the institutional requirements of republican freedom differ from ones promoting pure negative freedom depends on defining republican freedom so that it demands a more robust protection against uncontrolled interference than is compatible with promoting pure negative freedom. More robust here means the effectiveness of the protection being less dependent on whether agents prefer to interfere with one another or not. I show how enhancing robustness means introducing institutional

Popular Control, and Collective Action"; Skinner, *Liberty before Liberalism*, 77–99, and "Freedom as the Absence of Arbitrary Power"; Viroli, *Republicanism*, 8–12, 40–41.

[7] Carter, *A Measure of Freedom*; Kramer, *The Quality of Freedom*; Steiner, *An Essay on Rights*.

[8] Pettit, "Republican Freedom." See also Skinner, "Freedom as the Absence of Arbitrary Power."

[9] I say *liberal* neutrality, as this neutrality between conceptions of the good is commonly associated with political liberalism. Pettit also notes that "republicanism joins with liberalism" on the issue of neutrality (*Republicanism*, 120).

constraints that offer citizens greater protection at the expense of reducing their range of available courses of action. More robust protection therefore implies more interference, but republicans can justify such interference as under popular control. Their freedom ideal will conflict with the promotion of pure negative freedom insofar as it justifies more interference than is compatible with maximizing the number of courses of action citizens can pursue.

Achieving this higher level of protection against uncontrolled interference depends on citizens committing to a higher level of political engagement to keep powerholders virtuous. In section 5, I show that such protection requires citizens to hold distinctly republican preferences. This requirement makes republicanism incompatible with certain comprehensive doctrines and therefore incompatible with the neutrality Pettit takes his republican theory to be compatible with.[10] I therefore, in section 6, reach the conclusion that a state promoting a conception of republican freedom that conflicts with the promotion of pure negative freedom cannot be neutral. It must introduce measures to make citizens endorse a comprehensive doctrine compatible with the level of political engagement republican freedom requires. To the extent that it requires such measures, republican freedom conflicts with pure negative freedom but also with liberal neutrality. Conversely, to the extent that republican freedom is understood to conflict with such measures, republicans can maintain a commitment to neutrality but not without promoting pure negative freedom.

## 1. Two Concepts of Freedom

To provide a clear view of how republican and pure negative freedom differ, I begin by defining and comparing the two concepts. On the pure negative view, first, an agent, *A*, makes another agent, *B*, unfree to perform an action, *x*, by preventing *B* from doing *x*—that is, by making it physically impossible for *B* to do *x*. It is important to note here that whether *A*'s action toward *B* counts as a prevention does not depend on *B*'s preferences. *A* makes *B* unfree to do *x* regardless of whether *B* would have preferred to do *x*. Pure negative freedom, therefore, differs significantly from Thomas Hobbes's view that *A* can only make *B* unfree to do *x* if *B* has a will to do *x*.[11] If *A* locks the door to the tennis court, Hobbes says, *A* makes *B* unfree to play tennis only if *B* has formed the will to play tennis. On this definition of freedom, *B* can make himself free by

[10] A comprehensive doctrine, Rawls explains, is a set of convictions about how to live, which includes conceptions of the good, how we ought to treat others, and "much else that is to inform our conduct, and in the limit to our life as a whole" (*Political Liberalism*, 13).

[11] Hobbes, "Selections from *The Question Concerning Liberty, Necessity, and Chance*," 81.

adapting his preferences to the constraints *A* has imposed on him.[12] To avoid this counterintuitive implication, later accounts of freedom from interference, or prevention, take it that *A*'s constraint makes *B* unfree regardless of whether *B* wants, or has formed a will, to perform the action *A* makes unavailable to *B*.

On the republican account of freedom, on the other hand, not all kinds of interference are sources of unfreedom. *A*'s interference does not make *B* unfree as long as *B* has instructed the interference and possesses the control to make sure *A* cannot interfere with him in any other way.[13] In Pettit's example, *B* gives *A* the key to *B*'s alcohol cupboard with the instruction of giving it back only on twenty-four hours' notice when *B* later asks for it.[14] When *B* then asks for the key and *A* refuses to give it back, *A* interferes with *B*, but not in a way that makes *B* unfree since *A* acts as *B* instructed. *A* therefore makes *B* unfree in the pure negative sense but not in the republican sense. Analogously, Pettit argues, the government does not interfere with the citizens in an uncontrolled manner when it acts to promote their common interests. These are interests citizens are ready to avow in public without embarrassment because they are compatible with treating others as free and equal members of the society.[15] More specifically, Pettit argues, these are interests in the government providing the resources and protection each citizen needs to effectively exercise the basic liberties.[16]

But for *B* in the alcohol cupboard example to possess the control republican freedom requires, *A* must interfere because *B* instructed it. It cannot merely be a happy coincidence that *A* happens to want to interfere in accordance with *B*'s instruction. In Pettit's terms, "an act of interference will be nonarbitrary to the extent that it is forced to track the interests and ideas of the person suffering the interference."[17] *A*'s interference is then robustly in accordance with *B*'s interests.

[12] Berlin, *Liberty*, 32.

[13] In one way, of course, *A*'s interference with *B* takes away *B*'s freedom to perform the action he otherwise could have performed. Pettit prefers to think of cases of uncontrolled interference as making *B* "non-free but not unfree" (*Republicanism*, 26n1).

[14] Pettit, *On the People's Terms*, 57.

[15] Pettit, *A Theory of Freedom*, 156–60.

[16] Pettit, *On the People's Terms*. For Pettit, basic liberties are the liberties that meet three conditions. First, they can be exercised without thereby preventing any others from exercising them. Second, they are widely considered within a society to have an important role in the lives of normal people. And third, the set of basic liberties are limited only by these first two conditions. This definition leads Pettit to a list of basic liberties, which includes at least freedoms of thought, expression, religious practice, association, assembly, personal property, employment, movement, and to take part in public life as a voter, candidate, or critic. See Pettit, *On the People's Terms*, 103, and "The Basic Liberties," 220.

[17] Pettit, *Republicanism*, 55.

Analogously, free citizens, in the republican sense, have the power to make sure their government robustly promotes their common interests.[18]

This robustness comes about by virtue of systematic protection making uncontrolled interference "inaccessible."[19] "Inaccessible," for Pettit, does not mean "sufficiently improbable" but rather "not within the agent's power."[20] Inaccessibility requires the presence of institutions protecting *B* against *A*'s uncontrolled interference "in a range of possible worlds associated with the available options, however unlikely some of those worlds may be."[21] Institutions set up to prevent uncontrolled interference therefore do not *cause* nondomination, as that would suggest reducing the probability. Instead, they *constitute* nondomination; nondomination "comes into existence simultaneously with the appearance of the appropriate institutions."[22]

The meaning of "inaccessible" interference is a controversial issue in the literature on republican freedom. Several critics of republican freedom have taken it to mean that institutions must make it impossible for *A* to interfere with *B* in a way *B* does not control, which would be unrealistically demanding.[23] If *B*'s freedom to do *x* depends on everyone being made unable to prevent him from doing *x*, then *B* can never be free to do *x* since it will always be possible for someone to prevent him from doing *x*. On this interpretation, therefore, *B* can never be free to do anything whatsoever since there is always a possibility that someone will interfere with him.

However, Pettit and Quentin Skinner stress that the republican objective is not to make interference impossible but instead to establish institutions that make no one capable of interfering with others with impunity.[24] So, uncontrolled interference being "inaccessible," as Pettit says, means it is impossible without having to pay a high price for it. But as Keith Dowding notes, people often get away with breaking the law.[25] If perfect law enforcement is what republican freedom requires, then it remains impossibly demanding. Sean

[18] Pettit, *On the People's Terms*, ch. 3.

[19] Pettit, "Republican Freedom," 104.

[20] Pettit, *On the People's Terms*, 60, and *Republicanism*, 88.

[21] Pettit, *On the People's Terms*, 67.

[22] Pettit, *Republicanism*, 107.

[23] Carter and Shnayderman, "The Impossibility of 'Freedom as Independence'," 139–40; Dowding, "Republican Freedom, Rights, and the Coalition Problem"; Gaus, "Backwards into the Future"; Goodin and Jackson, "Freedom from Fear"; Kramer, "Freedom and the Rule of Law," 843–44, and "Liberty and Domination," 45–46; Simpson, "The Impossibility of Republican Freedom."

[24] Pettit, *Republicanism*, 22; Skinner, *Liberty before Liberalism*, 72.

[25] Dowding, "Republican Freedom, Rights, and the Coalition Problem," 311.

Ingham and Frank Lovett label this view "strong republicanism," which they understand to be "generally impossible."[26]

Pettit, however, says freedom requires only that citizens be sufficiently protected to pass "the eyeball test."[27] To pass this test, Pettit explains, "the safeguards should enable people, by local standards, to look one another in the eye without reason for fear or deference."[28] People must "have this capacity in the absence of what would count, even by the most demanding standards of their society, as mere timidity or cowardice."[29] Pettit is aware of his own vagueness on this matter and sees the required level of law enforcement as akin to what we find in epistemology when we determine how many possible worlds in which a true belief must be present for something to be called knowledge.[30] Ingham and Lovett understand this account to fit into the category of "moderate republicanism," which requires that it be common knowledge that people are generally punished for uncontrolled interference and that people can therefore relate to one another "as if" they know that uncontrolled interference comes at a high cost no one is prepared to take.[31] Institutional protection will have the required deterrent effect.

But by weakening this law enforcement requirement, republicans still encounter the problem that legal systems protecting citizens against uncontrolled interference inevitably rely on the wills of government officials.[32] This protection is therefore not particularly robust. Pettit says the citizens must make sure these officials behave as they are supposed to, and their freedom therefore depends on norms that induce citizens to remain vigilant and resist any decision not tracking their common interests.[33] Under such conditions, Lovett and Pettit argue, no collection of individuals can coordinate so as to take control of legal institutions and gain the power of uncontrolled interference.[34] But it remains the case that the required social norms underpinning such effective institutions can change, as norms often do, and people therefore remain dependent on each other's wills.[35]

[26] Ingham and Lovett, "Republican Freedom, Popular Control, and Collective Action," 778.

[27] Pettit, *Just Freedom* and *On the People's Terms*.

[28] Pettit, *On the People's Terms*, 47.

[29] Pettit, *On the People's Terms*, 84.

[30] Pettit, *On the People's Terms*, 68n38.

[31] Ingham and Lovett, "Republican Freedom, Popular Control, and Collective Action," 779.

[32] Kramer, "Freedom and the Rule of Law," 842–44.

[33] Pettit, *On the People's Terms*, 173.

[34] Lovett and Pettit, "Preserving Republican Freedom."

[35] Simpson, "Freedom and Trust."

The republican robustness requirement remains unclear, but I shall proceed on the understanding that freedom as nondomination at least demands some institutional protection vaguely specified by the eyeball test.

## 2. Promoting Freedom

To Pettit, no account of freedom as noninterference requires that the eyeball test be met since it focuses on the probability, rather than the accessibility, of constraint.[36] Pure negative freedom, in Pettit's view, is therefore compatible with the implausible strategy of making yourself free by currying favor with an agent possessing the power to interfere with you in a way you do not control. To the extent that such slavish behavior reduces the probability of interference and therefore enables you to perform an action, it makes you free in the pure negative sense. Such "liberation by ingratiation," Pettit says, does not work on a republican understanding of freedom since it does not enable you to look the powerful agent in the eye without fear or deference.

If Pettit's understanding of freedom as noninterference is correct, then promoting pure negative freedom will differ in important ways from promoting republican freedom, as the former will require less institutional protection. Institutions will not even be necessary insofar as ingratiation can enhance your pure negative freedom. To test Pettit's claim, we need to understand the role of probability in the measurement of pure negative freedom. Note first that on the pure negative account, you have a *specific* freedom to perform any particular action that no one prevents you from performing.[37] Your specific freedoms do not exist by degrees but are either possessed or not possessed, depending on whether another agent prevents you from exercising them or not.[38] Probability comes in when we consider the probability of possessing a specific freedom—that is, the probability of another agent preventing you from performing a particular action.

This probability measure is relevant for the measurement of *overall* freedom, which does come in degrees. A person's overall freedom, Kramer explains, "is largely determined by the range of combinations-of-conjunctively-exercisable-liberties available to her."[39] If *A* is likely to prevent *B* from doing *y* if *B*

[36] Pettit, "The Instability of Freedom as Noninterference," 704–11.

[37] Here I follow Carter and Steiner's bivalence view of freedom. On Kramer's trivalence account, on the other hand, *A*'s freedom to do *x* does not just depend on *A* not being prevented from doing *x* but also on *A* being able to do *x*.

[38] Carter, *A Measure of Freedom*, 228, 233; Kramer, "Why Freedoms Do Not Exist by Degrees"; Steiner, "How Free," 78.

[39] Kramer, *The Quality of Freedom*, 137.

does *x*, as in a case where *A* has credibly threatened *B*, then *A* interferes with *B*'s conjunctively exercisable actions.[40] We measure overall freedom in terms of the probability of being able to choose any option in one choice situation *and* any option in a subsequent choice situation. The higher the probability of such interference, the greater is the loss of overall freedom.[41]

Let us now return to Pettit's critique of probability in a measurement of freedom. When *B* ingratiates himself with *A* so as to be able to do *x*, he might reduce the probability of *A*'s interference. But note that he does not gain a specific freedom to do *x*. If the ingratiation makes *B* able to do *x* at time *t*, then he is always free to do *x* at *t*; he just has to curry favor with *A* first. This point is illustrated by Kramer's example where "Barry the large Bully" prevents Ernest from eating a russet apple at time $t_1$.[42] But Barry lets Ernest eat the apple at $t_6$ after Ernest has curried favor with Barry. So, Ernest is unfree to eat the apple at $t_1$, but he is free at $t_1$ to eat it at $t_6$. And if he follows a pattern of self-abasing behavior between $t_1$ and $t_6$, he is free at any intermediate stage to eat the apple at $t_6$. His ingratiation, therefore, does not gain him a new freedom.

Furthermore, *B*'s having to curry favor with *A* to be able to do *x* means that *A* reduces *B*'s overall pure negative freedom since *A* limits the frequency with which *x* is included in the combinations of conjunctively exercisable actions available to *B*. *A* restricts the range of other actions *B* can perform conjunctively with doing *x*. *A*'s power over *B* therefore restricts *B*'s number of available courses of action and thereby reduces *B*'s overall freedom. *A* prevents the conjunctive exercisability of some of *B*'s freedoms by making *x* conjunctively exercisable only with ingratiation. *A* thus interferes with *B* and compromises his freedom.

A plausible way to reduce the probability of such restrictions on citizens' conjunctively exercisable actions, and thus enhance their overall pure negative freedom, is to establish legal institutions that enforce individuals' rights. And the possession of reliably enforced rights should enable citizens to meet Pettit's eyeball test. As Joel Feinberg notes, "having rights enables us to 'stand up like men,' to look others in the eye, and to feel in some fundamental way the equal of anyone."[43] The institutional protection required for passing the eyeball test and consequently realizing republican freedom, as Pettit understands it, therefore seems practically indistinguishable from the institutional arrangement necessary for promoting pure negative freedom. As the heuristic guiding

[40] Carter, *A Measure of Freedom*, 240, and "How Are Power and Unfreedom Related?"; Kramer, *The Quality of Freedom*, 174–78.

[41] Carter, *A Measure of Freedom*, 233–45; Kramer, *The Quality of Freedom*, 174–78.

[42] Kramer, "Liberty and Domination," 50–56, and *The Quality of Freedom*, 144–48.

[43] Feinberg, "The Nature and Value of Rights," 252.

republicans in prescribing social institutions, the eyeball test will therefore not do to distinguish the institutional requirements of republican freedom from institutions promoting pure negative freedom.

Pettit also says that passing the eyeball test requires a government that reliably enforces citizens' rights, especially their rights to exercise the basic liberties. Unlike Pettit, negative-freedom theorists treat the legal coercion involved in protecting citizens' ability to exercise the basic liberties as a source of unfreedom.[44] This difference follows the conceptual difference that while negative-freedom theorists consider any act of interference a source of unfreedom, republicans think only uncontrolled interference can make an agent unfree. But negative-freedom theorists can nonetheless justify such coercion as a way of promoting overall freedom insofar as it gives citizens opportunities they otherwise would not have had. And the basic liberties are especially important for overall pure negative freedom since they are bases for many other freedoms. It is true that a protected freedom of movement, for example, means *A* will be penalized for preventing *B*'s movement, and the number of actions *A* can perform conjunctively with preventing *B*'s movement is therefore restricted. However, this restriction will very likely give both *A* and *B* more opportunities they otherwise would not have had, thus increasing their overall freedom.

To make the institutions constitutive of republican freedom differ from those promoting pure negative freedom, republicans must conceptualize freedom so that it requires the protection of certain specific freedoms without thereby enhancing individuals' opportunity sets. Christian List points toward such a conception of republican freedom by understanding it to require more robust protection against uncontrolled interference than on Pettit's interpretation.[45] On List's account, republican freedom requires that society be organized so that *A* will in no "socially possible world" get away with interfering with *B* in a way *B* has not instructed.[46] A socially possible world, List explains, is "a particular combination of preference orderings across agents in the society."[47] The set of all socially possible worlds is defined by positive—as opposed

[44] To clarify, a law against doing *x* will not take away people's specific freedom to do *x* since they can violate the law. If that were not the case, there would be no criminals. See Steiner, *An Essay on Rights*, ch. 2. An exception is law enforcement by anticipatory preventive measures that close off opportunities to break the law. See Kramer, "Why Freedoms Do Not Exist by Degrees," 233.

[45] List, "Republican Freedom and the Rule of Law." Given this difference between List and Pettit's accounts, it is puzzling that Pettit thinks List "offers a wonderfully clear (and to me, congenial) view of how the liberal and republican approaches compare in their treatment of possibility" ("Freedom and Probability," 207n3).

[46] List, "Republican Freedom and the Rule of Law," 209.

[47] List, "Republican Freedom and the Rule of Law," 212.

to normative—social laws. These laws are regularities in human behavior, such as the law of supply and demand.[48]

On List's account, then, republican freedom requires that uncontrolled interference will not occur regardless of what preferences the individuals constituting society might have. The absence of uncontrolled interference thus becomes a positive social law. Such interference will always remain biologically, physically, and logically possible, but it cannot be socially possible. We might imagine that measures required for achieving this level of robust absence of unchecked power include an effective police force, extensive surveillance, and a highly vigilant citizenry. However, such measures can never be so effective that they rule out the possibility of uncontrolled interference. The impossibility of such protection therefore means that no one can ever have republican freedom, as List defines it, to perform any action.

List's interpretation is impossibly demanding, but it nonetheless points toward a way of separating republican freedom from the promotion of pure negative freedom. In the next section, I show how a conception of republican freedom stronger than Pettit's moderate eyeball-test understanding requires greater institutional protection than is compatible with the promotion of pure negative freedom. This conception is more demanding on the citizens than is Pettit's conception, and therefore closer to List's strong understanding, but it need not be impossibly demanding. What is required for defining republican freedom so that it conflicts with the promotion of pure negative freedom is only that it be made compatible with measures intended to stimulate a higher level of vigilance in the citizenry than is necessary for maintaining institutions that promote individuals' pure negative freedom. But as we shall see, such a stronger understanding of republican freedom conflicts with liberal neutrality.

## 3. The Trade-Off

To see how a stronger interpretation of republican freedom differs from Pettit's moderate one and how the former conflicts with the promotion of pure negative freedom, I introduce two dimensions of freedom: scope and robustness.[49] The scope of a definition of freedom indicates the extent to which it requires the absence of interference. Gerald MacCallum famously treats freedom as a triadic relation, where "*x* is (is not) free from *y* to do (not do, become, not

[48] List, "Republican Freedom and the Rule of Law," 203–5.

[49] In defining these dimensions, I draw on List, "Republican Freedom and the Rule of Law" and "The Impossibility of a Paretian Republican?"; Pettit, "Capability and Freedom."

become) *z*."[50] Here *x* refers to some agent, *y* to a constraint, restriction, or interference, while *z* refers to an action. Scope concerns the *y* variable in this formula. The more types of constraint are considered compatible with freedom, the lesser is the scope of that conception of freedom, and vice versa. A definition that treats all kinds of interference as a loss of freedom, such as pure negative freedom, has maximal scope.

Robustness, on the other hand, refers to the extent to which freedom is understood to require protection against interference. A definition of freedom with maximal robustness requires the absence of interference in *all* socially possible worlds—that is, regardless of the preferences of other agents—whereas one with minimal robustness requires such absence in only the actual world. There may be many intermediate positions between these extremes, and we shall see that republican freedom will occupy one of them.

A trade-off between scope and robustness is unavoidable.[51] To see why, notice that protecting two or more agents against each other's interference itself involves interfering with them. That is, protecting *A* and *B* against each other's interference means preventing them from interfering with one another. We thus see that scope and robustness are inversely related: the greater the protection freedom requires, the more acts of interference it must be compatible with.[52] Any move up along the robustness dimension thus implies a corresponding reduction in scope. By understanding freedom to require more than minimal robustness, we must therefore identify a kind of interference that is not a source of unfreedom, and that means reducing the scope of freedom.[53]

In fact, any definition of freedom with maximal robustness, regardless of its scope, makes freedom impossible.[54] There is no way of ensuring people's ability to perform an action in all socially possible worlds—that is, regardless of

[50] MacCallum, "Negative and Positive Freedom."

[51] Moen, "Freedom and Its Unavoidable Trade-Off."

[52] To be precise, the two dimensions are only roughly inversely related. An asymmetry becomes apparent when we notice that while maximal scope is compatible with minimal robustness, maximal robustness is incompatible with any scope at all. Giving freedom maximal robustness therefore implies that no one is ever free to do anything whatsoever.

[53] Geoffrey Brennan and Alan Hamlin also make this rather obvious observation: "More resilient liberty will necessarily be at the expense of less liberty" (Brennan and Hamlin, "Republican Liberty and Resilience," 54).

[54] List proves the weaker result that an implication of Amartya Sen's liberal paradox is that freedom with maximal robustness conflicts with the Pareto principle. The Pareto principle demands that a collective decision procedure must favor an alternative, *x*, to another alternative, *y*, if all individuals prefer *x* to *y*. See List, "The Impossibility of a Paretian Republican?"; Sen, "The Impossibility of a Paretian Liberal." Surprisingly, List and Valentini have recently proposed a definition of freedom—"freedom as independence"—that maximizes both

others' preferences. Maximal robustness is therefore not even compatible with minimal scope. List gives republican freedom maximal robustness by taking it to require the absence of uncontrolled interference in all socially possible worlds. His account of republican freedom is therefore impossible.

A definition of freedom with maximal scope, however, is possible, but only if combined with minimal robustness. After all, an agent can be said to be free to perform any action, *x*, that no one prevents her from performing insofar as there is no requirement that she can do *x* in other possible worlds. This is indeed the pure negative view. There is thus an asymmetry between scope and robustness that means they are only roughly inversely related. By treating any prevention as a source of unfreedom, pure negative freedom has maximal scope. And it has minimal robustness since only prevention itself, and not the mere possibility of prevention, can make you unfree to perform an action. Only prevention, that is, can take a specific freedom away from a person. We have seen, however, that increasing overall pure negative freedom involves protecting individuals against interference.

Republican freedom, on any interpretation, has a smaller scope than pure negative freedom has since it specifies a kind of interference—controlled interference—that does not contribute to unfreedom. Its scope can therefore be reduced so as to be compatible with a more than minimal level of robustness. How much more than minimal will determine whether it is possible or not. We have seen that Pettit is explicitly vague about the robustness requirement of his moderate account of republican freedom, but I have proceeded on the assumption that his eyeball-test level is realizable. The interference necessary for achieving the required level of robustness is considered controlled and therefore compatible with freedom from domination.

We have also seen, however, that pure negative freedom and moderate republican freedom's different trade-offs are irrelevant when it comes to promoting freedom. Promoting overall pure negative freedom requires the same institutional protection as does Pettit's moderate republican freedom. List is, therefore, wrong when he says that "perhaps [negative-freedom theorists] are also concerned with robustness, but if they are, that concern stems not from their commitment to pure negative freedom [as noninterference] itself, but from their commitment to other desiderata beyond freedom."[55]

List notes that republicans can treat any interference necessary for achieving the level of robustness specified in their definition of freedom as controlled

dimensions ("Freedom as Independence"). For a critique of List and Valentini's freedom concept, see Carter and Shnayderman, "The Impossibility of 'Freedom as Independence.'"

[55] List, "Republican Freedom and the Rule of Law," 218.

and therefore as compatible with freedom. On List's very strong interpretation, this means interfering to an extent that no one can do anything whatsoever. However, for republicans to define freedom so that its institutional requirements conflict with those of pure negative freedom, they need not endorse such a strong understanding. The level of robustness must be greater than on Pettit's moderate account, but it need not go all the way up to List's impossibly demanding interpretation. For republican freedom to conflict with the promotion of pure negative freedom, it must only be given enough robustness to imply a reduction in the total number of actions citizens can perform. Promoting republican freedom will then be to reduce people's pure negative freedom. The way to realize Pettit's view that republicans demand more robust institutional protection than do proponents of pure negative freedom is therefore to enhance the robustness requirement of republican freedom beyond Pettit's moderate level and, consequently, to expand the set of controlled interference.

## 4. SPECIFYING REPUBLICAN PREFERENCES

We have seen that the scope restriction in Pettit's moderate interpretation of republican freedom is compatible with measures for increasing citizens' range of courses of action, which implies promoting pure negative freedom. To avoid this result and achieve his desired break from freedom as noninterference, and pure negative freedom in particular, Pettit must reduce the scope of republican freedom further by making more kinds of interference compatible with freedom, thus restricting the range of courses of action citizens can pursue. In this section, I consider the preferences compatible with this stronger interpretation of republican freedom. This is an important step toward showing why a conception of republican freedom that conflicts with the promotion of pure negative freedom will also conflict with liberal neutrality.

On any account of republican freedom, the citizens are required to keep an eye out for behavior they perceive as incompatible with their common interests and be ready to contest such behavior and make sure it ceases. With respect to the government–citizen relationship, the government's power is not freedom-reducing as long as the citizens will successfully resist decisions not tracking their common interests.

Pettit also acknowledges that no matter how well-designed formal institutions are, their success in consolidating popular control depends on individuals' behavior.[56] Citizens must "always insist on the authorities going through the

[56] Pettit, *Republicanism*, ch. 8.

required hoops in order to prove themselves virtuous."[57] "People must be on the watch for proposals or measures that are not suitably supported … and they must be ready to organize in opposition to such policies."[58] They must be ready to contest decisions of government officials—elected or unelected—via channels such as the courts, the press, demonstrations in the streets, or by contacting their representative in parliament or an ombudsman.[59]

For Pettit, two factors indicate the level of popular control and, consequently, the extent to which citizens enjoy republican freedom: first, how disposed people are to resist perceived abuses of governmental power; and second, how disposed government officials are to be inhibited by actual or potential resistance.[60] Citizens must continuously give government officials reasons to use their power for the good of society so that a pattern of government action for the good of society remains robust, whether officials are virtuous or not.[61]

It is important to note that popular control is achieved by external constraints, without which the powerholders will not rule in the citizens' interests because the citizens have instructed it. In the absence of such constraints, government officials might rule in accordance with common interests as a matter of goodwill but not due to popular control. Such goodwill might reduce the probability of uncontrolled interference but not its accessibility. Popular control, Pettit stresses, is based on society's resistive character, not on the goodwill of government officials.[62] Without denying that government officials can be genuinely virtuous, Pettit takes the relevant constraint to be the external constraint citizens impose on the government officials and not the officials' inner constraint—that is, their moral commitment to promoting common interests.[63]

On his moderate understanding, Pettit takes republican freedom to require citizens' "virtual control" of their government, which means citizens can go about their lives as they wish as long as they are ready to blow the whistle should they become aware of power abuse.[64] They can remain in standby mode while being ready to speak out if "the red lights go on," as Pettit says.[65] Virtual control is consequently not particularly demanding on the citizens.

[57] Pettit, *Republicanism*, 264.

[58] Pettit, *On the People's Terms*, 226.

[59] Pettit, *On the People's Terms*, 237, and *Republicanism*, 193.

[60] Pettit, *On the People's Terms*, 174.

[61] Pettit, *On the People's Terms*, 124.

[62] Pettit, *On the People's Terms*, 174.

[63] Pettit, *Republicanism*, 211.

[64] Pettit, "Republican Freedom," 103, 111–13.

[65] Pettit, *On the People's Terms*, 136n5.

And importantly for the purposes of this paper, by not making active political engagement part of the ideal, moderate republicanism is compatible with the neutrality between conceptions of the good that Pettit favors. This is also why Rawls sees republicanism as neutral in this sense.[66]

On a stronger interpretation of republican freedom, on the other hand, the commitment to robust protection goes beyond virtual control. Greater commitment to virtuous behavior will always have a positive effect on the two factors indicating the level of republican freedom people enjoy. The more vigilant and ready to contest the citizens are, the greater is their protection against uncontrolled government interference. Their disposition for vigilance and contestation will increase the firmer their commitment is, and government officials will have a better reason to feel inhibited. "Active control" requires that citizens devote more of their lives to protecting their society against unchecked power. They cannot just report abuse of power whenever they happen to come across it; they must actively search for it.

To contribute to the robustness of the protection against unchecked power, active control must itself be robust—that is, citizens must be vigilant and ready to contest in a wide range of socially possible worlds. Such robustness is achieved by social norms that shape citizens' preferences. Pettit also notes that people's freedom depends on "the power of established norms."[67] Without norms motivating citizens to keep political power in check, legitimate government will remain "an unattainable ideal," he says.[68] Social norms are constituted by people's expectations of one another to conform to certain behavior and their desire to meet these expectations, as well as their approval of such conformity and disapproval of deviations. Under norms required for a stronger conception of republican freedom, citizens will expect each other to act so as to maintain a more active form of control—that is, to be actively vigilant and ready to contest. Deviations from this behavioral pattern will be met with social disapproval. In Pettit's own vocabulary, we may say that republican freedom requires an "economy of esteem" in which citizens give each other esteem for acting in accordance with active control and disesteem for not doing so.[69]

We therefore see that a stronger account of republican freedom requires that citizens adopt preferences compatible with a more active form of control than

[66] Rawls, *Justice as Fairness*, 142–44; Rawls, *Political Liberalism*, 205–6. Elsewhere, I argue against Pettit that there are no significant conflicts between his "liberal republicanism" and Rawls's "political liberalism." See Moen, "Eliminating Terms of Confusion" and "Republicanism as Critique of Liberalism."

[67] Pettit, *Republicanism*, 246.

[68] Pettit, *On the People's Terms*, 262.

[69] Brennan and Pettit, *The Economy of Esteem*.

does Pettit's moderate understanding. These are preferences compatible with a higher level of vigilance and readiness to contest one's government, as well as of monitoring one's fellow citizens to make sure they, too, are committed to active popular control.

## 5. Neutrality

Adjusting the scope dimension to strengthen robustness is to block the pursuit of some possible conceptions of the good. Making citizens form the republican preferences required by a stronger than moderate version of republican freedom is to make them conform to a comprehensive doctrine compatible with active control. A stronger account of republicanism is therefore incompatible with liberal neutrality, which requires the state to give no special advantages or disadvantages to rival conceptions of the good.[70] This may not be so surprising. After all, the ancient and Renaissance city-states commonly associated with republicanism were characterized by a conformist population, especially compared to today's large, pluralistic societies. Cass Sunstein also notes that republicans have traditionally argued that the polity should inculcate civic virtue in its population, which modern observers might see as an impermissible "imposition of a 'comprehensive doctrine' on the population."[71]

Pettit, however, takes a modern approach by looking at "political institutions with quite a different attitude from that of premodern republicans."[72] He understands freedom as nondomination as a primary good—that is, it is a good everyone would want no matter what her or his conception of the good

[70] Frank Lovett and Gregory Whitfield also argue that republicanism is incompatible with such neutrality of treatment, as well as neutrality of effect, which is more obvious since it is, as they say, "widely regarded as chimerical" ("Republicanism, Perfectionism, and Neutrality," 125). They further argue that republicanism cannot be impartial, and take impartiality to require that "public policies, institutions, and so forth be justifiable to all persons, regardless of their conception of the good, provided they are ready and willing to engage in social cooperation with others on fair terms" (129). Given the similarities I have identified between Pettit's republicanism and Rawls's political liberalism, Lovett and Whitfield implicitly also question the neutrality and impartiality of the latter. Kramer indeed denies that liberalism can be impartial in his *Liberalism with Excellence*, ch. 3. But Lovett and Whitfield take republicanism to satisfy a "principle of toleration," according to which "public policies, institutions, and so forth should impose no special disadvantages on any worthwhile conception of the good" (124). Lovett and Whitfield use the modifier "worthwhile" to exclude conceptions with "no possible benefit for those who hold them" (125). But promoting a stronger kind of republican freedom that conflicts with the promotion of pure negative freedom will involve imposing "special disadvantages" on certain "worthwhile conceptions of the good."

[71] Sunstein, "Republicanism and the Preference Problem," 181.

[72] Pettit, *Republicanism*, 95.

might be.[73] Pettit can therefore understand republicanism to seek "a relatively neutral brief for the state—a brief that is not tied to any particular conception of the good."[74] And this view seems plausible in the sense that having a right to exercise one's basic liberties is in everyone's interests.[75] But while everyone might benefit from robust protection of the basic liberties, it is doubtful that promoting nondomination in a way that conflicts with the promotion of pure negative freedom would be in everyone's interest.

In a modern, pluralistic society, many citizens will be content with a sufficiently low probability of someone restricting their ability to exercise the basic liberties—sufficiently low to meet the eyeball test, we might conjecture. But the more extensive vigilance required by a stronger conception of republican freedom will to some people involve a too-costly sacrifice of personal projects. These people want to pursue ends that conflict with devoting a significant part of their lives to making sure no one gets away with interference conflicting with their common interests. So, while the basic liberties may constitute a primary good, nondomination, on this stronger understanding, does not.[76]

Nondomination looks different from the perspective of Pettit's moderate republicanism, of course, since he takes it to require no more than virtual control. Since virtual control lets people lead their lives in accordance with a wide range of comprehensive doctrines, Pettit can understand his theory as compatible with liberal neutrality. It meets his criterion that "any plausible political ideal must be an ideal for all."[77] The republican ideal, he says, is "capable of commanding the allegiance of the citizens of developed, multicultural societies, regardless of their more particular conceptions of the good."[78] "Multicultural concerns," he says, "can be supported by an appeal to [freedom as nondomination]."[79] This view clearly conflicts with a stronger understanding of republican freedom, the pursuit of which involves restricting individuals' available

[73] Pettit, *Republicanism*, 90–92.

[74] Pettit, *Republicanism*, 120.

[75] For Pettit's account of the basic liberties, see note 16 above. Rawls also considers "the basic rights and liberties" as one kind of primary good (*Justice as Fairness*, 58).

[76] Lovett and Whitfield suggest a different reason for thinking nondomination is no primary good: some reasonable persons, they argue, manage to pursue their conceptions of the good while being dominated, such as women who cannot do so "unless subordinate to the unaccountable authority of a husband" (Lovett and Whitfield, "Republicanism, Perfectionism, and Neutrality," 131). Lovett and Whitfield do not themselves, however, deny that nondomination is a primary good. They regard this only as evidence for republicanism not being impartial.

[77] Pettit, *Republicanism*, 96.

[78] Pettit, *Republicanism*, 96.

[79] Pettit, *Republicanism*, 144.

courses of action and therefore their ability to develop and pursue different conceptions of the good.

While a stronger interpretation that conflicts with liberal neutrality is necessary for distinguishing republican freedom from the promotion of pure negative freedom, it is worth considering whether such an interpretation is conceptually possible. In addition to its conflict with neutrality, I have found two reasons Pettit gives for rejecting a stronger active-control view of republican freedom. First, not being constantly monitored gives government officials a feeling of being trusted, which motivates them to make good decisions in the interests of society.[80] The trustee wants the good opinion of the trustor, or of others witnessing the act of trust, and will therefore be motivated not to let the trustor down.[81] In other words, if *B* trusts *A*, then *A* will be more motivated to act on *B*'s instructions than if *B* had not shown *A* that he trusts her. Active control is incompatible with such showing of trustworthiness and therefore prevents this beneficial effect.

Whether or not trust actually has this beneficial effect is, of course, an empirical question. But I need not go into that here because I can respond conceptually by pointing out that trust can add nothing to the robustness of government officials acting for the good of society. Trust might reduce the probability of uncontrolled interference, but it imposes no external constraint on the government officials so as to make their power abuse absent from other socially possible worlds. Trust might lower the probability of government officials exercising unchecked power, but it does not deny them such power.

Pettit's second reason for objecting to active control is that it would be pointless to try to make citizens more virtuous than they actually prefer to be since that would cause more domination than it prevents.[82] Interference to stimulate the required virtue in the citizens will therefore be uncontrolled, as it will cause a greater loss of overall nondomination than it gains in robust protection against unchecked power. Pettit therefore argues for individual rights to protect citizens against being forced to contribute to ends they do not endorse. So, if an agent, *A*, faces a choice between doing *x* or *y*, and only *x*-ing is compatible with active control, republicans can still grant *A* the right to choose for herself whether to do so or not.

This view is further strengthened by Pettit's observation that forcing people to contribute to ends they might not endorse could make them less motivated

[80] Pettit, *Republicanism*, 268–69.

[81] Pettit, "The Cunning of Trust."

[82] Pettit, *Republicanism*, 173.

to contribute than if they had the opportunity to do so voluntarily.[83] After all, there is no point in trying to maximize the range of possible worlds without uncontrolled interference by interfering contrary to people's instructions in the actual world. Republicanism requires that the state be forced to track the citizens' interests, and these interests will probably conflict with protecting their society against unchecked power to the extent required by active control.

But Pettit's second response just says that the commitment to vigilance and contestation must be voluntary. It does not say that republicanism must be neutral. Republicans may grant individuals legal rights that protect them against uncontrolled interference, but they can still be concerned with how citizens use the freedoms granted by these rights.[84] So, if a right makes *A* free to choose *x* or *y*, and *x* will best promote nondomination, then the republican should look for noncoercive ways of making *A* voluntarily choose *x* rather than *y*. Republicans should interfere in subtle ways to alter citizens' preferences to make them willingly commit to the behavioral pattern of active control. Well-designed interventions might include nudges, which organize agents' opportunity sets so as to make them more likely to choose beneficial options without undermining their sense of autonomous decision-making.[85] Other measures, such as compulsory civic education, might also serve this end.

A very strong account of republican freedom might follow American revolutionary Benjamin Rush and "convert men into republican machines." A republican pupil, Rush said, should "be taught that he does not belong to himself, but that he is public property."[86] However, support for such extensive measures is

[83] Pettit, *Republicanism*, 256.

[84] I doubt that support for legal rights is sufficient for satisfying Lovett and Whitfield's principle of toleration (see note 70 above). But if it is, then I agree with them both that republicanism is incompatible with neutrality and impartiality, and that it is compatible with toleration.

[85] A nudge is meant to serve the interests of the person who is nudged. See Thaler and Sunstein, *Nudge*. In this case, the nudge is meant to benefit society as a whole. However, republicans have traditionally thought of the individual's best interests and society's best interests as inseparable. Acting against society's best interests is the definition of corruption in the republican literature. "Corruption," as Skinner explains, "is simply a failure of rationality, an inability to recognise that our own liberty depends on committing ourselves to a life of virtue and public service" ("The Republican Ideal of Political Liberty," 304). For a classic expression of this republican view, see Cicero, *On Obligations*, esp. 7–8.

[86] Quoted in Wood, *The Creation of the American Republic*, 427. One reason why Lovett and Whitfield deny that republicanism can be neutral is that it can support "education or other policies designed to inculcate a patriotic love of republican institutions" (Lovett and Whitfield, "Republicanism, Perfectionism, and Neutrality," 127). Weithman, similarly, points out that republicans can only support measures intended to stimulate political participation and civic virtue on the basis of a perfectionist argument for why such civic-mindedness

not necessary for republicans to defend institutions that differ from institutions promoting pure negative freedom. Republican institutions only need to reduce the number of permissible courses of action to enhance the protection of these favored courses of action. This implies favoring some comprehensive doctrines at the expense of others. The strength of the conception of republican freedom—that is, the extent to which it prioritizes robustness over scope—will determine how many comprehensive doctrines it is compatible with. But no conception strong enough to conflict with the promotion of pure negative freedom can be justified on the basis of neutrality; it must instead be based on a view of some comprehensive doctrines as more valuable than others.[87]

Now, Pettit might object by arguing that in a modern, pluralistic society, any attempt—coercive or noncoercive—to promote a particular conception of the good conflicts with republican freedom. But by accepting this constraint, which Rawls seems to do by viewing republicanism as compatible with political liberalism, republicans are bound to promote pure negative freedom.[88] Republicans can take Pettit's moderate line and defend liberal neutrality, but we have seen how this implies condoning the same institutional arrangements as do proponents of pure negative freedom.

## 6. Conclusion

Pettit faces a dilemma. His republicanism cannot both be neutral and conflict with the promotion of pure negative freedom. He can maintain neutrality by requiring a moderate, eyeball-test level of robustness, but that means promoting pure negative freedom. Alternatively, he can sustain his view that pursuing republican freedom conflicts with the promotion of pure negative freedom by reducing the scope of republican freedom. But on that account, promoting republican freedom involves promoting some comprehensive doctrines at the expense of others, which means undermining neutrality.

Pettit rejects stronger accounts of republican freedom by condemning interference intended to change people's preferences so that they willingly commit to active control. By saying we should let people decide for themselves how

constitutes an intrinsic good for individuals ("Political Republicanism and Perfectionist Republicanism"). With Lovett, Pettit argues that promoting the commitment to civic virtue necessary for republican freedom "requires a fairly robust program of civics education" (Lovett and Pettit, "Neorepublicanism," 23). Given his commitment to neutrality, it is unlikely that Pettit has in mind an educational program anything like the one Rush proposed.

[87] I elaborate on "comprehensive republicanism" in Moen, "Republicanism as Critique of Liberalism."

[88] Rawls, *Justice as Fairness*, 142–43, and *Political Liberalism*, 205–6.

much to contribute to the protection against unchecked power, he consequently rejects a kind of interference that we would understand to enhance popular control, and therefore not perceive as a source of unfreedom, on an account of republicanism that conflicts with pure negative freedom. By condemning such interference as at odds with freedom, Pettit condones institutions promoting pure negative freedom.

It is not hard to see why Pettit prefers moderate republicanism to a stronger, more demanding account. A political ideal, he says, must be achievable by democratic means in an actual society.[89] And the pluralism characterizing modern society might make it unlikely that we can transparently and democratically adopt policies intended to alter citizens' preferences in the way the robustness condition of a strong, but not necessarily impossibly demanding, conception of republican freedom requires. By making his theory sensitive to this fact of pluralism, Pettit comes up with an ideal that may well be an attractive aim for a modern society. But by doing so, he does not just give republican freedom a modern interpretation—he also promotes a freedom concept he has repeatedly claimed to reject: freedom as noninterference.[90]

*University of Vienna*
*lars.moen@univie.ac.at*

## REFERENCES

Berlin, Isaiah. *Liberty*. Edited by Henry Hardy. Oxford: Oxford University Press, 2002.

Brennan, Geoffrey, and Alan Hamlin. "Republican Liberty and Resilience." *The Monist* 84, no. 1 (January 2001): 45–59.

Brennan, Geoffrey, and Philip Pettit. *The Economy of Esteem: An Essay on Civil and Political Society*. New York: Oxford University Press, 2004.

Carter, Ian. "How Are Power and Unfreedom Related?" In Laborde and Maynor, *Republicanism and Political Theory*, 58–82.

[89] Pettit, "Political Realism Meets Civic Republicanism," 341–43, and "The Domination Complaint," 89.

[90] I am grateful to Keith Dowding and Philip Pettit for their comments on an early draft of this paper. I also thank Matteo Boccacci, Ian Carter, Annalisa Costella, Michael Garnett, Ilkin Huseynli, and Elena Icardi for their helpful feedback at a workshop at the University of Pavia. My work on the paper has in part been funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement number 740922).

———. *A Measure of Freedom*. Oxford: Oxford University Press, 1999.

Carter, Ian, and Ronen Shnayderman. "The Impossibility of 'Freedom as Independence.'" *Political Studies Review* 17, no. 2 (May 2019): 136–46.

Cicero, Marcus Tullius. *On Obligations*. Translated by P. G. Walsh. Oxford: Oxford University Press, 2008.

Dowding, Keith. "Republican Freedom, Rights, and the Coalition Problem." *Politics, Philosophy & Economics* 10, no. 3 (August 2011): 301–22.

Feinberg, Joel. "The Nature and Value of Rights." *Journal of Value Inquiry* 4, no. 4 (December 1970): 243–60.

Gaus, Gerald F. "Backwards into the Future: Neorepublicanism as a Postsocialist Critique of Market Society." *Social Philosophy and Policy* 20, no. 1 (January 2003): 59–91.

Goodin, Robert E., and Frank Jackson. "Freedom from Fear." *Philosophy and Public Affairs* 35, no. 3 (Summer 2007): 249–65.

Hobbes, Thomas. "Selections from *The Question Concerning Liberty, Necessity, and Chance*." In *Hobbes and Bramhall on Liberty and Necessity*, edited by Vere Chappell, 69–90. Cambridge: Cambridge University Press, 1999.

Ingham, Sean, and Frank Lovett. "Republican Freedom, Popular Control, and Collective Action." *American Journal of Political Science* 63, no. 4 (October 2019): 774–87.

Kramer, Matthew H. "Freedom and the Rule of Law." *Alabama Law Review* 61, no. 4 (2010): 827–45.

———. *Liberalism with Excellence*. Oxford: Oxford University Press, 2017.

———. "Liberty and Domination." In Laborde and Maynor, *Republicanism and Political Theory*, 31–57.

———. *The Quality of Freedom*. Oxford: Oxford University Press, 2003.

———. "Why Freedoms Do Not Exist by Degrees." *Political Studies* 50, no. 2 (June 2002): 230–43.

Laborde, Cécile, and John Maynor, eds. *Republicanism and Political Theory*. Oxford: Blackwell, 2008.

List, Christian. "Republican Freedom and the Rule of Law." *Politics, Philosophy and Economics* 5, no. 2 (June 2006): 201–20.

———. "The Impossibility of a Paretian Republican? Some Comments on Pettit and Sen." *Economics and Philosophy* 20, no. 1 (April 2004): 65–87.

List, Christian, and Laura Valentini. "Freedom as Independence." *Ethics* 126, no. 4 (July 2016): 1043–74.

Lovett, Frank, and Philip Pettit. "Neorepublicanism: A Normative and Institutional Research Program." *Annual Review of Political Science* 12, no. 1 (2009): 11–29.

———. "Preserving Republican Freedom: A Reply to Simpson." *Philosophy* *and Public Affairs* 46, no. 4 (Fall 2018): 363–83.

Lovett, Frank, and Gregory Whitfield. "Republicanism, Perfectionism, and Neutrality." *Journal of Political Philosophy* 24, no. 1 (March 2016): 120–34.

MacCallum, Gerald C. "Negative and Positive Freedom." *Philosophical Review* 76, no. 3 (July 1967): 312–34.

Moen, Lars J. K. "Eliminating Terms of Confusion: Resolving the Liberal–Republican Dispute." *Journal of Ethics* 26, no. 2 (June 2022): 247–71.

———. "Freedom and Its Unavoidable Trade-Off." *Analytic Philosophy* (forthcoming). Published online ahead of print, March 27, 2023. https://doi.org/10.1111/phib.12301.

———. "Republicanism and Moralised Freedom." *Politics, Philosophy and Economics* 22, no. 4 (November 2023): 423–40.

———. "Republicanism as Critique of Liberalism." *Southern Journal of Philosophy* 61, no. 2 (June 2023): 308–24

Pettit, Philip. "The Basic Liberties. In *The Legacy of H. L. A. Hart: Legal, Political, and Moral Philosophy*, edited by Matthew H. Kramer, Claire Grant, Ben Colburn, and Antony Hatzistavrou, 201–22. Oxford: Oxford University Press, 2008.

———. "Capability and Freedom: A Defence of Sen." *Economics and Philosophy* 17, no. 1 (April 2001): 1–20.

———. "The Cunning of Trust." *Philosophy and Public Affairs* 24, no. 3 (July 1995): 202–25.

———. "The Domination Complaint." In *Nomos 46: Political Exclusion and Domination*, edited by Melissa S. Williams and Stephen Macedo, 87–117. New York: New York University Press, 2005.

———. "Freedom and Probability: A Comment on Goodin and Jackson." *Philosophy and Public Affairs* 36, no. 2 (Spring 2008): 206–20.

———. "The Instability of Freedom as Noninterference: The Case of Isaiah Berlin." *Ethics* 121, no. 4 (July 2011): 693–716.

———. *Just Freedom: A Moral Compass for a Complex World*. London: W. W. Norton and Co., 2014.

———. *On the People's Terms: A Republican Theory and Model of Democracy*. Cambridge: Cambridge University Press, 2012.

———. "Political Realism Meets Civic Republicanism." *Critical Review of International Social and Political Philosophy* 20, no. 3 (2017): 331–47.

———. "Republican Freedom: Three Axioms, Four Theorems." In Laborde and Maynor, *Republicanism and Political Theory*, 102–30.

———. *Republicanism: A Theory of Freedom and Government*. Oxford: Oxford University Press, 1997.

———. *A Theory of Freedom: From the Psychology to the Politics of Agency.*

Cambridge: Polity, 2001.

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

———. *Political Liberalism*. Expanded ed. New York: Columbia University Press, 2005.

Sen, Amartya. "The Impossibility of a Paretian Liberal." *Journal of Political Economy* 78, no. 1 (January–February 1970): 152–57.

Simpson, Thomas W. "Freedom and Trust: A Rejoinder to Lovett and Pettit." *Philosophy and Public Affairs* 47, no. 4 (Fall 2019): 412–24.

———. "The Impossibility of Republican Freedom." *Philosophy and Public Affairs* 45, no. 1 (Winter 2017): 27–53.

Skinner, Quentin. "Freedom as the Absence of Arbitrary Power." In Laborde and Maynor, *Republicanism and Political Theory*, 83–101.

———. *Liberty before Liberalism*. Cambridge: Cambridge University Press, 1998.

———. "The Republican Ideal of Political Liberty." In *Machiavelli and Republicanism*, edited by Gisela Bock, Quentin Skinner, and Maurizio Viroli, 293–309. Cambridge: Cambridge University Press, 1991.

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

———. "How Free: Computing Personal Liberty." *Royal Institute of Philosophy Supplement* 15 (March 1983): 73–89.

Sunstein, Cass R. "Republicanism and the Preference Problem." *Chicago-Kent Law Review* 66, no. 1 (April 1990): 181–203.

Thaler, Richard H., and Cass R. Sunstein. *Nudge: Improving Decisions about Health, Wealth and Happiness*. London: Penguin, 2009.

Viroli, Maurizio. *Republicanism*. Translated by Antony Shugaar. New York: Hill and Wang, 2002.

Weithman, Paul. "Political Republicanism and Perfectionist Republicanism." *Review of Politics* 66, no. 2 (Spring 2004): 285–312.

Wood, Gordon S. *The Creation of the American Republic, 1776–1787*. Chapel Hill, NC: University of North Carolina Press, 1998.

*Journal of Ethics and Social Philosophy*
vol. 26, no. 2 · december 2023

*https://doi.org/10.26556/jesp.v26i2.1433*

# RESOLUTION AND RESOLVE

## RATIONALLY RESISTING TEMPTATION

*Abigail Bruxvoort*

Folk wisdom tells us that if we have any hope of achieving something difficult, we will have to make a resolution to do the thing in question. We do not stumble by accident into running ultramarathons, being good romantic partners, or writing books: we first make resolutions about these things.[1] After all, in the absence of a resolution to do a difficult thing, we can simply change our mind and decide to do something else at any point. Resolutions, in other words, draw a line between intentions that we can revise on a whim and those that we cannot rationally revise without special justification.

But are resolutions appropriate philosophical fodder? Might they be instead a matter for empirical study? After all, one natural question to raise about resolutions is whether they are of any *use*. This is clearly an empirical question, and an important one at that. However, prior and adjacent to this empirical question about resolution are philosophical questions. They are:

1. What is a resolution?
2. Why should we grant resolutions rational authority over our actions, given that the moment of temptation involves the (re)evaluation that it would be best to act on the temptation?
3. Are resolutions a uniquely rational way to resist temptation? Are there any uniquely rational means of resisting temptation?

The first question matters but is to my mind the least interesting of the three questions. It is clear that resolution is some kind of extra-committed intention: I might form the intention to have fried rice for lunch because I like fried rice and I have leftovers readily available in the fridge, or I might form the intention to have fried rice for lunch because I am committing to a gluten-free diet and fried rice is the only gluten-free meal I have available in the home. In the case

[1] Of course, many cases of romantic partnership involve something beyond a resolution: a promise to another person. However, it is also common and natural to make resolutions regarding important relationships.

where I form the intention to eat fried rice just because I like it and it is available, I clearly remain open to changing my mind. Perhaps lunchtime rolls around and I decide to have quesadillas instead. But in the second case, where I resolve to have fried rice, it is clear that my intention carries with it the further thought that I should not change my mind or go back on this intention unless there are extenuating circumstances.

When we ask what a resolution is, we are trying to understand this element of "extra commitment." Does the extra commitment take the form of a second-order intention, a first-order intention plus the intention to not reconsider? Or an intention-desire pair, where we have a first-order intention and a desire to not reconsider? It seems to me that philosophers are most likely to answer question one in light of their answer to question three. I will say more about question three shortly, but if our focus is on giving an account of why resolution would help an ideally rational agent to accomplish her goals (without committing to empirical claims about how actual, nonideal agents manage to accomplish their goals), then it is natural to allow our answer to question three to drive our answer to question one, as I think is the case with the two accounts of resolution I will consider below.

The second question is the one that preoccupies most of the existing philosophical literature on resolutions.[2] For example, the following case posed by Bratman is representative of the cases motivating this literature:

> Consider Ann. She enjoys a good read after dinner but also loves fine beer at dinner. However, she knows that if she has more than one beer at dinner she cannot concentrate on her book after dinner. Prior to dinner Ann prefers an evening of one beer plus a good book to an evening with more than one beer but no book. Her problem, though, is that each evening at dinner, having drunk her first Pilsner Urquell, she finds herself tempted by the thought of a second: For a short period of time she prefers a second beer to her after-dinner read. This new preference is not experienced by her as compulsive.[3]

The question is whether it would be rational for Ann to act on a prior preference given that her preferences have now changed. However, for the most part I will set aside this problem in this article: I do not assume that our preferences

[2] The current philosophical literature on resolution is concerned primarily with this question. See, for instance, Andreou, "Temptation, Resolutions, and Regret"; Bratman, "Toxin, Temptation, and the Stability of Intention," "Temptation Revisited," and "Temptation and the Agent's Standpoint"; and Paul, "Diachronic Incontinence Is a Problem in Moral Philosophy."

[3] Bratman, "Toxin, Temptation, and the Stability of Intention," 74.

in each moment determine what is most rational for us to do in that moment. I will not provide a full argument for this claim here, but in brief, it seems to me that this way of thinking uncritically adopts many of the assumptions of, for instance, rational choice theory in economics, and moral philosophers should not feel beholden to adopt such assumptions, or at least not adopt them uncritically.

The final question, then, is the primary object of my attention in this paper: Should we expect resolutions to be of any use in resisting temptation? As I have already mentioned, this question is closely related to empirical questions about the efficacy of resolutions. However, there is space to consider this question philosophically. We might for instance hold fixed the idea that resolutions are at least of *some* use in resisting temptation and then attempt to give an account of why resolutions help rational agents resist temptation. Or, we might think in terms of an ideally rational agent: not so ideally rational that they do not experience temptation at all, but ideally rational enough to respond to temptation in a fully rational way. While a less rational agent might need to use commitment devices or other such strategies that tackle temptation "sideways," as it were, we might hope that resolution is a way for this ideally rational agent to tackle temptation head on.[4]

In the paper, I will first consider two dueling accounts of resolution, those put forth by Richard Holton and Alida Liberman. Ultimately, I think both accounts go awry in their failure to think about the nature of tempting desire. Tempting desires involve the presentation of reasons, and we cannot use resolution understood as an intention, reason, or desire to simply "hold off" temptation. Instead, in the final section of the paper I argue that the most rational way to resist temptation is to cultivate one's character and agency such that you are not tempted.

## 1. The Nature of Resolution

In his book *Willing, Wanting, Waiting*, Richard Holton introduces the idea of a resolution, writing, "Resolutions serve to overcome the desires or beliefs that

[4] The idea of a commitment device is most often used in economics, psychology, and public health. Formally speaking, "a commitment device [is] an arrangement entered into by an agent who restricts his or her future choice set by making certain choices more expensive, perhaps infinitely expensive, while also satisfying two conditions: (a) The agent would, on the margin, pay something in the present to make those choices more expensive, even if he or she received no other benefit for the payment, and (b) the arrangement does not have a strategic purpose with respect to others" (Bryan, Karlan, and Nelson, "Commitment Devices," 673). A classic example of a commitment device is Odysseus tying himself to the mast of the ship in order to hear the sirens.

the agent fears they will form by the time they come to act, desires or beliefs that will inhibit them from acting as they now plan."[5] So the purpose of a resolution is to hold off beliefs and desires that might prevent me from acting as I now intend. But what is a resolution? According to Holton, a resolution is a pair of intentions, one first order and one second order. It is "both an intention to engage in a certain action, and a further intention to not let that intention be deflected."[6] So someone who resolves to quit smoking forms a first-order intention to quit, and a second-order intention to stick with the first-order intention. This second-order intention does not generate new reasons (we would have a bootstrapping problem if it did), but instead entrenches the first-order decision in response to reasons for not reconsidering that decision.[7]

There is something intuitive about this view of resolution, since it captures the phenomenology of resolving: deciding to do something, and furthermore deciding to not let yourself be distracted or deterred from your goal. However, in the following I will consider an objection raised against this account of resolution by Alida Liberman in her article, "Reconsidering Resolutions."

Liberman argues that resolutions understood as two-tier intentions are not effective in resisting temptation. She summarizes her objection as follows:

> It seems that the second-order intention should succumb to the same temptation to which the first-order intention is susceptible.... Why do the very same considerations that tempt you toward watching yet another episode of your favorite TV show—say, your burning desire to find out what happens next, and your aversion to working—not *also* tempt you to reconsider your resolution to turn off the TV and get to work on your paper?[8]

Liberman develops this objection through an argument she calls "Temptation Transmission." The argument comes in two stages. The first stage lays out two background principles, and the second gives the actual argument, which consists of three premises.[9] The details of her argument are compelling, but for our purposes we need only consider her conclusion. She writes:

[5] Holton, *Willing, Wanting, Waiting*, 77.

[6] Holton, *Willing, Wanting, Waiting*, 11.

[7] Holton, *Willing, Wanting, Waiting*, 146.

[8] Liberman, "Reconsidering Resolutions," 4.

[9] The first background principle is: "Temptation Claim: Temptation works by altering the appearances in favor of there being a reason to do the tempting thing, from the agent's perspective" (Liberman, "Reconsidering Resolutions," 6). The other background principle is simply a version of means-end transmission: "Rational-Means Reasons Transmission (RMRT): Where *E* is an intentional action, if it appears to a rational agent *A* that (1) there

> Since the temptation to $\phi$ leads to the appearance of an *equally strong* reason to abandon the first-order intention to $\phi$ *and* to abandon the resolute second-order intention, the second-order intention cannot do any meaningful work in blocking temptation and preventing judgment shift.[10]

In short, Liberman argues that resolutions do not work and that the second-order intention is not an effective source of rational resistance against temptation because that intention is itself vulnerable to temptation's effects.

Liberman goes on to offer her own solution to the efficacy objection. The third premise in her argument relies crucially on the rational requirement to avoid *akrasia*, and so Liberman thinks that we must identify a mental state that is not subject to an "anti-*akrasia* norm" to block temptation. She turns to desire to play this role.

> *Second-Order Desire Account* (SODA): Resolving to $\phi$ involves intending to $\phi$, and *desiring* not to reconsider the intention to $\phi$.[11]

On this view, then, a resolution is a (first-order) intention coupled with a (second-order) desire. For example, if I resolve to clean my office, I intend to clean *and* I desire to not reconsider my intention. Liberman thinks that there are two notable advantages to her view over Holton's. One, it does not fall prey to the Temptation Transmission argument because the argument responds specifically to resolution understood as a two-tier intention, and two, it better captures data about resolutions.[12]

is a reason of strength *X* for an agent *A* to attain end *E*, and (2) *M* is the only rationally permissible way to attain *E*, then there will appear to *A* to be a reason of strength *X* for *A* to do *M*" ("Reconsidering Resolutions," 6). Then, the actual argument appears as follows: "(1) When $\phi$ ing is an action about which an agent has formed a resolution, an apparent reason to $\phi$ (stemming from temptation) necessarily leads to an apparent reason to intend to $\phi$. (2) An apparent reason to intend to $\phi$ necessarily leads to an apparent reason to reconsider the intention to avoid $\phi$-ing (call this intention 'Intention 1'). (3) An apparent reason to reconsider Intention 1 necessarily leads to an apparent reason to reconsider the intention not to reconsider Intention 1 (call this intention 'Intention 2')" ("Reconsidering Resolutions," 9).

[10] Liberman, "Reconsidering Resolutions," 15.

[11] Liberman, "Reconsidering Resolutions," 18.

[12] Liberman notes that resolution comes in degrees and that some are stronger than others, and "appealing to desire as a necessary component of a resolution gives us an easy and efficient explanation of how resolutions can vary in strength. . . . We can explain the strength of a resolution as a direct result of the strength of the agent's desire to avoid reconsideration" ("Reconsidering Resolutions," 20). She further claims that the strength of the resolution desire determines the degree to which the agent is successful in resisting temptation. She writes:

> In general, the degree to which an agent is resolute in $\phi$-ing seems to depend not on how strong the temptation to $\phi$ is, but on how much the agent cares about

The chief problem with SODA is that we can apply the general gist of the Temptation Transmission argument to Liberman's own positive account. Why should we expect our second-order desires to persist through temptation, when our first-order desires so frequently fail in the face of temptation? Temptation paradigmatically comes in the form of desire, swamping other relevant desires. So how then is desire supposed to play the role of resistor to (tempting) desire?

In order to see how this works, consider an example Liberman uses in defense of SODA: the resolution not to eat donuts. In this case, we have three pertinent desires:

> *Desire 1*: The desire *not* to eat donuts, which corresponds to an intention not to eat donuts (*Intention 1*)—first-order desire.[13]
>
> *Desire 2*: The desire to persist in the intention not to eat donuts—second-order desire.
>
> *Desire 3*: The desire to eat donuts—tempting desire.

Imagine that I form an intention not to have a donut on the grounds of Desire 1, and furthermore I desire to persist in this resolution. But then I enter the break room and a colleague has brought in fresh donuts, and I am tempted to eat a donut after all. In the face of this tempting desire, imagine that Desire 1 drops away. I no longer desire not to eat a donut; in fact, I desire the opposite. Why should we expect Desire 2 to persist after Desire 1 has disappeared? Liberman's answer to this is as follows:

> The second-order desire can persist when the first-order desire does not because the second-order desire is held for additional reasons. I might desire to avoid eating a donut because I do not want to ruin my supper, or because I do not want to get powdered sugar on my shirt, or because I want to heed my doctor's advice to consume less sugar, etc. I desire to remain firm in my intention to avoid eating donuts for another reason:

> whether she $\phi$'s.... Suppose I resolve not to drink any beer at a party tonight, and I care very much about whether I keep this resolution. In such a case, it seems that even extremely tempting beer ... will not be very likely to sway me to break my resolution. (32)

[13] It may seem odd to speak of a first-order desire to not eat donuts, since normally we associate the temptation to eat donuts with desire, but we do not really associate the resolution to give up donuts with desire. However, Liberman herself speaks this way ("Reconsidering Resolutions," 22–23), and I think this language is natural in that we have a (placeholder) desire to $\phi$ whenever we intend to $\phi$. However, I do think that Liberman ought to draw a distinction between placeholder and substantive desires, since there are interesting and relevant differences between the two categories.

> because I care about carrying out my donut-avoidance plan and being an effective agent regarding the baked goods I consume . . . my desire to carry out my plan is a desire about what kind of agent I want to be; this sort of desire is resistant to temptations that press on the content of the plan itself.[14]

The agent is likely to lose Desire 1, the first-order desire not to eat donuts, in the face of temptation. Why not think that the tempting desire also puts pressure on the second-order desire, the desire to persist in the intention not to eat donuts? Liberman claims that the agent's second-order desire not to reconsider is supported by her reason to be an effective agent, someone who follows through on her plans and intentions.[15] Our interest in being an "effective agent" thus is an additional reason, separate from our first-order reasons, and it is this additional reason that is supposed to bolster second-order desires in the face of temptation.

However, the heart of Liberman's critique of Holton's views is that temptation applies just as much to our second-order intentions as it does to first-order intentions. Although she defends her own view from this objection by appealing to "being an effective agent" as an independent, additional reason that bolsters the second-order desire, I think this appeal is ultimately unsuccessful.

First, although being an effective agent, the sort of person who follows through on his commitments, is a worthwhile aim, being an effective agent does not require us to follow through on every single resolution. Take the example of resolutions regarding difficult athletic pursuits that require a demanding training regimen. One reason people undertake such pursuits is to prove to themselves their own capability, and in this sense, being an effective agent is among their motives. However, too much rigidity in following one's training plan is a detriment to meeting one's goal. Obsessive adherence to one's training plan is likely to lead to injury or burnout. Rather, what is needed—and is arguably harder to achieve—is flexible consistency. The point holds in general as well: to be a generally effective agent, yes, you must be willing to stick to your resolutions. But perfect adherence to one's resolutions is not effective agency. It is obsessiveness that is likely to backfire.

[14] Liberman, "Reconsidering Resolutions," 22–23.

[15] Although Liberman does not highlight this aspect of being an effective agent, it seems to me to share similarities with accounts that emphasize the temporally extended nature of practical rationality. For instance, Thomas Nagel writes that one sees "oneself as a temporally extended being for whom the future is no less real than the present" (*The Possibility of Altruism*, 69). As Michael Bratman points out, simply recognizing that one is temporally extended demands some concern for one's future ("Toxin, Temptation, and the Stability of Intention," 85–86).

Furthermore, if our response to temptation is to count the reasons in favor of not reconsidering versus the reasons in favor of reconsidering, it is not clear that the plentiful reasons on the side of not reconsidering will be decisive in favor of staying resolute. Given that temptation paradigmatically involves the presentation of reasons, it may not matter that we have independent, additional reason for the second-order desire (or the second-order intention, on Holton's account). We cannot weigh our reasons in a neutral deliberative space, because temptation makes certain reasons more salient and thus affects our ability to weigh reasons objectively. In the moment of temptation, our interest in being an effective agent may not count for much.

So whether we understand them as two-tier intentions or an intention-desire combo, resolutions are supposed to entrench or freeze our reasons for our original intention by forestalling reconsideration of that intention. The problem with this strategy is that temptation itself involves the presentation of reasons, and those reasons affect our first- and second-order intentions and first- and second-order desires. When tempted, we will feel that we lack good reason to act as we initially intended to act, and this is a key part of why temptation corrupts our rational agency so easily: it involves the appearance of reasons. Part of being rational is responding to a landscape of changing reasons and updating one's intentions accordingly. This means that, from the first-person perspective, resolutions are not going to be consistently effective, since when tempted, we are faced with reasons to do as we are tempted *and* reason to give up the second-order desire or intention not to reconsider.

## 2. THE NATURE OF TEMPTATION

In short, one reason both Holton's and Liberman's accounts go awry is because they fail to attend carefully to temptation in its own right. Thus, in order to give an account of how to rationally resist temptation, we must have in hand a clear and internally consistent account of the "enemy." Although both Holton and Liberman do briefly define temptation—gesturing in their accounts toward Scanlon's view of desire as a state involving the appearance of reasons—they also take on the additional and conflicting idea of desire being an "urge" or "pull."[16] A charitable interpretation of this move is that they recognize the limitations of Scanlon's account when it comes to motivational pressure, i.e., the felt sense of it being difficult to resist acting on a tempting desire, but I think appealing to "urge" language is not the right corrective to this gap in Scanlon's

[16] Liberman for her part adds the idea that desire comes in degrees of "strength" in discussion of SODA ("Reconsidering Resolutions," 16, 20). For Scanlon's views on desire, see *What We Owe to Each Other* and "Reasons and Passions."

account. I will start by presenting Holton's view of temptation, and then move to defending my own account of temptation and how tempting desires affect our deliberation.

Holton's view begins with temptation as it relates to judgment shift, since the two are inseparable on his account. He explains his view in several different passages, writing,

> I argue that temptation frequently works not simply by overcoming one's better judgment, but by corrupting one's judgment. It involves what I call judgment shift. . . . This in turn gives rise to the problem of understanding how one can resist [temptation]: the impetus to resist cannot come from the judgment that resistance is best.[17]

> The change in valuation [judgment shift] is not the *origin* of the process that leads to the subjects yielding to temptation: it is rather itself caused by [their] awareness that they are likely to yield. . . . If the change in valuation is not the source of the process that leads to yielding, what is? What causes the subjects to yield is desire, in one sense of that rather broad term.[18]

Then, discussing a particular case of temptation studied in experiments done by psychologists Karinol and Miller, Holton applies his view as follows:

> So, to sum up, what I think is happening in [the experiments] is this: the tempted children *find their attention focused* on the immediately available sweet; as a result they *find themselves with a strong urge* to ring the bell to get it; and, as they become aware that *they are likely to succumb to this urge, they change the evaluation of their options* so as to avoid cognitive dissonance.[19]

Scanlon's influence is clearly present in this description: "the children find their attention focused on." But Holton departs substantially from Scanlon as well in his characterization of tempting desire as a pull or urge, and in the claim that we change our judgment in response to the recognition that we are going to succumb to the urge.[20]

[17] Holton, *Willing, Wanting, Waiting*, 97.

[18] Holton, *Willing, Wanting, Waiting*, 101–2.

[19] Holton, *Willing, Wanting, Waiting*, 102 (emphasis added).

[20] At one point Holton directly states, "What is missing in Scanlon's characterization is the idea that desire *pulls* me to a course of action: that I have an *urge*, or, in more extreme cases, a *craving*, something that moves me to do it" (*Willing, Wanting, Waiting*, 102).

In short, I think there are two problems with this broad view of desire. First, Holton's understanding of desire is at odds with the rest of what he says about temptation and resolution. Consider again his description of how temptation works in the study:

1. The children find their attention focused.
2. As a result of this focused attention, they experience a strong urge.
3. They realize that they are going to succumb to this urge.
4. They shift their judgment in favor of temptation in order to avoid the cognitive dissonance of not doing that which they judge to be best.

How is resolution supposed to be effective, on this account? As I understand Holton, resolutions work by holding off reconsideration, since the second-order intention involved in reconsideration is the intention not to reconsider the first-order intention. The problem with this is that there is no reconsideration in the summary above. It would naturally fall in step two or step three: perhaps we are prompted to reconsider after having our attention focused, or perhaps the urge of temptation just *is* the urge to reconsider, and so after experiencing the urge, we are likely to reconsider. If there is no moment of reconsidering whether or not to act as resolved, there is no moment at which to choose whether to act as resolved or as tempted, and we simply slide into judgment shift and succumb to temptation, or resist temptation simply because the tempting urge is too weak to be a real threat. This leads me to my second criticism.

Setting aside the consistency of Holton's view, the more significant reason for rejecting an "urge" conception of tempting desires is that this conception renders us passive in the face of our desires. This approach conceptualizes desire or inclination as a force that acts upon us, which removes our agency in the face of desires. In other words, a view of inclination as an urge makes our ability to resist that inclination entirely contingent upon the strength of the inclination.[21]

Just like a current is a force outside of me with which I struggle, so is an urge-desire something outside of me, something against which I struggle. But this is not in fact what inclinations are like. Inclinations are not forces outside of us that act upon us. They arise from within our agency, and when we struggle with an inclination, we are struggling with ourselves, not something foreign to us. Grappling with our own inclinations is not the same as wrestling with

[21] Going forward I will tend to use "inclination" rather than desire, since I follow Tamar Schapiro in finding "desire" often unhelpfully vague. By inclination, I mean a desire that pressures the agent to act, i.e., a desire upon which it is easier to act rather than not, and a desire that requires effort to resist.

another person or a strong wind. Complicated though the relationship may be, our inclinations are part of us.

Consider for instance being moved to act as a result of an inclination. If having an inclination is like being pushed and pulled by an external force, then whatever results from that inclination is not properly understood as our action but is instead mere effect or behavior. About this, Tamar Schapiro writes, "If my desire pushes me around like an ocean tide, then it is hard to see how its effects can, in principle, count as my actions, unless action is just a way of being pushed around."[22] As Schapiro points out elsewhere, if inclination is a brute force like a tide or wind, it is not clear how we could ever act *on* an inclination.[23] I can act in light of the tide or the wind, but I cannot act on them in the same way I can act on a strong desire to scream or eat cake or start dancing.

Furthermore, construing inclinations as brute forces does not just create problems for acting *on* inclinations, it also creates a problem when it comes to resisting tempting desires. Take a current in water: my ability to swim against a current is ultimately not up to me. It is up to me whether I have learned how to swim, or whether I try to resist. But there are some currents so strong that even very skilled swimmers cannot resist despite their best efforts. Similarly, conceiving of inclinations as forces that act upon us makes inclinations like currents: some of them will be perfectly manageable forces we can resist. But others will simply be too strong, and we will be helpless in the face of them. This means that viewing inclinations as urges renders us unfree in the face of inclination, and whether we are able to resist the inclination is not up to us but is instead contingent upon the force of the inclination.[24]

It is not an accident, in other words, that Holton's view of judgment shift holds that temptation causes judgment shift because we *predict* that we will succumb to temptation. In other words, on his account of judgment shift, I view myself from the outside and realize that the tempting inclination is too strong to overcome, and so I predict that I will succumb and change my judgment about what is best to do to be in keeping with my prediction. But this is an odd and problematic account of how tempting desire affects us. Relating to our inclinations in this way involves abdicating responsibility for ourselves as agents, viewing ourselves from a third-personal perspective instead

[22] Schapiro, "What Are Theories of Desire Theories of?," 4.

[23] Schapiro, *Feeling Like It*, 49.

[24] I intend my remarks in this section to be neutral with respect to views on free will. Although I suppose some determinists might argue that the correct way to conceive of our agency is to view ourselves as predicting what we will do, rather than deciding what to do, I take it that this is a minority position and would generally be regarded as a *reductio* of the view in question.

of occupying our agency from its own perspective, the first-personal.[25] Take the following scenario:

> Vinny resolves to spend his Saturday catching up on a complex project for work. He then learns that several of his friends are planning to drive into the country and visit Vinny's favorite vineyard on Saturday, and that he's invited to join them. Vinny loves to get out of town on the weekends, and he furthermore enjoys the food and wine at this particular vineyard. When he reflects on the tempting desire to skip his work and go instead to the vineyard, he realizes that he's probably going to succumb to the temptation. Given his prediction that he will succumb, Vinny just decides in advance to give up on his resolution and tells his friends he will join them on Saturday.

The problem with this scenario is that Vinny replaces the first-personal perspective of the deciding mind with the third personal. But even this is not quite strong enough: it is not just that his instinctive mind has made a *prima facie* decision to go to the vineyard and his deciding mind goes along with it. Rather, Vinny decides what to do *on the basis* of his prediction about what he will do, which is to say that he does not properly decide. He does not settle the practical question of "What should I do?" but rather substitutes it with a theoretical question, "What will I do?" He takes the perspective of an observer, not the perspective of an agent responsible for his own action.

Still, one might object that the problem here is not with the conception of inclination as something that acts on us, necessitating a third-person perspective, predicting stances toward ourselves, but rather that the problem lies in Vinny's failure to distinguish between predicting and deciding. In other words, the problem is that he confuses the two activities. This objection holds that sometimes inclinations really *are* too strong to be resisted. In such cases, we ought to recognize that we are unlikely to be able to resist the inclination, although we should not treat this prediction as good reason for then deciding that the best thing to do is to act on the tempting desire, as Vinny does. On this line of thought, perhaps Vinny should keep trying to maintain his resolution and just wait and see what happens on Saturday: maybe he will be successful in working, maybe he will not. Or perhaps Vinny should take his prediction that he will succumb as a sign that his initial resolution was poorly formed and unrealistic and revise on the grounds that his initial resolution was ill thought.

I find both practical recommendations dissatisfying, since both continue to treat Vinny as a bystander to his actions. However, the objection helpfully

[25] My way of framing this issue is drawn in part from Marušić, *Evidence and Agency*, 122–36.

highlights the fact that when we predict that there is a good chance we will succumb to a tempting desire, we typically either alter our plan for executing the intention or resolution in question and/or go on to make a decision that will alter the context in which we decide, in order to avoid succumbing to the tempting desire after all. Take Vinny: imagine instead that Vinny decides to give up his Friday night leisure time and gets his work done then, in order to free up his Saturday for a trip to the vineyard. This would be an instance of altering his plan for executing the resolution. Or perhaps Vinny instead makes plans with his spouse to go out for dinner at a local restaurant on Saturday, knowing that he could not make it back from the vineyard in time for dinner but that he will nonetheless have plenty of time for working. If he takes this option, he alters the context or ecology in which he will deliberate and act on Saturday afternoon. If he makes plans with his spouse, he may still feel tempted to drive to the vineyard on Saturday. But having made plans, he will be able to resist his temptation because this new situation or context will make the vineyard less tempting.

In short, I think we should reject the idea that inclination is a force that acts upon us. For one, construing inclination as a force makes inclination out to be something external to our agency, and inclinations are part of our agency. Second, I do not think that our ability to resist a given inclination is entirely contingent on the strength of the inclination being sufficiently weak.

In contrast, on my view inclinations are moves in deliberation as opposed to brute forces inasmuch as they involve the appearance of reasons, but they nonetheless do pressure our deciding mind because they purport to settle our action. So why is it difficult to resist temptation? Not because temptation is a force that acts on us from without, but because when we are tempted our own agency is in tension, part of it directing us to act as it wills and the other part asking, "But should I really $\phi$?" or even more open-endedly, "What should I do?" But this means that tempting desire is not arational. Part of the difficulty of resisting tempting desire is the difficulty of resisting the reasons latent within it. Furthermore, strong tempting desires can make certain reasons very salient, so that it seems as if we have very strong or decisive reason to do something, even though that may not in fact be true.

When we are specifically tempted to give up on a resolution, we may lose our grasp on the reasons we had for forming that intention, but that does not mean that we lose those reasons altogether. Rather, it is as if we "forget" or "lose sight."[26] For instance, if I resolve to exercise more but then am tempted to stay on the couch when it comes time to go to the gym, it is not that in being

[26] Thus, succumbing to temptation often seems subjectively rational in the moment but later occasions regret.

tempted my reasons for exercising are no longer relevant to my situation. No, the reasons that led me to resolve to exercise more in the first place are still relevant considerations; I have just lost sight of them because the temptation made other reasons salient.[27] When tempted, it is not just that my instinctive mind is figuratively yelling imperatives at my deciding mind and the imperatives have no sticking power. Rather, inclinations appear as imperatives for which we have good reason. In the gym case, my inclination to stay on the couch will include my being drawn to the comfy couch, the annoying long drive to the gym, and the physically strenuous and unpleasant workout waiting for me upon arrival at the gym.

But this leads us back to the place at which the discussion began: if temptation makes certain reasons very salient, then the saliency of those reasons affects our ability to maintain our resolutions in the face of temptation. Forming a resolution is not an automatic out from being affected by temptation. Tempting desires are throwing reasons for consideration into the ring, and they may appear to be excellent reasons even in circumstances where we have formed a resolution to the contrary. After all, as I said above, the deciding mind does not tally our reasons from a neutral or dispassionate perspective. We can try to ignore our inclinations when deciding but tempting reasons will affect our overall tally of what we have best reason to do unless we specifically intervene to wholly discount tempting reasons in our deliberation.

Imagine for instance that Bri has resolved to stay home and eat simple homemade meals over the weekend in order to save money. But then on Friday afternoon she gets a text from a friend inviting her to join a group of people at her favorite restaurant for dinner. It would be natural for Bri to feel tempted to join them: she would have the company of friends, her favorite food, and no work preparing for or cleaning up after dinner. Although it might be rational for Bri to refuse to reconsider her dinner plans given her initial reasons for resolving to stay in and any independent, additional reason she has for refusing to reconsider, it is not clear why these reasons will be compelling in the face of temptation. After all, when her deciding mind is weighing what to do, all of

[27] Perhaps surprisingly, this account of how temptation affects us is compatible with how Holton understands the effort involved in resisting temptation. He writes, "One maintains one's resolution by dint of effort in the face of the contrary desire" (*Willing, Wanting, Waiting*, 118), and then later adds that "the effort involved has to be a kind of mental effort. It is the mental effort of maintaining one's resolutions; that is, of refusing to revise them. And my suggestion here is that one achieves this primarily by refusing to reconsider one's resolutions" (121). In other words, according to Holton's own lights, it is difficult to resist temptation not because resisting temptation is like swimming against a current, but because we must set aside the reasons temptation makes salient and instead affirm our resolutions. In other words, it seems that Holton's account of tempting desire is confused.

the reasons to act as she is tempted will be part of the deliberative milieu. Furthermore, she cannot just pull out the effective agency card and say to herself, "Ah, compelling though these are, I must rule out and ignore all of the tempting reasons because I want to be an effective agent." You can be an effective agent without following through on every single one of your resolutions. So why not deliberatively choose to downgrade effective agency on this occasion and go for the pleasure of going out with friends instead?

Furthermore, it seems to me that this difficulty applies just as much to Liberman's SODA account as it does to Holton's two-tier resolutions. Although it is possible to want conflicting things at the same time, it would be unusual to experience a tempting desire *and* a desire not to reconsider what you have previously decided to do because you want to be an effective agent. Why is this unusual? For the same reasons listed immediately above. Temptation makes us think that we no longer have good reason to do as we previously intended, and in the absence of these reasons, it will be hard to maintain a desire not to reconsider that prior intention. Furthermore, although I do think it is possible for the instinctive mind to be conflicted and thus for us to have inclinations for two conflicting things at once, I think this is a case in which it is important to distinguish between desire in the mere sense of finding something good in a way that would make action intelligible, and desire in a more substantive sense, i.e., inclination or desire that makes it easier to act as we so desire. I suppose it is technically possible to be inclined to be an effective agent, but this is rather odd as an object of inclination. Consider how odd it is for instance to speak of "it being easier to be an effective agent rather than not."[28]

Granted, as Liberman argues, there is independent reason for the desire to not reconsider, and that reason is the aim of being an effective agent. In other words, Liberman claims that Bri's "first-order" reasons for desiring not to reconsider her first-order intentions may disappear in the face of temptation, but the desire to be an effective agent will persist and protect the first-order intention from reconsideration. This, however, seems tenuous. I do not see why the temptation's capacity to affect what we see as good reason extends only to first-order reasons and not also to second-order desires, including desires motivated by "additional" reasons like being an effective agent. We cannot appeal to this reason as a special consideration that is somehow immune from pressure by temptation.

Furthermore, as discussed above, being an effective agent does not require *never* changing your mind or giving in to temptation. Sometimes we make

[28] Arguably the problem is that "being an effective agent" is not an action you can undertake. It is the accomplishment of one's actions, and the way to accomplish one's actions is just to perform a particular action. You cannot generically perform the action of accomplishing one's actions.

foolish resolutions and sometimes the situation changes so significantly as to make our original resolution inapt. It is also perfectly acceptable for a normally resolute and effective agent to occasionally give up on a resolution just because. Our values and goals regarding the kind of agency we wish to have extend beyond mere efficacy. This is a point Sarah Paul makes, writing:

> Most of us do not care about *perfect* self-governance, even as an ideal. We also care about things like existential spontaneity, losing control, rolling the dice and letting the world decide, and other more Romantic ideals. For an agent with these multifaceted values, a life that is perfectly self-governed would not in fact be successful relative to her varied concerns.[29]

Given this, we should not rely too heavily on the idea of effective agency as a solution for resisting temptation. Sure, sometimes we resist tempting desires in order to persist in our goals and be effective as agents, but in other cases we prefer the ideal of being a flexible, spontaneous, or even rebellious agent.

Instead, I think that the power of Liberman's positive account derives from her emphasis on desire. When we want to do what we have resolved to do, it is much harder for temptation to get a foothold in our consciousness.[30] When a desire to do as we have resolved is making our reasons to act on the resolution very salient, tempting reasons will have a harder time crowding them out. For instance, compare someone who enjoys running resolving to push themselves to run their first marathon in contrast to a self-identified "couch potato" who resolves to run a marathon only on a dare from a friend. Although it is certainly possible for the second person to successfully finish a marathon, it seems more likely that the first person will complete the race, and furthermore, it is likely that the first person will have an easier time with their training. This person identifies as someone who enjoys physical exertion, and they view their training as something they want to do. The second person will relate to running the marathon as something they "have" to do. It is not an accident, in other words, that many people come to like and/or teach themselves to like something in order to fulfill a resolution to do that thing.

[29] Paul, "Diachronic Incontinence Is a Problem in Moral Philosophy," 345.

[30] This idea is well-supported by empirical research on temptation that shows that having "want-to goals," goals that reflect our "genuine interest and values and are personally important and meaningful," helps us focus on our goals and not get distracted by distracting and tempting alternatives (Milyavskaya et al., "Saying 'No' to Temptation," 679). See also Deci and Ryan, "Facilitating Optimal Motivation and Psychological Well-Being across Life's Domains," 14–23; and Werner and Milyavsksaya, "Motivation and Self-Regulation," 1–14.

However, although I think desire is a powerful tool for achieving our goals and resolutions, it does not thereby follow that making nonreconsideration the object of our desire is the most effective and rational way to achieve our goals. Rather, note that, in the example above, the object of desire was the goal itself: wanting to run the marathon. This is importantly different from desiring to maintain a resolution or desiring to not reconsider. Thus, if desire will help us resist temptation, we will find the most support from desires to do what we intend to do, i.e., first-order desires, and not in the second-order desire to resist temptation to which Liberman appeals.

In short, it seems that we finally have the answer to the question I posed at the outset of the paper: resolutions understood as a special two-tier intention are not a reliably effective sources of resistance to temptation. Neither Holton or Lieberman is committed to the claim that resolutions are effective in every instance, but one way to describe their project is the attempt to give an account of why we should expect resolutions to be effective at resisting temptation in a rational agent. But given the nature of temptation, I think we must abandon this aim. Temptation is not a force outside of rationality such that if we put up appropriate rational bulwarks, we will be free from temptation's pressure. Temptation occurs *within* our rational nature. Temptation is an inner conflict, not a conflict in which one party to the conflict assails us from without.

However, it is important to clarify at this juncture that resolutions nonetheless play an important conceptual role when it comes to temptation. We should not discard the idea of resolutions altogether, because resolutions mark the difference between intentions that are open to easy revision and those for which there ought to be a high bar for revision. Furthermore, this conceptual difference does make a difference in what is most rational for us to do. If on vacation I intend to spend my afternoon watching TV but then change my mind and go for a mystery novel instead, there is no sense in which this is a poor decision or a moment of weakness of will. I just changed my mind about what I wanted to do. If on the other hand I give up on my resolution to spend my afternoon working in favor of reading a mystery novel, I have probably failed to act as I ought.

This might seem inconsistent with my claim that resolutions are not of special use in resisting temptation: How could it be that resolutions mark what is (normally, in the absence of other special reasons or notable changes in circumstance) most rational for us to do but are not of any special use in resisting temptation? Are we really that insensitive to what we have most or best reason to do? In some cases, yes. This just is the problem of weakness of will or *akrasia*. If we always did what is most rational for us to do, there would be no need for this paper.

However, having said that, I want to immediately walk the claim back to some extent. I am not saying that resolutions are of no use, that our understanding of what our reasons are is useless in the face of the force of tempting desire. I am just denying the strong claim that simply forming a resolution understood as either a two-tier intention or two-tier intention and desire pair is sufficient for resisting temptation. I will take for granted that very often the presence of a resolution means that we *ought* to resist temptation. But simply having a second-order intention or desire present is not sufficient as a *strategy* for rationally resisting temptation.

Insofar as resolutions help make us aware of the excellent reason we have to $\phi$, they probably help rational agents resist temptation. But this is not a function that is unique to resolutions. This is just a point about what it is to be a rational agent who is responsive to reasons that bear on practical questions. And certainly this claim does not show that forming a resolution is some special strategy that will take away the power of tempting desire altogether.

Where does this leave us, then? In order to give an account of how to rationally resist temptation, we need to broaden our perspective and move our focus beyond resolution and the discrete desire to be an effective agent to a practical virtue that encompasses our inclinations themselves.

## 3. MANAGING TEMPTING DESIRES

Above I have tried to emphasize that tempting inclinations are not just brute urges that determine our actions from without. On the other hand, our inclinations are not rational if by "rational" we mean immediately responsive to what we judge to be our reasons. Simply judging that one has best reason to act as they have resolved does not mean that a desire or inclination to so act will follow. In light of this, perhaps we should just give up on the idea of resolutions and the project of resisting temptation in a uniquely rational way altogether, and instead focus on distancing ourselves from our inclinations in order to manage them.

In the penultimate chapter of his book, Holton introduces the idea that rationally resisting temptation requires a general policy of nonreconsideration, a policy of not reopening the deliberative question when temptation threatens a resolution.[31] What exactly Holton means by this is sometimes difficult to trace, but one natural thought is that reconsideration is not a strategy of rationally resisting temptation but instead a strategy of mere management: since we cannot get our inclinations to respond to our judgment about what we have best reason to do, and since furthermore our tempting desires can affect our judgment about what we have best reason to do, perhaps we should focus on simply ignoring tempting desires altogether.

[31] Holton, *Willing, Wanting, Waiting*, 140.

However, Holton denies that nonreconsideration is to straightforwardly ignore one's tempting desires. He writes:

> In saying that agents do not reconsider, I do not mean that they do not think about the issue at all; as we have seen, some thought will typically be necessary for effective monitoring. Nonreconsideration only requires that they do not seriously reopen the issue of what to do, and seriously arrive at a new judgment.… That judgment [that it would be best, all things considered, to abandon the resolution] involves not just an evaluative judgment, but a comparison: a *ranking* of one option as better than the others.… [Such a ranking] is not the kind of thing that simply arrives unbidden.[32]

In other words, for Holton, nonreconsideration is supposed to protect the rationality of resisting temptation: it is not a concept introduced to solve the problem of efficacy but is rather introduced in order to preserve the rationality of acting on one's resolutions in the face of temptation. If we were to reconsider in the face of temptation, Holton reasons, then we would end up making the judgment that it would be best for us to act on our tempting desire and thus we would have a sort of reverse *akrasia* problem in which it would be irrational for us to act as we initially resolved.

More practically speaking, in terms of what it actually looks like to adopt a policy of nonreconsideration, in the above quote Holton suggests that reconsideration does involve thinking about one's resolution, just not "seriously reopening the issue of what to do." There is something compelling about this reply, since there is an important difference between merely thinking about an alternative course of action as opposed to reopening a deliberative question and actively ranking one's options. However, this reply also requires a delicate balancing act, and it is not clear that this balance is possible in practice. About this, Paul writes:

> [Holton] denies that what he is recommending is weathering temptation by making oneself irrational, or even arational; we are meant to be able to see ourselves as in rational control of our actions when implementing a prior resolution. At the same time, his proposal requires ignoring one's own evaluative ranking at the time of action and refusing to reconsider a resolution one knows it would be rational to revise if one did [reconsider in view of the tempting reasons]. Holton therefore needs a cognitive state to exist in which the agent takes her present action to be up to her, maintains awareness of her resolution and the considerations

[32] Holton, *Willing, Wanting, Waiting*, 150.

> supporting it, undergoes a shift in evaluative judgment in the light of which those considerations appear comparatively weak, and yet sees no open practical question. This strikes me as a very difficult state of mind to consciously maintain, bordering on bad faith.[33]

Furthermore, I think we can add to these concerns. It is natural to interpret the idea of nonreconsideration as just being the policy of refusing to take tempting reasons into consideration, steadfastly ignoring them, refusing to engage with or think about them. This is after all a strategy people take with respect to temptation: say that I have resolved to forgo all fun purchases for the remainder of the month. Essential purchases only. I might reasonably refuse to engage with any tempting thoughts in the course of carrying out this resolution: delete or block all emails about sales, immediately dismiss proposals from friends to go out for the night, etc.

However, in keeping with Paul's remarks, this is unhelpful as a singular long-term strategy, since it results in a kind of alienation from one's inclinations that I think cannot be sustained for long. Furthermore, this strategy is open to the objection that in so acting the agent is irrational, closing herself off to a set of perfectly good reasons to revise her earlier intention. And these points are ultimately the problem with adopting distance-and-manage-by-ignoring as our overall strategy with respect to tempting desires. Refusing to consider tempting reasons may be a way of managing one's tempting inclinations, but to adopt this position is to treat one's inclinations as something that happens to you that you must work around. It may be necessary or appropriate to do this in the short run or in certain extenuating circumstances, but it is not the overall outlook we should adopt toward tempting inclinations because it is not sustainable or appropriate to perpetually live in a state of divided agency. In other words, rather than skirt or bypass our inclining nature altogether in order to forestall inclination from interfering with resolution, the better option is to recognize that inclinations are among the attitudes that constitute our agency, and as such our stance toward them should not be denial but rather cultivation.

## 4. Resolve as Practical Virtue

An effective response to temptation must address temptation itself. One often hears that it is better to address to source or root of a problem, rather than simply try to manage or mitigate its effects, and the adage holds in the case of temptation as well as in home repairs. Rather than respond to temptation by counting up reasons, a process that will be prone to distortion by highly salient

[33] Paul, review of *Willing, Wanting, Waiting*, 890–91.

tempting reasons, or permanently adopt an alienated stance toward one's tempting desires, we ought instead to strive to resist temptation by attending to what and how we generally desire.

In other words, if we are to rationally resist temptation and uphold our resolutions, we must focus on mitigating or even preventing temptation altogether, not resisting it. This however requires a practical virtue, since a practical virtue shapes the nature of agency, forming habits of deliberating and desiring.[34] I suggest that we call the practical virtue relevant to resisting temptation "resolve."[35] I will first sketch out what I imagine the resolute person would look like before transitioning to address the question of how we cultivate resolve.

The resolute person is good at taking the long view, good at remembering why they resolved to do the thing in the first place, and good at anticipating how they would feel in the future if they abandoned their resolution. Furthermore, such a person will have dispositions to desire that which they have resolved to do and minimize the effects of tempting desire. Of course, there are no absolutes here. Sometimes it is normal or even good to be tempted, and so I am not claiming that a resolute person will never experience conflict over a decision or desire something opposed to their resolution. Rather, the resolute person is the kind of person whose inclinations are generally in keeping with her judgments about what is worth desiring and doing, and furthermore is good at delaying gratification, not getting easily distracted by desires for immediately available pleasant things, the pursuing of which will prevent her from following through on her other resolutions.

Return to the example of Bri and consider what this might look like in practice. Above, I claimed that when Bri is tempted to go out to eat with her friends, her temptation will make her reasons for going out very salient. One

[34] By practical virtue, I mean an excellence of agency. I take no position here on how the idea of a practical virtue maps onto Aristotle's distinction between moral and intellectual virtue.

[35] One might wonder if resolve is simply another name for continence. I do think these is a strong connection between resolve and continence, but I think resolve involves more than mere continence, since continence involves merely controlling how one acts in the face of one's desires as opposed to shaping them. I also think there may be a connection between resolve and the concept of "trait self-control" in social psychology, which is self-control that involves automatic behaviors as well as conscious and effortful exercise of control. About this, psychologists Gillebaart and de Ridder write, "These findings give credit to the idea that people high in trait self-control make the desired choice in an automatized, effortless manner, suggesting that trait self-control does not so much involve effortful resistance of immediate urges on single occasions, but rather involves the ability of not being tempted or distracted by such urges at all" ("Effortless Self-Control," 90). For more on the idea of self-control as a trait or disposition, see Tangney, Baumeister, and Boone, "High Self-Control Predicts Good Adjustment, Less Pathology, Better Grades, and Interpersonal Success," 271–322; and de Ridder et al., "Taking Stock of Self-Control," 76–99.

response to this temptation would be to tally her reasons for and against going out. The problem with this response, as previously delineated, is that a strong temptation to go out will make all of her tempting reasons very salient, and the initial resolution reasons will pale in comparison, and so her reasons on balance may favor going out to eat, even though this is opposed to her resolution. In contrast, on the view I am advancing, Bri's capacity to resist tempting desires will depend on her broader tendencies regarding desire and deliberation. Is she easily distracted by the inclination to go out with friends and does the tempting desire swamp out all other relevant desires? Or does she remain mindful of her desire to be more financially disciplined, to increase her savings? When deliberating, does she account for the fact that her resolution was formed in order to resist temptations exactly like these, thereby downplaying the apparently good tempting reasons and refusing to actively reconsider her resolution?[36] Or does she take the tempting reasons as new pieces of information that call for full-scale reconsideration of her resolution?

Some of this description might seem relatively obvious, but notice how I am *not* describing resolve: resolve is not forming a resolution that will form a "wall" around one's future deliberation and prevent temptation from taking hold. Neither is resolve strong willpower, the ability to punch down or overcome any temptation that comes one's way. This way of conceiving resolve sees resolve as a virtue exclusive to the rational deciding mind, a strong capacity to resist the inclinations. On the contrary, I think resolve is a practical virtue of the whole agent, which means that it encompasses both our capacity for inclination and our capacity for rational decision making. Resolve shapes our desire as much as our deliberation.

Reading these descriptions, it may begin to seem as though the person of resolve just does not experience temptation. This is true in some sense. It is not true insofar as a resolute person should be able to feel or recognize the force of conflicting considerations. Bri, for instance, might remain fully resolute, fully committed to her budget, and yet acknowledge the presence of conflicting reasons and even "feel" their force, by which I mean seeing them as compelling reasons on which she could act, as opposed to considerations that she merely

[36] However, it is clearly irrational to *never* reconsider one's resolutions or intentions. So any such policy will have to take this into account. Holton proposes the following as guidelines: "It is rational to have a tendency not to reconsider a resolution: if one is faced with the very temptations that the resolution was designed to overcome; if one's judgment will be worse than it was when the resolution was formed. It is rational to have a tendency to reconsider a resolution: if the reasons for forming the resolution no longer obtain; if circumstances turn out to be importantly different from those anticipated; if one made an important mistake in the reasoning that led to the resolution" (*Willing, Wanting, Waiting*, 160).

recognizes as potentially reason giving for another person although they have no draw on her, at least not in the current context. Furthermore, I think having the virtue of resolve is compatible with experiencing conflict (understood broadly) over a decision or commitment. However, it is true that the resolute person does not experience temptation in that the resolute will take pleasure in and thus be inclined to do that which she intends to do and has resolved to do.

Notice that the idea of resolve as a practical virtue encompasses key ideas Liberman appealed to in her account. The best understanding of Liberman's SODA highlights the centrality of desire *for* that which we have resolved to do, and the resolute person will either naturally desire the object of his resolution or actively work to cultivate desire for it. What about the desire to be an effective agent, which played a prominent role in avoiding reconsideration for Liberman? Will this desire be present in the resolute person? Yes, although I think this desire will not play an especially prominent role in the psychology of the resolute person. If the resolute person was constantly being resolute to prove his effectiveness as an agent, this would be a desire to be effective for its own sake, and this hardly seems like an excellence of agency. However, I do think a resolute person would desire to be an effective agent, and this desire might be especially important for those in the process of cultivating resolve.

Having seen an outline of what resolve looks like, arguably the more difficult question is how one becomes a person of resolve. The reason tempting desire poses a problem in acting as we have resolved to act is because it is recalcitrant to our judgments about what we have best reason to do in a given moment. How can I then insist that the rational way to relate to temptation is to cultivate one's inclination in the right way?

First of all, perhaps surprisingly, we can use management to cultivate our inclinations. This is surprising because managing or acting on is a stance of alienation: when I take this stance toward my inclinations, I seem to be regarding them from a distance rather than inhabiting them. However, when we manipulate our own attitudes with the result that the attitudes themselves are changed, the result is a change in the attitudes we inhabit and potentially a change in our own character. Take a specific example: say that I am trying to teach myself to love running, and so I decide to make running more fun by listening to my favorite music for dancing while running. I think this is best described as a case of manipulation or management because it is an attempt to shape one's inclinations, not by directly altering our inclinations, but by associating something I wish to be inclined to do with something I already take pleasure in and am inclined to do. But, if the result is that I develop the inclination to run, I think this strategy of management is a form of cultivation insofar as it results in me having cultivated a new inclination.

Still, the natural objection is that unless my new inclination involves some sense of the goodness of running—the inchoate or latent reason of inclination—it will not be an inclination proper but rather a brute urge. This then is where the role of attention comes in. If all we did to cultivate inclinations was bribe or trick ourselves into acting, cultivating inclinations would be no deeper than forming new associations. For instance, in the example above, I would simply form an association between something I was already inclined to do and something new, thus not properly acquiring a new inclination.[37] In cases of especially dull or odious action, this may be the best we can hope for.

If our ideal end result is instead a grasp of the goodness of running, then I think there are two primary ways in which I might come to find running good. One is of course through the practical cognition of judging something good. The problem is that in the running case, I probably do judge that running is good, and I am simply trying to bring my inclinations in alignment which what I judge to be good. What we are inclined to do is not directly up to us, and so there is no way to guarantee or force oneself to have an inclination, although I suspect that the right kind of appreciative attention can at least encourage us to take pleasure in something. Again, take running: if I am attentive to my running and actively thinking about the good embedded in the activity, I might begin bribing myself to run with the promise of my favorite music, but along the way, through attention to running and its goods, I learn to appreciate running for other reasons, for example the mental clarity, the fun of pushing oneself to run farther or faster, and the simple joy of movement. Appreciating these new reasons does not mean that I will have an inclination to run at every waking moment—one can after all appreciate reasons in a motivationally cold way—but it does make it possible for me to have the inclination to run at all.

Although I have for the most part set aside questions about the rationality of resolution in this paper, I want at this point to briefly address the rationality

[37] If the activity in question is one for which only certain motivations count as good motivations, then we can raise the further objection that merely associating an existing inclination with this new activity will not generate the right kind of motivation. I have in mind the chess-playing-child case raised by MacIntyre (*After Virtue*, 188). There he imagines bribing a seven-year-old child with candy to play chess, but points out that we can reasonably hope that over time the child will come to appreciate the goods internal to chess, and so desire to play chess for its own sake. This has the further important effect that the child will no longer be willing to cheat, since if the child can cheat successfully in order to get candy, he has every reason to do so as long as candy is his only reason for playing chess. But if and when he begins to play chess out of appreciation of the goods internal to chess, he will no longer be willing to cheat because to do so would be to violate the goods he appreciates in chess. The example is supposed to be a metaphor for the acquisition of virtue.

of resolve. Is it really rational for an agent to have habits that support resisting temptation and maintaining resolutions? One intuitive answer would claim that, objectively speaking, it is normally rational to act as we have resolved, perhaps because resolution is the product of the cool and deliberative deciding mind in contrast to hasty and often misguided inclination. However, although I think there are cases in which there is decisive objective reason for an agent to act as she has initially resolved, I do not think these reasons undergird the rationality of resolve in general. After all, in other cases agents may have good reason to act as they are tempted. Rather, the rationality of resolve stems chiefly from the authority of our decisions. That is not to say that we can never rationally revise our decisions. But as Paul writes:

> We may see our decisions as to some degree up to us, but we must also see the act of deciding as a matter of relinquishing our authority to change them whenever we like. For, otherwise, they would not be the kind of thing that can do the job of settling an open practical question.[38]

That is, the rationality of resolve does not stem from balancing of reasons to see whether resolution or temptation has better reason on its side but is rather located in our authority to settle practical questions.[39] As cross-temporal agents, we need to be able to make plans that settle current and sometimes future practical questions, and it is for this reason that we are often justified in refusing to reconsider our resolutions in the face of temptation.

But this answer simply pushes the question back. We can still ask why it is rational to regard our decisions as authoritative. One answer would hold that it is rational to view our intentions as blocking overeager reconsideration because practical rationality demands it: perhaps practical rationality requires that we take a long-range view about our preferences, or perhaps we will fail to be instrumentally rational in achieving our ends if we are constantly reopening the deliberative question. However, another answer maintains that granting past decisions special authority is not a strict requirement of practical rationality but is rather rational in the virtue-theoretic sense: it is a wise or good way to act and live in the world. This is the conclusion at which Paul arrives, writing:

[38] Paul, “Diachronic Incontinence Is a Problem in Moral Philosophy,” 349–50.

[39] By this, I do not mean that intentions have a reason-giving force of their own, but rather that intentions play a particular role when it comes to how we relate to our reasons. Namely, they settle our answer to a question and in so doing they block off reconsideration, or more accurately, once an intention has been formed, reconsideration must be justified and not be undertaken on just any whim.

> Diachronic continence [or resolve] is in many respects a virtue but not a rational requirement. That is, from the point of view of living well, treating one's intentions as having default stability and refusing to reconsider one's plans too frequently are highly recommended. The appropriate criticisms to make of someone who fails to be stable in this way will be that she is irresolute, flaky, wanton, always wondering where the better party is. But there are failures from the point of view of human excellence or virtue, not the philosophy of action.[40]

This is a substantive claim, and there may well be individuals or cultures who reject this vision of agency, but for my part I think Paul is right. Thus resolve is rational in two important senses. On one hand, it is uniquely rational as a method of resisting temptation. Efforts to become resolute might require nonrational (or arational) methods of resisting temptation, like refusing to buy a dozen donuts at the grocery store for fear that one will overeat them upon arriving home, but resolve consists in a set of dispositions of rational agency and is thus essentially rational in its function. However, resolve is furthermore rational as an excellence of agency. Being a resolute agent enables excellence in desiring, deliberating, acting, and, ultimately, living. We should aim to be the kind of agent whose decisions are generally resistant to tempting desires because the alternative is to be fickle, to flit from project to project, commitment to commitment, and follow through on few or none of them.

*Tyler Junior College*
*abigail.bruxvoort@tjc.edu*

## REFERENCES

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

Bratman, Michael. "Temptation and the Agent's Standpoint." *Inquiry* 57, no. 3 (April 2014): 293–310.

———. "Temptation Revisited." In *Structures of Agency*, 257–82. Oxford: Oxford University Press, 2007.

———. "Time, Rationality, and Self-Governance." *Philosophical Issues* 22 (2012): 73–88.

[40] Paul, "Diachronic Continence Is a Problem in Moral Philosophy," 354. For a contrary view on which diachronic continence is a requirement of practical rationality, see Bratman, "Time, Rationality, and Self-Governance," 73–88.

———. "Toxin, Temptation, and the Stability of Intention." In *Faces of Intention*, 58–90. Cambridge: Cambridge University Press, 1999.

Bryan, Gharad, Dean Karlan, and Scott Nelson. "Commitment Devices." *Annual Review of Economics* 2, no. 1 (June 2010): 671–98.

Deci, Edward L., and Richard M. Ryan. "Facilitating Optimal Motivation and Psychological Well-Being across Life's Domains." *Canadian Psychology* 49, no. 1 (2008): 14–23.

De Ridder, Denise T. D., Gerty Lensvelt-Mulders, Catrin Finkenauer, F. Marijn Stok, and Roy F. Baumeister. "Taking Stock of Self-Control: A Meta-Analysis of How Trait Self-Control Relates to a Wide Range of Behaviors." *Personality and Social Psychology Review* 16, no. 1 (February 2012): 76–99.

Gillebaart, Marleen, and Denise T. D. de Ridder. "Effortless Self-Control: A Novel Perspective on Response Conflict Strategies in Trait Self-Control." *Social and Personality Psychology Compass* 9, no. 2 (February 2015): 88–99.

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

Liberman, Alida. "Reconsidering Resolutions." *Journal of Ethics and Social Philosophy* 10, no. 2 (May 2016): 1–26.

MacIntyre, Alasdair. *After Virtue.* Notre Dame, IN: University of Notre Dame Press, 2007.

Marušić, Berislav. *Evidence and Agency.* Oxford: Oxford University Press, 2015.

Milyavskaya, Marina, Michael Inzlicht, Nora Hope, and Richard Koestner. "Saying 'No' to Temptation: *Want-to* Motivation Improves Self-Regulation by Reducing Temptation Rather Than by Increasing Self-Control." *Journal of Personality and Social Psychology* 109, no. 4 (October 2015): 677–93

Nagel, Thomas. *The Possibility of Altruism.* Oxford: Oxford University Press, 1971.

Paul, Sarah K. "Diachronic Incontinence Is a Problem in Moral Philosophy." *Inquiry* 57, no. 3 (2014): 335–55.

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

Scanlon, T. M. "Reasons and Passions*."* In *The Contours of Agency,* edited by Sarah Buss and Lee Overton, 165–83. Cambridge, MA: MIT Press, 2002.

———. *What We Owe to Each Other.* Cambridge, MA*:* Harvard University Press, 1998.

Schapiro, Tamar. *Feeling Like It*. Oxford: Oxford University Press, 2021.

———. "What Are Theories of Desire Theories of?" *Analytic Philosophy* 55, no. 2 (May 2014): 131–50.

Tangney, June P., Roy F. Baumeister, and Angie L. Boone. "High Self-Control Predicts Good Adjustment, Less Pathology, Better Grades, and

Interpersonal Success." *Journal of Personality* 72, no. 2 (October 2008), 271–322.

Werner, Kaitlyn M., and Marina Milyavsksaya. "Motivation and Self-Regulation: The Role of Want-to Motivation in the Processes Underlying Self-Regulation and Self-Control." *Social and Personality Psychology Compass* 13 (January 2019): 1–14.

*Journal of Ethics and Social Philosophy*
vol. 26, no. 2 · december 2023

*https://doi.org/10.26556/jesp.v26i2.1886*

# THE ESSENCE OF STRUCTURAL IRRATIONALITY

## THE IMPOSSIBILITY OF ATTITUDINAL SUCCESS

*Julian Fink*

It is commonly accepted that there is a wide range of combinations of attitudes that make a person irrational. These so-called structural cases of attitude-based irrationality include contradictory beliefs and intentions, failing to intend something you deem necessary for your intended ends (instrumental incoherence), failing to intend what you judge you ought to do (akratic incoherence), preferring *a* to *b*, *b* to *c*, but also *c* to *a* (cyclical preferences), etc.

Until recently, it was only tentatively assumed that these diverse patterns of attitudes belong to one and the same domain of irrationality. The tentative nature of this assumption has been challenged, however, by two major attempts to unify the domain of structural irrationality. Errol Lord and Benjamin Kiesewetter both propose that structural irrationality is a matter of combining attitudes that jointly guarantee the violation of a decisive normative reason.[1] Alex Worsnip, by contrast, suggests that structural irrationality is a matter of combining attitudes a person is constitutively disposed not to combine.[2]

Lord and Kiesewetter's proposal comes in two parts. First, they suggest that one is *substantially* irrational whenever one fails to respond correctly to the normative reasons that are epistemically available. For example, suppose you believe that the cat is on the mat, despite possessing sufficient evidence that this is not the case. Then, assuming that sufficient evidence translates to having decisive reason, you believe something you have decisive reason not to believe.[3] You therefore fail to respond correctly to the reasons available. This makes you substantially irrational.[4]

[1] Lord, "What You're Rationally Required to Do and What You Ought to Do" and *The Importance of Being Rational*; Kiesewetter, *The Normativity of Rationality*.

[2] Worsnip, "What Is (In)Coherence?" and *Fitting Things Together*.

[3] On the point of sufficient evidence translating to having decisive reason, see Kiesewetter, *The Normativity of Rationality*, 180–85.

[4] Cf. Kiesewetter, *The Normativity of Rationality*, ch. 7.

Second, *structural* irrationality is then understood as a subdomain of substantial irrationality. What creates this distinctive domain is the hypothesis that structurally irrational patterns of attitudes alone suffice to guarantee a failure to respond correctly to available reasons.[5]

By contrast, Worsnip offers a naturalistic alternative to the reasons-based account of structural irrationality.[6] Put roughly, structural irrationality resides in combinations of attitudes that a person is disposed not to sustain. More precisely, two or more attitudes are structurally irrational if and only if you are necessarily (and appropriately) disposed to abandon at least one of them once you become aware of holding them together.[7]

No doubt, both proposals have plenty of merit. They pick out aspects that are deeply symptomatic of many instances of structural irrationality. I am convinced that, generally speaking, structural irrationality tends to be unsustainable under awareness. This is particularly true of paradigmatic cases such as contradictory beliefs and intentions and means-end irrationality.

Likewise, there are paradigmatic instances of structural irrationality where you will necessarily violate a decisive reason. It is quite plausible, for example, that if you have sufficient evidence that you ought to *p*, then you will have decisive reason to intend to *p*. Also, if you lack sufficient evidence that you ought to *p*, then you have decisive reason not to believe that you ought to *p*. And so, since, necessarily, you either have or lack sufficient evidence that you ought to *p*, you indeed inevitably violate a decisive reason whenever you believe that you ought to do something without intending to do it.[8]

[5] Kiesewetter, *The Normativity of Rationality*, 236; Lord, *The Importance of Being Rational*, 27. For example, suppose you believe *p* and you believe not-*p*. This is structurally irrational because, necessarily, either (you have decisive reason not to believe *p*) or (you have decisive reason not to believe not-*p*). More generally: whenever you adopt a structurally irrational pattern of attitudes, you necessarily have at least one attitude that is substantially irrational (i.e., you have decisive reason not to have it). It is this necessity that distinguishes structural irrationality from other (and more general) forms of irrationality.

[6] Worsnip, "What Is (In)Coherence?" and *Fitting Things Together*.

[7] Worsnip argues that the disposition to abandon an attitude must be sensitive to a *constitutive* aspect of the attitude in question. This is what I mean by "appropriately disposed": "That is, human agents are disposed such that they are (at least normally) not able to (or at least find it difficult to) psychologically sustain such combinations of attitudes under conditions of full transparency" (Worsnip, "What Is (In)Coherence?" 188). For example, simultaneously intending *p* and intending not-*p* is structurally irrational because, necessarily, once you become aware of having these two intentions, you are appropriately disposed either to give up your intention to *p* or give up your intention to not-*p*. Accordingly, structural irrationality consists of combinations of attitudes that tend not to survive cognitive transparency.

[8] Cf. Kiesewetter, *The Normativity of Rationality*, 233 and sec. 9.5; Lord, *The Importance of Being Rational*, sec. 2.4.5.

Nevertheless, and despite these merits, I argue that both accounts fail to identify the *essence* of structural irrationality. Worsnip himself admits, for instance, that akratic incoherence (i.e., failing to intend what you believe you ought to do) poses a hard case for his account, since there is "widespread consensus that clear-eyed akrasia is possible."[9] Thus, the inability (or even the disposition not) to sustain structural instances of irrationality under awareness does not seem to be a strictly necessary condition for structural irrationality.[10]

Moreover, there are also cases where the inability to sustain a combination of attitudes under transparency does not seem to be sufficient for identifying structural irrationality. Suppose you have attitude *A*, yet you lack a belief that you have attitude *A*. For many types of attitudes, this clearly fails to be structurally irrational. If, for instance, you desire to go skiing yet you lack a belief that you desire to go skiing, you are not necessarily structurally irrational. However, once you become fully aware that you have attitude *A*, you are certainly disposed to believe that you have attitude *A*. Consequently, Worsnip's account would incorrectly qualify these combinations as structurally irrational.[11]

Analogously, looking at Lord and Kiesewetter's proposal, violating a decisive reason turns out to be neither necessary nor sufficient for structural irrationality. Suppose you intend to heal from a deep trauma and you are normatively permitted to intend so. Also, you intend not to smoke and, again, you are permitted to intend so. However, you also believe that you cannot heal from the trauma without smoking. Due to quirky circumstances, suppose you also have sufficient evidence for the truth of this belief and are thus permitted to believe so.

In these circumstances, you are structurally irrational: you fail to intend something you deem necessary for your intended ends. However, your attitudes do not violate a decisive reason. You are permitted (and thus *lack* decisive

[9] Worsnip, "What Is (In)Coherence?" 198; see also Worsnip, *Fitting Things Together*, sec. 5.4.3.

[10] Worsnip tries to counter this problem by arguing that the disposition not to sustain structurally irrational patterns of attitudes comes in degrees and that the weaker the disposition, the less irrational the pattern in question. So the fact that "clear-eyed akrasia is possible" just indicates that, in general, akrasia represents a weaker form of structural irrationality: "The most incoherent sets of mental states are ones whereby the disposition is so strong that it cannot be blocked; these sets of states will be impossible to sustain jointly under conditions of full transparency. But in less incoherent cases, such as akrasia, the disposition is weak enough to sometimes be blocked" (Worsnip, "What Is (In)Coherence?," 200; see also Worsnip, *Fitting Things Together*, sec. 5.4.3).

[11] For a more complete version of this criticism see Fink, "What (In)Coherence Is Not." See also Worsnip, *Fitting Things Together*, sec. 5.4.4.

reason not) to have each attitude. Consequently, the violation of a decisive reason does not qualify as a necessary condition for structural irrationality.[12]

These shortcomings reveal that we still lack an account of the essence of structural irrationality. This paper attempts to redeem this situation. I shall offer an original, reductive, and unified account of structural irrationality. The core of the account can be stated as follows: a set of attitudes is structurally irrational if and only if it is metaphysically impossible for those attitudes to be jointly successful. I will show that this account can fully explain the irrationality of some of the paradigmatic instances of structural irrationality.

While it is original, it is important to notice that my proposal also incorporates key aspects of the two accounts discussed above. Lord and Kiesewetter claim that a set of attitudes is structurally irrational if and only if it is impossible for those attitudes to be jointly substantially rational. This is similar to the view I offer in this paper, except that my account picks out structural irrationality not in terms of the impossibility of joint substantial rationality but rather in terms of the impossibility of joint attitudinal success. Also, my account (and its understanding of attitudinal success) incorporates a key element of Worsnip's account. I shall identify necessary success conditions for an attitude via that attitude's constitutive dispositions. That is, *s* is a necessary success condition for an attitude if and only if *A* constitutively aims at *s*.[13]

This paper is organized as follows. Sections 1–4 establish the core conditions of the account I wish to defend. I demonstrate how the idea that the attitudes in an irrational set of attitudes cannot possibly be jointly successful can unify an explanation of the irrationality of contradictory beliefs and intentions, cyclical preferences, and akratic and instrumental incoherence.[14] In section 5, I then add three subjective conditions to the account. Roughly, my proposal will be as follows: a set of attitudes is irrational if and only if it is transparent to the person who has them that it is impossible for them to be jointly successful.

[12] In addition, as I have argued elsewhere, the violation of a decisive reason does not qualify as a *sufficient* condition for structural irrationality (Fink, "Structural Irrationality Does Not Consist in Having Attitudes You Ought Not to Have"). Lord and Kiesewetter's reasons-violation view requires a number of assumptions that imply that you violate a decisive reason whenever you intend something you take to be normatively optional. However, intending to *p* and believing that it is neither the case that you ought to *p* nor that you ought to not-*p* does not qualify as a structurally irrational pattern of attitudes. The shortcomings of the two prevalent approaches make it necessary to seek an alternative account of the essence of structural irrationality.

[13] I thank an anonymous reviewer for highlighting these similarities.

[14] For a first and underdeveloped version of the account developed here, see Fink, "A Constitutive Account of 'Rationality Requires,'" sec. 8.

## 1. A Simple Account

This paper aims to establish the essence of structural irrationality. I seek to determine what it is to be structurally irrational and thus what unifies the specific domain of structurally irrational attitudes.

To begin, I will stipulate a simple view of what makes a set of attitudes structurally irrational. This view locates structural irrationality exclusively in the impossibility of the collective success of the attitudes in question. I will refer to this view as the Simple Impossible Success Account (SISA). I will explain what I mean by "success" in section 2, and I will turn to specific examples of irrational sets of attitudes in section 3.

Informally, SISA can be defined as follows:

> *SISA Informal*: A set of attitudes is irrational if and only if it is impossible for that set to be successful.[15]

Accordingly, irrationality consists in the impossibility of concurrent attitudinal success. Put differently, it is the necessity of attitudinal failure that makes that set of attitudes irrational. As we will see, SISA will turn out to be overly restrictive and inclusive at the same time. It both identifies as irrational sets of attitudes that are *not* irrational and fails to account for the irrationality of attitudes that *are* irrational. Nevertheless, it is worth gaining a detailed understanding of this account as it brings to light an essential aspect of structural irrationality.

In what follows, I will use "A set of attitudes is irrational" as shorthand for "Necessarily, if you adopt this set of attitudes, then you are not fully rational." By "a set of attitudes" I mean any possible set of a person's present and/or absent attitudes. By "present attitude" I mean a mental relationship between a person and a particular object. By "absent attitude" I mean the lack of a particular mental relationship between a person and a particular object. I will use "attitudes*" to refer to both present and absent attitudes.[16] Unless specified to the contrary, I will also assume that all attitudes* are contemporaneous.

In order to represent attitudes*, I will resort to the following schemas:

> <*A*: *x*> and <not-*A*: *x*>,

where *A* stands for the *type* of attitude (e.g., belief, intention, preference, fear, admiration, hope), *x* stands for the *object* of the attitude (e.g., a proposition,

[15] To save space, I use "set" here as shorthand for the members of that set taken together (not the abstract entity of the set itself). Thus, "it is impossible for the set to be successful" is meant to express that "it is impossible for the attitudes in that set to be jointly successful."

[16] When I speak of "attitudes," I will only refer to present attitudes, i.e., attitudes I assume one has. When I speak of "absent attitudes," I will refer to attitudes one does not have.

state of affairs, person), and "not" signifies the absence of an attitude. For reasons that will become apparent later, I will assume that "<$A$: $x$>" and "<not-$A$: $x$>" stand not for attitudes* themselves but for the *propositions* that represent those attitudes*.

Accordingly, "<belief: *the cat is on the mat*>" stands for the proposition "You believe that the cat is on the mat." Likewise, "<not-intention: *you go to school*>" stands for the proposition "It is not the case that you intend to go to school."

SISA Informal refers to the impossibility of the *success* of a set of attitudes. I will stipulate a general and a substantive view of attitudinal success in section 2. Here, I can say how I will model attitudinal success formally.

I will assume that every attitude comes with a particular set of success conditions. Success conditions relate to attitudes as follows. An attitude picked out by

<$A$: $x$>

is successful only if all success conditions obtain. Suppose, for example, that the propositions

$s_1, s_2 \ldots s_n$

denote the success conditions of the attitude picked out by

<$A$: $x$>.

Then, the success of the attitude represented by <$A$: $x$> requires the truth of

$s_1, s_2 \ldots s_n,$

which I shall refer to as "success propositions."

I need to add one further clarification. SISA Informal says that a set of attitudes* is irrational if and only if it is impossible for its members to be jointly successful. I interpret this as follows. First, by "impossible" I mean "metaphysically impossible." Second, I assume that it is metaphysically impossible for a set of attitudes to be jointly successful if and only if the success propositions of the attitudes in question cannot be true without a contradiction's being true.[17]

[17] By saying that the success propositions of the attitudes in question cannot be true without a contradiction's being true, I do not mean to say that all instances of metaphysical impossibility *formally* entail a contradiction. For example, (1) "Peter is a bachelor" can be true only if (2) Peter is unmarried. However, 1 does not formally entail 2. (It would do so, of course, if we were to add "If Peter is a bachelor, then Peter is unmarried" to 1 and 2.) This distinction between requiring the truth of a contradiction and formally entailing a contradiction will become important in section 7, when I add transparency conditions to the developed account of structural irrationality.

For example, suppose the success of your attitudes requires the truth of (1) "The numerical value $x$ is greater than $y$" and (2) "The numerical value $y$ is greater than $x$." There is no metaphysically possible world in which 1 and 2 are jointly true. In fact, the joint truth of 1 and 2 would require the truth of the flat contradiction that $x$ is greater than $y$ and it is not the case that $x$ is greater than $y$.

With these preliminaries in hand, I can now render SISA more precisely. Let $M$ be a set of attitude* propositions <$A$: $x$> and <not-$A$: $x$>. Then SISA purports to determine the irrationality of the attitudes* that $M$ represents as follows. First, assign all success propositions to every individual present attitude <$A$: $x$> that $M$ picks out. Second, form the complete set of success propositions for all attitudes <$A$: $x$> that $M$ picks out. Call this set "$S_M$." Third, examine whether it is metaphysically impossible for all propositions in $S_M$ to be jointly true. If it is metaphysically impossible, then $M$ represents an irrational set of attitudes. If it is not metaphysically impossible, then $M$ does not represent an irrational set of attitudes.

SISA can be formally stated as follows:

> *SISA Formal*: $M$ represents an irrational set of attitudes if and only if $S_M$ entails$_{\text{me}}$ a contradiction.

I use "entails$_{\text{me}}$" here as a technical term, meaning that it is metaphysically impossible for all propositions in $S_M$ to be true without a contradiction's being true. I will contrast "entails$_{\text{me}}$" with logical entailment or consequence later. This will become significant when it comes to introducing the kind of transparency required by my account of irrational attitudes.

I argue that SISA can explain and unify a core segment of structural irrationality. In particular, it manages to account for the irrationality of contradictory beliefs and intentions and, as I will show below, some forms of instrumental incoherence as well. Before I can demonstrate this in detail, however, I need to say more about attitudinal success. In particular, I need to state how we can correctly assign particular success propositions to particular attitudes.

## 2. ATTITUDINAL SUCCESS

Attitudes come with success conditions. This is a key assumption of my paper. I argue that understanding attitudinal success is essential to understanding structural irrationality. I operate with an essential or constitutive notion of success here. You may deem your intention to go to a bar a success if doing so results in your meeting the love of your life, but this is a non-constitutive kind of success. Intending to go to a bar and *not* meeting the love of your life (however regrettable) does not necessarily indicate that your intention was essentially defective.

I will distinguish between two types of success propositions. Both are equally relevant to establishing when a set of attitudes is irrational. One type of success proposition will pick out the success of an attitude *qua* its *correctness*. The other type picks out the success of an attitude *qua* its *executive performance*. I will explain below how these two types differ and how I arrive at them.

Before I justify my particular assignment of success propositions, I will first simply list a set of success propositions for the types of attitudes that are involved in the paradigmatic cases of structural irrationality, such as contradictory beliefs and intentions, instrumental and akratic incoherence, and cyclical preferences. Later, when explaining the irrationality of the paradigmatic cases of structural irrationality, I will rely exclusively on the success propositions in **bold**. As I will emphasize below, each listed success condition only states a necessary condition for the full success of the respective attitude. I am committed neither to conceiving of these conditions as sufficient nor to conceiving of them as contributory conditions for attitudinal success.

Let us start with belief. I will assume that a belief that *p* is fully successful only if

> ***p***.

An intention that *p* is fully successful only if

> it is not the case that you ought to not-*p*, and
>
> ***p***.

An all-things-considered ought judgment (that expresses a truth-apt cognitive attitude such as a belief) is successful only if

> **you ought to *p***, and
>
> *p* is possible, and
>
> **if you ought to *p*, then *p***.

By contrast, an all-things-considered ought judgment (that expresses a non-truth-apt, noncognitive attitude such as a desire or an intention) is successful only if

> *p* is possible, and
>
> ***p***.

If a preference is sensitive to a comparative judgment (i.e., having the preference depends on a judgment that *a* has more of a certain property, say *F,* than *b*), then the preference is successful only if

> ***a* is *F*er than *b*.**

If a preference subsists in what I shall call a "conditional intention" (i.e., roughly, one intends to [*a* and not-*b*] whenever one will either *a* or *b*), then the preference is successful only if the following material conditional holds true:

> **if (*a* or *b*), then (*a* and not-*b*).**[18]

What justifies this assignment of success propositions? I assume that many philosophers will share the intuition behind many of the success propositions specified above. That a false belief is not entirely successful is indeed uncontroversial. That an unrealized intention is not entirely successful is equally uncontroversial. However, some of the other conditions I have specified are in need of explanation. That the material conditional "If (*a* or *b*), then (*a* and not-*b*)" is a success proposition for a particular type of preference may not be immediately obvious.

I offer a method for identifying success propositions. This method is built upon the following principle:

> *Success*: Necessarily, *p* is a success proposition of <*A*: *x*> if and only if <*A*: *x*> constitutively aims at *p*.[19]

I follow Paul Katsafanas's definition of constitutive aims:

> Let *A* be a type of attitude or event. Let *G* be a goal. *A* constitutively aims at *G* iff (i) each token of *A* aims at *G*, and (ii) aiming at *G* is part of what constitutes an attitude or event as a token of *A*.[20]

In short, if an attitude of type *A* aims constitutively at *G*, then part of what makes that attitude a member of type *A* is that it is directed at *G*.

I suggest that *p* is a success proposition of <*A*: *x*> if and only if <*A*: *x*> constitutively aims or is directed at *p*. In general, there are two distinct ways in which an attitude can aim at something. First, an attitude can aim at *x* by instantiating a disposition to bring about *x*. This is the sense in which, for example,

[18] In order to limit the number of variables used in this paper, I use "*p*" (as well as "*a*" and "*b*") for a range of propositions *and* action types. I hope (and am confident) that the reader will be able to discern this dual use here. What is most important is that whenever "*p*," "*a*," and "*b*" are embedded in a complex syntactic construction (for example: "*a* is *F*er than *b*"), the resulting clause represents a proposition.

[19] I adopt Paul Katsafanas's suggestion that constitutive aims set up a fundamental and intrinsic standard of attitudinal success. That is: "If *X* [constitutively] aims at *G*, then *G* is a [fundamental or intrinsic] standard of success for *X*" (Katsafanas, *Agency and the Foundations of Ethics*, 39).

[20] Katsafanas, *Agency and the Foundations of Ethics*, 39.

an intention that *p* aims at its implementation, i.e., *p*. Second, an attitude can aim at *p* by having a propensity to be abandoned in the face of not-*p*. This is the sense in which, for example, a belief that *p* aims at truth, i.e., *p*. Of course, for these dispositions to ground a *constitutive* aim, they need to be essential to the attitude in question.

Here is a more precise characterization of these two dispositions. I assume that an attitude <*A*: *x*> aims constitutively at *p* if and only if either

1. <*A*: *x*> disposes* you to *p*, or
2. awareness of not-*p* disposes* you to <not-*A*: *x*>,

where "disposes*" signifies a disposition that is *essential* to having <*A*: *x*>.

The two dispositions differ in a directional sense. The first is an "attitude-to-world" disposition. Here the attitude constitutes a stimulus condition of the disposition. This type of disposition picks out the *executive aim* of the attitude. Failing to meet this aim implies an executive defect on the part of the attitude.

The second disposition is a "world-to-attitude" disposition. Here, the disposition manifests itself in response to (becoming aware of) how the world is. An aspect of the world (or awareness thereof) constitutes a stimulus condition of altering the mind. This type of disposition picks out a condition of *correctness*. Failing to meet this condition implies that the attitude is defective *qua* being incorrect.

Before I turn to justifying the success conditions specified above, let me add a crucial clarification. In order to make sense of these success conditions, it is essential to understand them as merely stating *necessary* conditions for an attitude's success. I only claim that an attitude is not entirely successful if one of its success conditions turns out to be false. However, I do not wish to make the additional claim that the truth of a success condition is sufficient for or necessarily contributes to the degree of an attitude's success.

Of course, this will likely be the case for certain success conditions. I would think, for example, that truth contributes to the success of a belief. But in other cases, this will not be so. For example, I identified the material conditional "If you ought to *p*, then *p*" as a success condition for a judgment that you ought to *p* (which I read as a material conditional). I understand this as saying that if that conditional turns out to be false (i.e., you ought to *p*, but you do not *p*), your ought judgment will not be (fully) successful. This strikes me as evident: you judge that you ought to *p*, you (in fact) ought to *p*, yet *p* is never realized. However, I do not mean to say that the truth of this conditional will necessarily contribute to the success of your judgment. This is obviously the case if the condition turns out true just in virtue of its antecedent's being false (i.e., it is not the case that you ought to *p*).

With this understanding of success propositions in hand, I will now justify the success propositions I assigned above. Let us first look at beliefs. Beliefs are, constitutively, truth-taking attitudes. If you believe that the cat is on the mat yet you become aware that it is not the case that the cat is on the mat, this will dispose* you to give up your belief that the cat is on the mat. The proposition "The cat is on the mat" thus qualifies as a success proposition of your belief. Or, in general, a belief that *p* is successful only if *p*.

I now turn to intentions. An intention is "a description of some future action, addressed to the prospective agent, and cast in a form whose point in the language is to make the person do what is described."[21] Intentions are, constitutively, truth-making attitudes; they aim at implementation. If you intend *p*, you are disposed* to make *p* true. An intention that *p* is therefore an executive success only if *p*; an intention that *p* is successful only if *p*.[22]

I now turn to all-things-considered ought judgments. Here, we face an initial difficulty. There are two principal views as to the type of attitude an ought judgment represents. An ought judgment can express a cognitive (and truth-apt) attitude, such as a belief, or it can express a noncognitive (and thus non-truth-apt) attitude, such as an intention or desire. If ought judgments are beliefs, then "You ought to *p*" is a success proposition of your judgment that you ought to *p*. This simply follows from the success conditions of ordinary nonnormative beliefs.

Moreover, it is plausible that ought beliefs also have executive success conditions. I assume that if you believe you ought to *p*, then you are disposed* to *p*. That is, like intentions, ought beliefs aim at implementation. In this case, *p* turns out to be a success proposition of a belief that you ought to *p*.

I will be slightly cautious with this condition, however. There may be circumstances where the executive disposition* of an ought belief does not give rise to this success proposition. Suppose you believe that you ought to *p*, yet you become aware that it is not the case that you ought to *p*. In this scenario, your awareness may cancel your disposition to *p* (in the very least, I am not in a position to exclude this). I will therefore offer a weaker proposal. Your ought

[21] Anscombe, *Intention*, 3.

[22] Moreover, I assume that "It is not the case that you ought to not-*p*" (i.e., the absence of an ought to the contrary) is a success proposition of an intention that *p* (although I will not rely on this assumption in explaining paradigmatic cases of structural irrationality). Nishi Shah supports this idea as follows: "My hypothesis is that the concept of intention includes a standard of correctness. Just as classifying an attitude as a belief entails applying to it the standard of being correct if and only if its content is true, likewise classifying an attitude as an intention entails applying to it the standard of being correct if and only if it is not the case that one ought not to perform the action that is its object" (Shah, "How Action Governs Intention," 12).

belief succeeds only if the material conditional "If you ought to *p*, then *p*" is true. That is, a necessary condition of the success of your ought belief is that *if* you ought to *p*, then *p*. So, from an executive point of view, your belief that you ought to *p* succeeds only if *p* or if it is not the case that you ought to *p*.

This weakening makes sense. The executive disposition to bring about *p* if you believe that you ought to *p* may be impaired if you are aware that it is not the case that you ought to *p*. I assume that awareness is factive. That is, if you are aware that it is not the case that you ought to *p*, then it is not the case that you ought to *p*. This is how, under such awareness, a belief that you ought to *p* can be successful even if you do not bring about *p*.

I now turn to a noncognitivist interpretation of ought judgments. On this interpretation, ought judgments are not truth apt. So, neither "You ought to *p*" nor "If you ought to *p*, then *p*" can qualify as a success proposition of your judgment that you ought to *p*. Nevertheless, for the noncognitivist, a key constitutive role of an ought judgment is to make you do what you judge you ought to do.[23] Like intentions, ought judgments are motivating states. They are truth-making attitudes: if you judge that you ought to *p*, you are disposed* to bring it about that *p*; *p* is thus a success proposition for the judgment that you ought to *p*. In fact, that ought judgments aim at implementation seems to be a corollary of the noncognitivist take on "ought."

I now turn to success with regard to preferences. Preferences, I take it, are dispositions to choose. So, as a minimal aspect of preferences, you prefer *a* over *b* only if you are in a mental state that is prone to cause you to *a* if you are prompted to choose between *a* and *b*.[24] On this view, there is an inherent difficulty in assigning success propositions to preferences. First, non-attitudinal mental states can constitute preferences (for example, a non-attitudinal perceptual state may dispose you to choose *a* over *b*). Second, there are a variety of attitudes that can constitute a disposition that amounts to a preference.

Suppose you prefer cycling to driving. Your preference may consist in a (comparative) desire. That is, you like cycling more than driving. Or it may be an intention-belief pair. You may intend to live healthily, and you believe that cycling is healthier than driving. Or this may be due to a comparative judgment, e.g., that cycling is healthier, better for the environment, or in any other comparative sense *F*er than *b*. Or it may be a directly comparative evaluative belief. You may believe that cycling is better than driving. Alternatively, it may be what I call

[23] Typically, noncognitivists treat this property of ought judgments as key evidence for their noncognitive status.

[24] Cf., e.g., Rabinowicz, "Value Relations" and "Modeling Parity and Incomparability."

a "conditional intention": you intend to cycle if you will either cycle or drive. I assume that all such attitudes can constitute a preference for cycling over driving.

David Hume declared that "a passion must be accompany'd with some false judgment, in order to its being unreasonable; and even then 'tis not the passion, properly speaking, which is unreasonable, but the judgment."[25] In a similar vein, I suggest that the success of a preference derives from the attitude that constitutes it. Thus, I do not think it necessary to list success propositions of preferences that are based on (normative) beliefs and/or (ordinary) intentions. If a belief and/or an intention amount(s) to a preference, then the success proposition of the resulting preference is simply the success proposition of the belief and/or the intention. For example, if a belief that *a* is healthier than *b* partly constitutes a preference for *a* over *b*, then one success proposition of this preference is the success proposition of the corresponding belief.

Thus far, I have refrained from ascribing success conditions to desires. This is a tricky issue. Extreme Humeans believe "that rationality allows a person to have any pattern of preferences whatsoever."[26] If this is correct, then this severely restricts the possibility of assigning success conditions to desires. If irrationality consists in the impossibility of success, then desires do not come with success conditions that could ever be necessarily incompatible. I will not dive into the extreme Humean abyss, however. Instead, I will put forward a moderate proposal. Take a preference for *a* over *b* that consists in a comparative desire. That is, you desire *a* more than *b*. I will assume that your preference comes with a disposition* to discard the preference if you become aware that not-*p*. So, analogously to how I identify success propositions above, *p* is a success proposition for that type of preference.

I will deliberately refrain from identifying any particular property here. I am only claiming that there is an evaluative and comparative property for which the discovery that *a* does not have more of that property than *b* would necessarily incline you to abandon your preference for *a* over *b*. For many philosophers, this suggestion will be more than natural. Anscombe suggested, for example, that desires aim at the good: "The conceptual connexion between 'wanting' … and 'good' can be compared to the conceptual connexion between 'judgment' and 'truth.' Truth is the object of judgment, and good the object of wanting."[27]

With these three assumptions in hand, it is only a small step to construe a success proposition for comparative desires. Suppose you prefer *a* to *b* based

[25] Hume, *A Treatise of Human Nature*, III.iii.3.

[26] Broome, *Ethics out of Economics*, 76.

[27] Anscombe, *Intention*, 76.

on a comparative desire. If the conceptual connection between "wanting" and "good" is analogous to the conceptual connection between "judgment" and "truth," it is natural to suppose that you are disposed* to give up your preference for *a* over *b* should you become aware that it is not the case that *a* is better than *b*. In this case, "*a* is better than *b*" turns out to be a success proposition of your preference for *a* over *b*. (Of course, this suggestion only works if the betterness relation holds independently of your preference; that is, "*a* is better than *b*" does not imply that you prefer *a* to *b*.)

In this paper, however, I need not rely on the assumption that desires aim at the good or that comparative desires aim at the better (although I am not denying this either). My proposal is slightly more modest and general. There only needs to be *some* relation "__ is *F*er than __" that holds between *a* and *b* for a comparative desire preference to succeed. That relation must be such that the realization of its absence necessarily disposes* you to drop the comparative desire. *F* may very well be some kind of goodness (desirability or choiceworthiness), but I will not commit myself to this assumption.

What about preferences that consist in a "conditional intention"? Following Ralph Wedgwood's suggestion, a conditional intention is a (complex) intention to do one thing—say *a*—and not another—say *b*—if you will do either *a* or *b*.[28] (I understand the "or" here as an exclusive disjunction, meaning that the choice between *a* and *b* is an exclusive one.) When going to a restaurant, for example, you may intend to eat monkfish and not chicken if you narrow down your choice to monkfish or chicken. Or when buying a used car, you may intend to buy a Saab 9-3 and not a Saab 9-5 if given a choice between the two. In this case, you have a conditional intention to buy a Saab 9-3. This intention exemplifies a preference: you prefer buying a Saab 9-3 to buying a Saab 9-5.

When do such preferences succeed? Suppose you prefer *a* to *b* because of a conditional intention to *a* given that *a* and *b* are your only options. With one exception, we cannot suggest (as I have done above for intentions *simpliciter*) that this intention is successful only if *a* is the case.[29] As a vegetarian, for example, you generally prefer eating no beef to eating beef. Yet when faced with the exclusive choice between beef and nothing (i.e., you will eat beef or nothing),

[28] "To form a preference for a proposition over the relevant alternatives is not necessarily to form a choice or intention to realize that proposition; it would at most be to form a *conditional* intention—in effect, the intention of acting in such a way that the proposition in question is true, rather than in such a way that the relevant alternative is true, *if one does either*" (Wedgwood, *The Nature of Normativity*, 120).

[29] Here is the exception: suppose you prefer *a* to not-*a*. Then the success proposition will be "If *a* or not-*a*, then *a* and not-not-*a*," which is equivalent to *a*. This is relevant to some of the examples I present below.

you prefer to eat beef to eating nothing. That does not mean, of course, that your preference succeeds *only if* you end up eating beef (i.e., eating beef is not a necessary condition for the success of this preference).

I propose "If (*a* or *b*), then (*a* and not-*b*)" as a success proposition for a preference (*a* over *b*) that is based on a conditional intention. This is a natural suggestion. A constitutive aim of such a preference is to bring about (*a* and not-*b*) whenever you either *a* or *b*. This is only the case if the following material conditional holds true: "If (*a* or *b*), then (*a* and not-*b*)" (where "or" and "and" signify the standard truth-functional (inclusive) disjunction and conjunction).

Here is the same point from another angle. When is a conditional intention (to *a* and not-*b* if you will do either *a* or *b*) *not* successful? The answer seems obvious. It is unsuccessful if *b* is true while *a* is not true. Suppose, again, that you have a conditional intention to eat beef if you will either eat beef or eat nothing, yet you end up eating nothing. Then your conditional intention will not be successful. That is why, when it comes to a preference for *a* over *b* that is based on a conditional intention, in order for it to succeed, "If (*a* or *b*), then (*a* and not-*b*)" must hold true (which requires that either both *a* and *b* are false or (*a* and not-*b*) is true).

It is fortunate, though, that we do not need to handle such a cumbersome success proposition. This is because "If (*a* or *b*), then (*a* and not-*b*)" turns out to be equivalent to "not-b." It is easy to show this. The truth of *b* suffices for the falsity of "If (*a* or *b*), then (*a* and not-*b*)," for it guarantees the truth of its antecedent and the falsity of the consequent. Likewise, the falsity of *b* suffices for the truth of "If (*a* or *b*), then (*a* and not-*b*)." If both *b* and *a* are false, then the antecedent is false and hence the conditional is true. If *b* is false and *a* is true, then both the antecedent and the consequent are true. So, the conditional is again true. Consequently, the negation of *b*—not-*b*—turns out to be logically equivalent to "If (*a* or *b*), then (*a* and not-*b*)." Thus, I will treat not-*b* as a success proposition for a preference for *a* over *b* that is based on a conditional intention to *a* and not-*b* if you will do either *a* or *b*.

To summarize this section, I have created a table of the success conditions proposed above (table 1). In this table, *O* refers to "You ought to . . ."; ◇ refers to "It is possible that"; and *F*er signifies a comparative relation. "—" signifies that I will refrain from specifying a particular success proposition (this is not to say, of course, that there is not one). ". . ." is meant to indicate that there could be more success propositions than those I have specified. Again, in explaining the irrationality of the paradigmatic cases of structural irrationality, I will rely exclusively on the success propositions in **bold**. I have listed the others for illustrative purposes only.

*Table 1*

| **<Attitude Type: Object>** | **Correctness Success** | **Executive Success** |
|---|---|---|
| <belief: $p$> | $\boldsymbol{p}, \ldots$ | — |
| <intention: $p$> | not-$O$ (not-$p$), … | $\boldsymbol{p}, \ldots$ |
| <ought judgment: $p$>$_{\text{<cognitive>}}$ | $\boldsymbol{Op}$, $\Diamond p, \ldots$ | **if $\boldsymbol{Op}$, then $\boldsymbol{p}$**, … |
| <ought judgment: $p$>$_{\text{<noncognitive>}}$ | $\Diamond p, \ldots$ | $\boldsymbol{p}, \ldots$ |
| <preference: $a$ over $b$>$_{\text{<comparative desire>}}$ | **$\boldsymbol{a}$ is $\boldsymbol{F}$er than $\boldsymbol{b}$**, … | not-$b$, … |
| <preference: $a$ over $b$>$_{\text{<conditional intention>}}$ | — | **not-$\boldsymbol{b}$**, … |

## 3. APPLYING THE SIMPLE ACCOUNT

Let us return to SISA. SISA says that a collection of attitudes* is irrational if and only if their success propositions entail$_{\text{me}}$ a contradiction. Recall that "entail$_{\text{me}}$ a contradiction" means that it is metaphysically impossible for a set of propositions to be true without a contradiction's also being true. In this section, I will demonstrate that this account can unify an explanation of the irrationality of contradictory beliefs, contradictory intentions, and cyclical preferences. In the next section, I will show how two modest modifications of SISA render it able to explain the irrationality of akratic incoherence and two types of instrumental incoherence.

Let us begin with contradictory beliefs:

$M_1$ {<belief: $p$>; <belief: not-$p$>}.

Let us form $M_1$'s set of success propositions. I assume that <belief: $p$> is successful only if $p$. Correspondingly, <belief: not-$p$> is successful only if not-$p$. $M_1$'s set of success propositions, $S_{M_1}$, reads as follows:

{$p$; not-$p$}.

This set most evidently entails$_{\text{me}}$ a contradiction. The success of the belief that $p$ thus precludes the success of the belief that not-$p$ (and *vice versa*). According to SISA, $M_1$ is thus irrational.

Next is $M_2$:

$M_2$ {<intention: p>; <intention: not-p>}.

Let us consider $M_2$'s set of success propositions. I assume that <intention: $p$> is successful only if $p$. Correspondingly, <intention: not-$p$> is successful only if not-$p$. $S_{M_2}$ thus reads as follows:

{$p$; not-$p$}.

$S_{M_1}$ and $S_{M_2}$ are identical. SISA also implies that $M_2$ is irrational. The success of an intention that *p* precludes the success of an intention that not-*p* (and *vice versa*).

I now turn to cyclical preferences. Recall $M_3$:

$M_3$ {<preference: *a* to *b*>; <preference: *b* to *c*>; <preference: *c* to *a*>}.

I said in section 2 that the success of a preference derives from the attitudes that constitute it. Suppose, first, that the preferences in $M_3$ are based on comparative desires. That is, you desire *a* more than *b*, *b* more than *c*, and *c* more than *a*. I assume that a comparative desire for *a* over *b* is successful only if *a* is *F*er than *b*, where *F* stands for a comparative property (i.e., anything *a* and *b* can have more or less of).

This generates the following set of success propositions $S_{M_3}$:

{*a* is *F*er than *b*; *b* is *F*er than *c*; *c* is *F*er than *a*}.

These propositions entail$_{\text{me}}$ a contradiction. The "___ is *F*er than ___" relation is necessarily acyclical.[30] The truth of any two of these propositions precludes the truth of the third. A set of cyclical comparative desire-based preferences is such that it is impossible for the set to succeed. In this sense, SISA can already explain the irrationality of cyclical preferences.

I now turn to preferences constituted in conditional intentions. Here, things are slightly more complicated. Suppose that $M_3$'s preferences are constituted by conditional intentions. I assume that a conditional intention to *a* and not-*b* if either *a* or *b* is successful only if not-*b* is true. Accordingly, for $M_3$, this generates the following set of success propositions $S_{M_3}$:

{not-*b*; not-*c*; not-*a*}.

Does this set entail$_{\text{me}}$ a contradiction? In fact, this hinges on one condition. Suppose that, as a matter of metaphysical necessity, at least one of *a*, *b*, or *c* is an exhaustive proposition. That is, at least one of them is necessarily true. Consequently, the propositions in $S_{M_3}$ could never be true together; their joint truth would require the truth of a contradiction, and so $S_{M_3}$ would entail$_{\text{me}}$ a contradiction.

Consequently, SISA implies that a set of cyclical conditional intentions is structurally irrational whenever we consider mutually contrary propositions. However, if it is possible for *none* of the options to be realized, then SISA does not predict structural irrationality. I admit that this is a considerable drawback for SISA. If conditional intentions are genuine preferences, then there seem to

[30] Cf. Broome, *Weighing Lives*, 51. For a famous disagreement, see Tempkin, "Intransitivity and the Mere Addition Paradox."

be sets of cyclical preferences that do not turn out to be structurally irrational. This is so for sets of cyclical preferences where it is possible not to realize any of the options involved.

However, I believe there are two somewhat satisfactory ways to fix this problem. Suppose you have a set of cyclical preferences. You prefer *a* to *b*, *b* to *c*, and yet *c* to *a*. Suppose your preferences consist of conditional intentions. In addition, let us assume one of two things: either you hold a meta-preference for *either* over *neither* of these options (i.e., you prefer [*a* or *b* or *c*] to not-[*a* or *b* or *c*]) or you believe that (*a* or *b* or *c*) will be the case.

Let us look at the meta-preference first. I will assume that it is a conditional intention, too. So, you intend to (*a* or *b* or *c*) and *not not*-(*a* or *b* or *c*) (double negation) if you will either (*a* or *b* or *c*) or not-(*a* or *b* or *c*). Since [(*a* or *b* or *c*) or not-(*a* or *b* or *c*)] is a tautology and *not not*-(*a* or *b* or *c*) is simply (*a* or *b* or *c*), this meta-preference comes down to a straightforward intention to (*a* or *b* or *c*). The success proposition for such an intention is (*a* or *b* or *c*). Second, suppose you believe that either *a* or *b* or *c* will be the case. The success proposition for such a belief is also (*a* or *b* or *c*).

Let us add this success proposition to $S_{M_3}$. This results in:

{not-*b*; not-*c*; not-*a*; *a* or *b* or *c*}.

This set entails$_{\mathrm{me}}$ a contradiction. If not-*b*, not-*c*, and not-*a* are true, then (*a* or *b* or *c*) is false. Likewise, if (*a* or *b* or *c*) is true, then either not-*b* or not-*c* or not-*a* is false. Hence, your preferences cannot succeed jointly. SISA thus explains the irrationality of conditional intention cyclical preferences *if* you also prefer *either* to *neither* of the involved options or you believe that at least one of the options will be realized.

What should we make of this result? By and large, I believe it makes sense. Suppose you have a set of cyclical preferences that are conditional intentions, but you prefer that they not be realized or you believe they will not happen. It seems to me that this would alleviate the irrationality of your intention-based cyclical preferences (at least from an executive point of view). However, if you do prefer any of these options to none of them or believe they will be realized, then you hold a set of attitudes that are structurally irrational.

## 4. Applying an Extended Account

I said that contradictory beliefs, contradictory intentions, cyclical preferences, and instrumental and akratic incoherence are paradigmatic of structural irrationality. Any credible account of structural irrationality needs to explain why these combinations of attitudes are irrational. However, contradictory beliefs

($M_1$), contradictory intentions ($M_2$), and cyclical preferences ($M_3$) demarcate the explanatory limit of SISA. I have shown that, by and large, SISA can explain and unify the irrationality of $M_1$, $M_2$, and $M_3$. With regard to akratic and instrumental incoherence, however, SISA remains inapt. I will therefore introduce and argue for an extension of SISA.[31]

I loosely defined akratic incoherence as a state in which you believe that you ought to do something without intending to do so. More precisely, you are akratically incoherent if and only if you are in a state where (1) you judge that, all things considered, you ought to *p*, (2) you believe that *p* only if you intend *p*, and (3) yet you have no intention to *p*. That is:

> ($M_4$) {<ought judgment: *p*>; <belief: *p* only if <intention: *p*>>; <not-intention: *p*>},

where "only if" is meant to cover any necessary condition that suffices to make $M_4$ structurally irrational. Here is why 2 is an integral part of akratic incoherence. Suppose you judge that you ought to relax. But you also know that intending to relax is not necessary for relaxing. In fact, an intention to relax will make it much less likely that you will relax. In such circumstances, you are not structurally irrational if you believe you ought to *p* yet you refrain from intending *p*.[32] Thus, to be genuinely akratic, you must affirm that you will not end up doing what you believe you ought to do unless you intend so.[33]

I roughly defined instrumental incoherence as failing to intend something you deem necessary for your intended ends (instrumental incoherence). More precisely, I will assume that one type of instrumental incoherence (I will discuss

[31] By this I do not mean that SISA can only explain the irrationality of $M_1$, $M_2$, and $M_3$. It may also explain the irrationality of incompatible credences, for example (or, more generally, the assignments of incompatible probabilities to propositions). Suppose (*a* or *b* or *c*) are mutually exclusive and exhaustive. Suppose you assign a particular probability to the truth of these propositions. Then it seems that rationality requires that *a*, *b*, and *c* must add up to one in your assignment of probabilities. One straightforward way to include this under SISA would be to say that the attitude that constitutes the assignment of probability to a proposition (i.e., the corresponding credence or belief) is successful only if it corresponds to the objective probability that the proposition will be true. If you now, for example, assign a probability of 0.3 to *a*, *b*, and *c*, then this will imply that at least one of your attitudes will not be fully successful. This holds, in fact, for all probability assignments that do not add up to one.

[32] Cf. Broome, *Rationality through Reasoning*, 170–72, and "Enkrasia," 433–34; Kiesewetter, *The Normativity of Rationality*, 192.

[33] I am also implicitly stipulating a restriction on your judgments that you ought to *p* (<ought judgment: *p*>) to judgments where you also believe that *p* is in your power and can be brought about by your intentions and/or actions (cf., e.g., Broome, "Enkrasia," 433–34).

another type later) consists in (1) intending *p*, (2) believing that an intention to *q* is necessary for realizing *p*, yet (3) not intending *q*. That is:

> ($M_5$) {<intention: *p*>; <belief: *p* only if <intention: *q*>>; <not-intention: *q*>}.

Let us consider $M_4$'s and $M_5$'s sets of success propositions. Let us first assume cognitivism about ought judgments. $M_4$'s set of success propositions—$S_{M_4}$—thus reads as follows:

> {*Op*; if *Op*, then *p*; *p* only if <intention: *p*>}.

If the ought judgment is noncognitive, $S_{M_4}$ reads:

> {*p*; *p* only if <intention: *p*>}.

Likewise, $M_5$'s set of success propositions—$S_{M_5}$—reads:

> {*p*; *p* only if <intention: *q*>}.

On the face of it, none of these sets entails$_{\text{me}}$ a contradiction. Unless some propositions within a set happen to be metaphysically incompatible with each other (which is, of course, not necessary), it is metaphysically possible for all propositions to be true (without the truth of a contradiction).

This demonstrates a significant limitation of SISA. $M_4$ and $M_5$ are structurally irrational, but for almost all common instances of $M_4$ and $M_5$, SISA cannot explain this. We need to extend SISA if we want to utilize its core idea in unifying an explanation of structural irrationality.

SISA says that *M* represents an irrational set of attitudes if and only if $S_M$ (i.e., the set of success propositions of the attitudes represented by *M*) entails$_{\text{me}}$ a contradiction. Let me apply a subtle extension of this. I propose that when determining whether it is possible for your attitudes to succeed, we need to consider not only current but also *absent* attitudes. Looking back at $M_5$, for example, SISA only considers the success propositions of the attitudes in $M_5$, i.e., {*p*; *p* only if <intention: *q*>}. However, it ignores the fact that $M_5$ also contains an absent attitude, i.e., <not-intention: *q*>. I argue that once we modify SISA to include this absent attitude, it will be able to explain the irrationality of $M_4$ and $M_5$ (as well as $M_1$, $M_2$, and $M_3$).

Spelled out slightly more formally, I propose that *M* represents an irrational set of attitudes* if and only if the *union* of *M* and $S_M$ entails$_{\text{me}}$ a contradiction. That is, if by merging the sets *M* and $S_M$ we gain a set for which it is metaphysically impossible that all propositions are true at the same time, then and only then is the set of attitudes* represented by *M* irrational. Let us call this the "extended impossible success account" (hereafter EISA). A formal version reads as follows:

> EISA *Formal*: $M$ represents an irrational set of attitudes* if and only if $M \cup S_M$ entails$_{\text{me}}$ a contradiction,

where ∪ symbolizes the union of two sets. Accordingly, a set of attitudes* represented by $M$ is irrational precisely when conjoining the propositions that represent those attitudes* with the propositions that represent the success of the attitudes in $M$ entails$_{\text{me}}$ a contradiction. Informally, EISA can be stated as follows:

> EISA *Informal:* A set of attitudes* is irrational if and only if holding those attitudes* precludes the joint success of the present attitudes in that set,

where "precludes" refers to metaphysical impossibility. That is, for $M$ to represent an irrational set of attitudes, it need not be *per se* impossible for the present attitudes represented by $M$ to succeed jointly. Rather, the idea is that it is impossible both for a person to hold the attitudes* picked out by $M$ and for the present attitudes represented by $M$ to succeed jointly.

Can this account explain the irrationality of $M_4$ and $M_5$? First, let us form the union of $M$ and $S_M$. We simply need to add <not-intention: *p*> to the two sets ("cognitive" and "noncognitive") of success propositions $S_{\text{M4}}$ and <not-intention: *q*> to the set of success propositions $S_{M_5}$. The result reads as follows:

{O*p*; if O*p*, then *p*; *p* only if <intention: *p*>; <not-intention: *p*>};

{*p*; *p* only if <intention: *p*>; <not-intention: *p*>};

{*p*; *p* only if <intention: *q*>; <not-intention: *q*>}.[34]

All three sets entail$_{\text{me}}$ a contradiction. Consider the first set: *Op* and if *Op*, then *p* formally entail *p*. Conjoining *p* with *p* only if <intention: *p*> formally entails <intention: *p*>. Conjoining <intention: *p*> with <not-intention: *p*> then entails the contradiction.

Here is what this means concretely. Consider *akratic incoherence* ($M_4$). Suppose you judge that you ought to *p* and you believe that you can *p* only if you intend *p*. I assume here that your ought judgment expresses a belief. Suppose that both the normative judgment and the belief are successful. That is,

you ought to *p*;

if you ought to *p*, then *p*;

and so *p* holds true. Moreover, it also holds true that

[34] Strictly speaking, the three sets do not represent the union of $M$ and $S_M$ because (to avoid unnecessary complexity) I have not included all present attitudes. The result remains the same, however.

> if $p$, you intend $p$.

However,

> it is not the case that you intend $p$.

This implies, necessarily, that as long as you hold the combination of attitudes* captured by $M_4$, either your normative judgment (that you ought to $p$) or your belief (that if $p$, you intend $p$) will not succeed. This is what makes $M_4$ irrational. Of course, the same result holds if we employ the noncognitive success proposition for ought judgments (compare the second set above).

The third set also entails a contradiction—namely, <intention: $q$> and <not-intention: $q$>. Suppose you intend $p$ and you believe that $p$ only if you intend $q$. Suppose that both the intention and your belief are successful. That is, both $p$ and ($p$ only if you intend $q$) hold true. The truth of these two propositions implies that you intend $q$. But you do not intend $q$. Hence, either your intention or your belief will not succeed. This is what makes *means-end absent intention incoherence* ($M_5$) irrational. In sum, EISA can track the irrationality of attitudinal combinations that contain absent attitudes.[35]

Finally, let us turn to another type of *instrumental incoherence*. Here, it is not an absent attitude that precludes the success of your attitudes, but rather a present attitude. Suppose you intend to go shopping. You believe that you will go shopping only if you do not intend to stay at home, and yet you intend to stay at home. Or more formally:

> ($M_6$) {<intention: $p$>; <belief: $p$ only if <not-intention: $q$>>; <intention: $q$>}.

I suggest that this pattern of attitudes is structurally irrational. Let us first form the set of success propositions $S_{M_6}$:

> {$p$; $p$ only if <not-intention: $q$>; $q$}.

As with $M_4$ and $M_5$, this set does not entail$_{\text{me}}$ a contradiction. From $p$ and $p$ only if <not-intention: $q$> we can derive <not-intention: $q$>. Conjoined with $q$, however, this does not entail$_{\text{me}}$ a contradiction. This holds under most replacements of $p$ and $q$. So, ordinarily, your attitudes do not undermine their own success.

[35] I am grateful to an anonymous reviewer for pointing out that John Brunero offers a similar explanation of the structural irrationality of means-end absent intention incoherence ($M_5$) (Brunero, *Instrumental Rationality*, 178, 197–98). However, unlike myself, Brunero remains skeptical of the claim that constitutive aims or the success of attitudes can explain the structural irrationality of other combinations of attitudes, such as *akratic incoherence* ($M_4$) or *cyclical preferences* ($M_3$) (*Instrumental Rationality*, 178, sec. 7.2.2).

Unlike $M_4$ and $M_5$, $M_6$ does not involve an absent attitude. But it involves another element that EISA considers—namely, *present* attitudes. EISA states that we should form the union of $M$ and $S_M$. This includes combining the attitude propositions of the present attitudes in $M$ with their success propositions.

$M_6$ contains three present attitudes (two intentions and a belief). For the sake of simplicity, we only need to add one of these to $S_{M_6}$ to generate a contradiction—namely, the intention to $q$. The resulting set reads as follows:

{*p*; *p* only if <not-intention: *q*>; *q*; <intention: *q*>}.

This set entails$_{\text{me}}$ a contradiction. From

*p*

and

*p* only if <not-intention: *q*>

we can infer <not-intention: *q*>. Conjoined with

<intention: *q*>

we arrive at a contradiction. So, according to EISA, $M_6$ turns out to be irrational.

What generates the irrationality of $M_6$? Again, unlike $M_1$–$M_3$, it is not the *success* of some of your attitudes that defeats the success of some of your other attitudes. Likewise, it is not the absence of an attitude (as with $M_4$ and $M_5$) that defeats an attitude's success. Instead, with $M_6$ it is the *presence* of an attitude that undermines the success of another attitude. The mere presence of <intention: *q*> makes the joint success of <intention: *p*> and <belief: *p* only if <not-intention: *q*>> impossible. So, in contrast to $M_4$ and $M_5$, you cannot overcome the irrationality of $M_6$ by adding another attitude to it. As with $M_1$–$M_3$, you need to eliminate at least one of your attitudes from the arrangement.

Let us take stock of where we are thus far. I have proposed that a set of attitudes* is irrational precisely when it is metaphysically impossible for the conjunction of the following to be true:

1. The propositions that pick out the present attitudes of the set
2. The propositions that pick out the absent attitudes of the set
3. The propositions whose truth is necessary for the success of the present attitudes of the set

In short, if a set of attitudes* is structured such that it is metaphysically impossible both to hold those attitudes* and for the present attitudes in that set to succeed, then (and only then) is that set of attitudes* irrational.

## 5. Structural Irrationality and Reflective Accessibility

I have shown that EISA manages to unify the six diverse yet paradigmatic instances of irrational patterns of attitudes* $M_1$–$M_6$. We now have *one* explanation of the irrationality of contradictory beliefs, contradictory intentions, cyclical preferences, and akratic and (two types of) instrumental incoherence. This is great progress indeed. EISA captures essential knowledge about what it *is* for a set of attitudes to be irrational.

However, does EISA as it stands state the real essence of structural irrationality? Or does this account require further refinements? Here is one issue on which this will depend.

Suppose you believe that 2 = 1 or you intend to be married to a bachelor:

<Belief: 2 = 1>;

<Intention: You are married to a bachelor>.

No doubt, your belief and your intention are substantially flawed. They are criticizable in many ways. You believe (or intend) a metaphysical impossibility. But does either of these attitudes suffice to make you structurally irrational (i.e., you are necessarily rendered structurally irrational if you adopt one of these two attitudes)?

If your answer is yes, then this supports the view that EISA captures the essence of structural irrationality. Consider the success proposition of your belief and your intention, respectively:

<2 = 1>;

<You are married to a bachelor>.

Both propositions entail$_{\text{me}}$ a contradiction: 2 = 1 requires not-(2 = 1); "You are married to a bachelor" requires the existence of a person who is married *and* not married (i.e., a bachelor).[36] Consequently, EISA implies that both attitudes are necessarily structurally irrational.

However, I believe there are good reasons not to count these attitudes as necessarily structurally irrational. There are mental environments (strange as they may be) in which such individual attitudes may turn out not to be structurally irrational. Suppose you have adopted a worldview with alternative axioms of arithmetic, or one that allows some contradictions to be true. Or you simply

[36] 2 = 1 | −1;
1 = 0
not-(1 = 0);
So: not-(2 = 1).

lack the conceptual or logical capacity to discern that you believe or intend something impossible. Indeed, I assume there are exceptional circumstances in which these may not be structurally irrational. EISA therefore turns out to be overly inclusive: it identifies as irrational combinations of attitudes that are perfectly rational.

In this final section, I will argue that there is one main reason for this. EISA does not presuppose that one has *reflective access* to the fact that one's attitudes cannot succeed jointly. For a set of attitudes to be structurally irrational, however, a person must, at least to some degree, be able to identify the fact that something about her attitudes is defective. Even if your attitudes cannot succeed, charging you with irrationality remains unwarranted as long as you are either genuinely unable or justified in failing to detect this.

In the remainder of this paper, I will therefore argue for three key refinements to EISA. I will defend three conditions that are necessary for the structural irrationality of a set of your attitudes:

1. The impossibility of joint success is transparent to the degree of a logically valid inference.
2. If you were to deploy your full logical abilities, you would be able to detect that the joint success of the attitudes implies a contradiction.
3. You do not have a justified paraconsistent belief that this contradiction is in fact true.

### 5.1. *The Traditional Solution*

Before I turn to defending these three conditions, I will briefly examine the traditional answer as to why individual attitudes do not count as structurally irrational. I will also explain why it is not a good idea to make EISA subject to the traditional view.

The traditional hypothesis is that structural irrationality is essentially *relational*. That is, structural irrationality, as the name indicates, can arise only from a mismatch among attitudes*. In an early contribution to this debate, Tim Scanlon explicates this point as follows. Claims of structural irrationality are

> structural because they are claims about the *relations* between an agent's attitudes that must hold insofar as he or she is not irrational, and the kind of irrationality involved is a matter of conflict between these attitudes.[37]

[37] Scanlon, "Structural Irrationality," 84–85 (emphasis added).

Or, as Niko Kolodny succinctly puts it: "Subjective rationality is a matter of the relations among one's attitudes."[38] More recent characterizations tie in with this: Worsnip specifies structural rationality as having "attitudes that are not jointly incoherent."[39] Kiesewetter writes that "structural irrationality is a kind of irrationality that we can detect simply by looking at a particular combination of attitudes that a person holds."[40] In sum, irrationality stems from a particular incoherence, disunity, mismatch, or lack of fit *between* attitudes.[41]

The core point of these characterizations seems to be this. For a set of attitudes to be irrational, that set must at least contain either (1) two *antagonistic* attitudes or (2) one attitude and the absence of a *complementary* attitude. For example, the set {<belief: *p*>; <belief: not-*p*>} serves as a fitting illustration of 1: <belief: *p*> and <belief: not-*p*> represent two antagonistic attitudes. The set {<ought judgment: *p*>; <belief: $\Diamond p \Rightarrow$ <intention: *p*>>; <not-intention: *p*>} serves as a fitting illustration of 2: the absence of <intention: *p*> represents the absence of a complementary attitude.

Should we accept this? First, it would satisfy the desideratum of excluding single attitudes from being irrational. By definition, a single attitude lacks (1) the presence of antagonistic and (2) the absence of complementary attitudes.[42] Second, it would be easy to make EISA sensitive to this assumption. We simply need to add the following condition to the account:

> *Multiple*: It is possible to generate the contradiction that $M \cup S_M$ entails$_{\text{me}}$ from multiple attitudes* (i.e., either at least two present or one present and one absent attitude) in *M*.[43]

[38] Kolodny, "Why Be Rational?," 530.

[39] Worsnip, review of *The Normativity of Rationality*.

[40] Kiesewetter, *The Normativity of Rationality*, 14.

[41] Reisner, "Is the Enkratic Principle a Requirement of Rationality?" Of course, interpreted strictly, Scanlon's point is perplexing. Irrationality cannot exclusively be a matter of conflict between attitudes. If (1) you believe that you ought to *p* and (2) you believe that *p* only if you intend *p*, yet (3) you do not intend *p*, you are irrational. But there is no conflict among your attitudes; 1 and 2 do not conflict, and 3 cannot "conflict" with any attitudes because it is the absence of an attitude.

[42] I consider a set with a present and an absent attitude to be a set with two attitudes*.

[43] It must merely be *possible* to generate the contradiction from $M \cup S_M$. Here is why. Suppose (1) you believe that 2 = 1. The success proposition of this belief generates a contradiction. But you are not necessarily irrational. However, suppose (2) you believe that 2 = 1 and you believe that not-(2 = 1). In this case, the two beliefs' success propositions generate the same contradiction as in 1. However, this time you are irrational. That is why there must merely be the *possibility* of generating a contradiction from $M \cup S_M$.

The consequence of this condition is clear. The fact that the attitudes in *M* cannot succeed jointly must be a consequence of how the attitudes* in *M* relate to each other. Individual attitudes thus cannot be irrational. As wrongheaded as your belief that 2 = 1 or your intention to be married to a bachelor may be, it is not necessarily irrational.

I agree that no individual first-order attitude should make a person necessarily structurally irrational. Nevertheless, I disagree that we should make EISA subject to Multiple. There are two chief reasons for this. First, the condition goes too far. There are cases where it exempts you from being irrational but where no such exemption should be granted. Second, it does not go far enough. It fails to exempt you from being irrational where such an exemption should be granted.

Here is my first objection. It shows that Multiple goes too far. Suppose you believe that (*p* and not-*p*). (Note that this is a first-order belief with a non-atomic content.) Many philosophers view such an individual belief in a contradiction as structurally irrational. If this is correct, then Multiple turns out to be a nonstarter. The success proposition of such a belief does entail$_{\text{me}}$ a contradiction, yet this contradiction is not generated from at least two attitudes*. Multiple thus precludes the irrationality of this belief, and there is no justification for that. This shows that Multiple does not state a prerequisite for structural irrationality.

Here is another aspect we should consider. I argue that making EISA subject to Multiple would be *ad hoc*. It does not help us to get to the core of structural irrationality. One and the same individual or object can figure in an attitude under various modes of representation. For example, one can represent the same person as "Batman" or "Bruce Wayne." Likewise, one can represent one and the same planet as "Hesperus" or "Phosphorus." Consider two examples where this becomes significant for structural irrationality. Suppose you believe that Batman is a hero and you believe that Bruce Wayne is not a hero. Or suppose you intend to observe Hesperus and you intend not to observe Phosphorus:

{<Belief: Batman is a hero>; <B: not-Bruce Wayne is a hero>};

{<Intention: You observe Hesperus>; <I: not-You observe Phosphorus>}.

First note that, according to EISA, these two sets are necessarily irrational. It is thus metaphysically impossible for Batman to possess a property that Bruce Wayne lacks. Likewise, it is metaphysically impossible to observe Hesperus without observing Phosphorus, and *vice versa*. Thus, the success propositions of both sets entail$_{\text{me}}$ a contradiction.

This reveals another defect of EISA. Subscribing to either of these two sets of attitudes does not suffice to make you irrational. You may blamelessly lack awareness that (Batman and Bruce Wayne) and (Hesperus and Phosphorus) are identical. The lack of awareness may not be a rational defect on your part. Again, EISA overshoots in ascribing irrationality.[44]

However, Multiple fails to provide an adequate solution to this problem. Multiple implies that irrationality can only arise for sets of attitudes* if one can generate the impossibility of success from the relationship that holds among those attitudes*. This is precisely the case for the two examples just discussed. Both represent *antagonistic* pairs of attitudes. Your belief that Batman is a hero succeeds if and only if your belief that Bruce Wayne is not a hero does not succeed. Your intention to observe Hesperus succeeds if and only if your intention not to observe Phosphorous does not succeed.

I conclude that Multiple turns out to be inadequate when it comes to adapting EISA appropriately. While it manages to exclude individual attitudes from being structurally irrational, it also excludes beliefs that are in flat contradiction from being structurally irrational. It also offers no solution for excluding attitudes that entail$_{\text{me}}$ a contradiction under various modes of representation from being structurally irrational. Multiple is not part of the essence of structural irrationality. We should not make ESIA subject to Multiple.

### 5.2. *Reflective Accessibility*

I propose an alternative, unified solution to the problems discussed in the previous section. I argue that irrationality presupposes *reflective accessibility*, as I shall put it. I will stipulate that reflective accessibility puts three fundamental constraints on any correct account of irrationality. First, it requires *objective transparency*. That is, a set of attitudes* is irrational only if it is objectively transparent that its members cannot succeed jointly when adopted. Second, it requires *subjective transparency*. That is, a set of attitudes* is irrational only if the person who holds the attitudes is able to infer that they cannot succeed jointly when adopted. Third, it requires *implicit approval*. That is, a set of attitudes* is irrational only if you are *not* justified in believing that the attitudes can succeed jointly when adopted. A correct modification of EISA must be sensitive to these three conditions, or so I shall argue in the following.

[44] Here is a more general expression of this problem. Suppose *p* and *q* are incompatible in the following sense: *p* entails not-*q*, and *q* entails not-*p*. However, suppose you are not in a position to know that. That is, even if you were to consider *p* and *q* while employing your full conceptual and logical abilities, you could not come to discover that *p* and *q* cannot be true together. EISA would still imply that you are irrational. Your attitudes cannot succeed together, yet this result is clearly untenable.

### 5.2.1. *Objective Transparency*

In this section, I will only focus on objective transparency. In the previous section, I mentioned two types of problems for EISA. First, some individual attitudes turn out to be necessarily irrational. Second, there are sets of attitudes that, due to their contents' mode or sense of representation, are not necessarily irrational, even though they cannot succeed jointly. Both types of problems can be resolved by introducing an objective transparency condition.

The basic idea behind objective transparency is this. Consider a set of attitudes* that cannot possibly succeed when adopted. Then, for these attitudes* to be irrational, this impossibility must be objectively transparent to a certain degree.

I assume that the degree to which the impossibility of success is objectively transparent depends on the (complexity of the) attitudes* and the inferential abilities one would have to deploy to ascertain that a set of attitudes* cannot succeed jointly. This relation is, of course, inverse: the more (complex) the attitudes and abilities needed to ascertain that a set of attitudes* cannot succeed jointly, the less objectively transparent the circumstances (and *vice versa*).

Thus defined, EISA presupposes an extremely low degree of objective transparency. In order to establish whether a set of attitudes is irrational, one needs to be able to ascertain virtually *all* metaphysically necessary falsities.[45] That is quite a tall order; no actual person satisfies this. Some metaphysical falsities may be too complicated for anyone to understand. Others may have yet to be discovered.[46] This exposes an elemental flaw in EISA. We should not accept an account of irrationality that allows for irrational combinations of attitudes that no actual person can identify as such. EISA presupposes insufficient objective transparency.[47]

[45] For example, I explained that EISA implies that intending to observe Hesperus while intending not to observe Phosphorous is irrational. In order to discern that, one needs to have a substantial and rather complex piece of information. One needs to be aware that Hesperus = Phosphorous. That is quite a substantial requirement.

[46] Here is a case in point: the axioms of arithmetic may either imply or contradict Goldbach's conjecture. So far, no one has settled the matter. Suppose, for the sake of argument, that the axioms of arithmetic contradict Goldbach's conjecture. Then, according to EISA, it would be irrational to believe both. But no one could actually show that this is the case (cf., Broome, *Rationality out of Reasoning*, 154).

[47] By contrast, consider a condition of objective transparency that would be too strong. Suppose that for a set of attitudes* to be irrational, the impossibility of success must be transparent to the degree of an explicit contradiction. By this I mean that for a set of attitudes $M$ to be irrational, conjoining $M$ with its success propositions $S_M$ must lead to a set that contains two explicitly contradictory propositions. In short, $M \cup S_M$ must contain "*p*" and "not-*p*." This condition is untenably strong. Looking at the paradigmatic examples of

I stipulate that the correct degree of objective transparency is best designated as "logical transparency." When considering a set of attitudes and their success propositions, one must be able to infer that not all attitudes in that set can succeed simultaneously *via a logically valid inference*. Here is a slightly more formal expression of this idea. Let $M$ again be the propositional representation of your attitudes*, and let $S_M$ be their success propositions. Then, in order for an account to pick out the irrationality of the attitudes $M$ represents correctly, the account must be subject to the following condition:

> *Objective Transparency*: $M \cup S_M$ entails a contradiction *qua* logically valid inference.

Again, by "contradiction" I mean a formal or explicit contradiction, i.e., a proposition and its negation ($p$ and not-$p$, for example). Moreover, it is critical to note here that I use "logically valid inference" in a rigid sense.

You may deem an inference logically valid if it necessarily preserves truth. In this sense, an inference from "Peter is unmarried" to "Peter is a bachelor," or from "Mary is an ophthalmologist" to "Mary is a doctor," would also be logically valid. Yet this is *not* my understanding of logical validity. I understand logical validity as strict formality or logical consequence. A logically valid inference is an inference the validity of which is entirely general, or topic neutral.[48] It preserves truth *qua* its logical form, not *qua* the substance or meaning of the proposition that constitutes it.[49] So, unlike the two abovementioned inferences regarding Peter and Mary, a logically valid inference remains valid even when abstracting from the particular meaning or semantic content of the objects that constitute it.[50]

I will rely on a rough-and-ready assessment to see whether an inference is logically valid. A logically valid inference remains strictly truth preserving even when one *anonymizes* the semantic content of the inference. For example, the inference "No fish is a mammal; some animals are mammals; so, some animals are not fish" satisfies this criterion. The inference remains truth preserving even when anonymized: "No $F$ is $M$; some $A$ is $M$; so, some $A$ is not $F$." The inference from "Peter is unmarried" to "Peter is a bachelor" does not remain

irrationality $M_1$–$M_6$, only $M_1$ and $M_2$ would qualify as irrational. These are the only two patterns of attitudes where $M \cup S_M$ contains two explicitly contradictory propositions. $M \cup S_M$ of $M_6$, for example, reads as follows: {$p$; $p$; <not-intention: $q$>; $q$; <intention: $q$>}; there is no explicit contradiction.

[48] MacFarlane, "What Does It Mean to Say That Logic Is Formal?"

[49] Cf. Beall and Restall, "Logical Consequence."

[50] Cf. MacFarlane, "What Does It Mean to Say That Logic Is Formal?"

truth preserving when anonymized: *P* is *U*; so, *P* is *B*. This inference preserves truth not through its logical form but through a semantic understanding of its contents.[51]

Let us now turn to how objective transparency contributes to establishing a unified account of irrationality. It fixes some of the defects of EISA and avoids the complications of Multiple.

In the previous section, I identified two distinct flaws of EISA. EISA overreaches in implying the necessary irrationality of attitudes that are not necessarily irrational. Recall, for example, the two individual attitudes mentioned at the beginning of the previous section:

<Belief: 2 = 1>;

<Intention: You are married to a bachelor>.

I have already shown how, in principle, Multiple can deal with this flaw. However, so can Objective Transparency. Consider the success propositions of your belief and intention:

<2 = 1>;

<You are married to a bachelor>.

Though both propositions entail$_{\text{me}}$ a contradiction, they alone do not constitute a logically valid inference that entails a contradiction. To show this, conceal everything that belongs to the semantic content of these propositions and leave only their formal structure. Assuming that the numerical values 2 and 1 are part of the semantic content of 2 = 1, an apt anonymization reads as the formal identity statement "$x = y$." Likewise, you may formalize "You are married to a bachelor" as "There exists one $x$: $Mx$ and $Bx$."

Trivially, neither the formal identity relation nor every existentially quantified conjunction implies a contradiction. The entailed contradiction is not a consequence of the formal structure of the statements. Neither statement implies a contradiction via a logical inference. Objective Transparency therefore corrects EISA in turning the attitudes in question into necessarily irrational ones.[52]

[51] Anonymizing a proposition involves a number of complicated issues. Of course, finding a clear-cut demarcation between semantic content and logical form is not always straightforward. Moreover, anonymizing must also be sensitive to the logical form that comes with a subject's representation of a proposition.

[52] This is not to say, of course, that by adopting either of these two attitudes you will not likely be irrational. I assume that most people believe that 2 ≠ 1 and that you cannot be married

Let us now turn to the second critical problem of EISA. Recall the following set of attitudes:

> {<Belief: Batman is a hero>; <Belief: not-(Bruce Wayne is a hero)>}.
>
> {<Intention: You observe Hesperus>; <Intention: not-(You observe Phosphorus)>}.

Since it is metaphysically impossible for your beliefs and intentions to succeed jointly, EISA picks them out as irrational.

I have already explained that Multiple is toothless when it comes to fixing the problem. Nevertheless, Objective Transparency manages to deal with it. Consider the success propositions of the attitudes just mentioned:

> <Batman is a hero>; <not-(Bruce Wayne is a hero)>.
>
> <You observe Hesperus>; <not-(You observe Phosphorus)>.

I treat "Batman is a hero" and "Bruce Wayne is a hero," as well as "You observe Phosphorous" and "You observe Hesperus," as two distinct propositions. Following Kripke, I take this to be a precondition for the possibility of your being able, for example, to intend (or believe) the one proposition without intending (or believing) the other.[53] Indeed, I take this to be a distinct possibility. It follows from the following relationship between accepting and believing:

> If an agent *A* sincerely, reflectively, and competently accepts a sentence *s* . . . , then *A* believes, at the time of *c*, what *s* expresses in *c*.[54]

I assume that one can "sincerely, reflectively, and competently" accept "You observe Phosphorus" while not accepting "You observe Hesperus." Thus, they are two distinct propositions.

This allows us to anonymize the success propositions of the patterns of attitudes above as follows:

> <*p*>; <not-*q*>;
>
> <*r*>; <not-*s* >.

to a bachelor. By adopting one of the attitudes in question, you will be subscribing to an irrational set of attitudes.

[53] "It also seems clear that there must be two distinct propositions or contents expressed by 'Cicero denounced Catiline' and 'Tully denounced Catiline.' How else can Tom believe one and deny the other? And the difference in propositions thus expressed can only come from a difference in sense between 'Tully' and 'Cicero'" (Kripke, "A Puzzle about Belief," 243). Cf. also McKay and Nelson, "Propositional Attitude Reports."

[54] McKay and Nelson, "Propositional Attitude Reports."

This makes it clear why, subject to Objective Transparency, EISA no longer implies the irrationality of the above attitudes. None of these pairs of success propositions licenses a logically valid inference to a contradiction. Given the absence of the insight that (Batman = Bruce Wayne) and (Hesperus = Phosphorus), no irrationality is on display here. So even if these attitudes cannot jointly metaphysically succeed, this is not sufficiently transparent to give rise to structural irrationality.

Unlike Multiple, Objective Transparency does not go too far here. It fixes the problem of assigning two attitudes, the incompatibility of which is disguised by the different senses of the attitudes' contents. This already suggests that Objective Transparency is preferable to Multiple as an appropriate weakening of EISA. There is further evidence that this is true, however. Recall that Multiple necessarily exempts a belief (or intention) in an explicit contradiction from being irrational. I have already indicated that such an exception can be granted in extraordinary circumstances (which I will define more closely in the next section). However, such an exception is far from *necessary*.

Objective Transparency does not imply a necessary exemption. Consider the anonymized success proposition of a belief in or intention to realize an explicit contradiction:

<*p* and not-*p*>.

Trivially, this entails a contradiction *qua* a logically valid inference. Thus, the requirement of *Objective Transparency* is consistent with the idea that single intentions and beliefs in explicit contradictions are attitudinally irrational. Objective Transparency, again, turns out to be superior to Multiple.

So far, I have shown three things. First, EISA goes too far in assigning irrationality to individual attitudes and combinations of attitudes with contents that differ in their mode of representation. Second, I have shown that the traditional solution to this problem, Multiple, is also inadequate. Multiple manages to avoid portraying individual attitudes as irrational, but it fails to deal appropriately with beliefs or intentions in flat contradictions and with attitudes that take different (or Fregean) attitudes toward one and the same thing or subject. Third, Objective Transparency turns out to be superior to Multiple. When considering the attitudes and their success propositions, the impossibility of success must be transparent to the degree of a logically valid inference. I suggested that this weakens EISA in the right way.

In order to satisfy Objective Transparency, EISA must be slightly adapted. I will call the adapted account the *Formal Impossible Success Account* (FISA). It reads as follows:

> FISA *Formal*: $M$ represents an irrational set of attitudes* if and only if $M \cup S_M$ entails$_{\text{fo}}$ a contradiction,

where "entails$_{\text{fo}}$" means "entails formally." I use "entails formally" as shorthand for "entails *qua* logically valid inference."

By replacing "entails$_{\text{me}}$" with "entails$_{\text{fo}}$" in EISA, we make the resulting account sensitive to Objective Transparency. The fact that an irrational set of attitudes makes joint success impossible must be transparent to the degree of a valid inference.

I have already shown that this adaptation manages to eliminate some of EISA's and Multiple's critical flaws. But how does it fare with regard to the irrationality of $M_1$–$M_6$? Can FISA still explain their irrationality? To begin, FISA straightforwardly preserves the irrationality of $M_1$–$M_2$ and $M_4$–$M_6$. For $M_1$ and $M_2$, this is virtually trivial; I will not demonstrate this here. As a proxy for $M_4$, $M_5$, and $M_6$, consider the union of $M$ and $S_M$ for $M_5$:

> {*p*; *p* ⇒ <intention: *q*>; <not-intention: *q*>}.

This set also entails$_{\text{fo}}$ a contradiction: *p* and *p* ⇒ <intention: *q*> entails$_{\text{fo}}$ <intention: *q*>, which, when conjoined with <not-intention: *q*>, entails a contradiction.

There is one outlier, however. Suppose you hold a set of cyclical preferences, as represented by $M_3$. Suppose these preferences are based on a set of conditional intentions. As discussed in section 2, I assume that a conditional intention to (*a* and not-*b*) if *a* or *b* is successful only if not-*b* is true. Accordingly, this generates the following set of success propositions:

> {not-*b*; not-*c*; not-*a*}.

According to EISA (and its predecessor SISA), there are situations in which these preferences turn out to be irrational. This is precisely the case if, and only if, *a*, *b*, and *c* are necessarily contrary propositions, i.e., it is impossible for all of them to be false. Then the three conditionals entail$_{\text{me}}$ a contradiction. According to FISA, however, this is no longer the case. The three propositions do not entail$_{\text{fo}}$ a contradiction. A set of conditional intention–based and cyclical preferences does not qualify for irrationality. Should we accept this?

In general, I believe we should. Suppose you have a cyclical set of preferences that is constituted by conditional intentions. Suppose you are innocuously unaware that there is no metaphysically possible world at which *a*, *b*, and *c* are collectively false. Perhaps you even justifiably believe that you can avoid failing to satisfy your conditional preferences simultaneously. Then you are not necessarily irrational. Moreover, FISA still permits a clear-cut scenario where a cyclical set of preferences makes you attitudinally irrational. As I explained in section 2, cyclical preferences of that kind make you irrational whenever you hold them in conjunction with a conditional intention–based preference to realize one of the options (i.e., *a*, *b*, or *c*) rather than none of the options (i.e., not-[*a*, *b*, and *c*]). The same is true if you also believe that it is impossible for none of the options to be implemented. As I explained in section 2, both your belief and your preference take "*a* or *b* or *c*" as a success proposition. So, by adding either this preference or this belief, the set of success propositions for your attitudes reads as follows:

{not-*b*; not-*c*; not-*a*; *a* or *b* or *c*}.

This set entails$_{\text{fo}}$ a contradiction. First, the first three propositions are jointly true only if *a*, *b*, and *c* are jointly false. However, the truth of the remaining success proposition depends on *a*, *b*, and *c not* being jointly false. Consequently, this set entails$_{\text{fo}}$ a contradiction. FISA can thus account for the irrationality of cyclical conditional intention preferences. But this is only the case if either you have a preference to realize at least one of these options or you have a belief that makes it objectively transparent to you (as it were) that your preferences cannot succeed jointly.

Let us now turn to cyclical preferences that are based on comparative desires. As discussed in section 2, I assume that a comparative desire–based preference for *a* over *b* succeeds only if *a* is (in some sense) *F*er than *b*. This generates the following set of success propositions $S_{M_3}$:

{*a* is *F*er than *b*; *b* is *F*er than *c*; *c* is *F*er than *a*}.

I have already explained that these propositions entail$_{\text{me}}$ a contradiction. The "___ is *F*er than ___" relation is necessarily transitive. Cyclical preferences that are based on comparative desires are thus irrational under EISA. But are they irrational under FISA as well? This poses an interesting question, for it depends on whether you consider the "___ is *F*er than ___" relation to be part of the semantic content or part of the formal structure of "*a* is better than *b*." Suppose it is entirely part of the semantic content. Then proper anonymizing would need to mask the semantic elements of "better than." We would need to represent "*a* is *F*er than *b*" as

*aRb*,

only conveying that *a* relates to *b*. The anonymized set of success propositions would thus look as follows:

{*aRb*; *bRc*; *cRa*}.

This does not entail$_{\text{fo}}$ a contradiction. Like the example above, this set entails$_{\text{fo}}$ a contradiction only if we add an attitude with the success proposition "(*aRb* and *bRc*) → not-*cRa*" to the above set. An example is the belief that if *a* is *F*er than *b* and *b* is *F*er than *c*, then *c* is not *F*er than *a*. Together with this belief, a set of desire-based preferences would turn out to be irrational.

Alternatively, we could suppose that at least the comparative part of the betterness relation belongs to the logical structure of the proposition "*a* is *F*er than *b*." Then "*a* is *F*er than *b*" already represents the anonymized proposition. The success propositions of desire-based cyclical preferences would read as follows:

{*a* is *F*er than *b*; *b* is *F*er than *c*; *c* is *Fer* than *a*}.

Arguably, this entails$_{\text{fo}}$ a contradiction, at least if we treat the inferences from "*a* is *F*er than *b*" and "*b* is *F*er than *c*" to "*a* is *F*er than *c*," as well as from "*a* is *F*er than *c*" to "not-(*c* is *F*er than *a*)" as formal—an assumption I do not find entirely implausible. As a result, a set of desire-based cyclical preferences would turn out to be irrational on its own.

### 5.2.2. *Subjective Transparency*

So far, I have defended the following picture of structural irrationality. For a set of attitudes* to be irrational, holding those attitudes* must preclude the joint success of the present attitudes in that set. But that is not enough. The mentioned impossibility must also be *reflectively accessible*. It must be entailed via a formally valid inference by the propositions that represent the attitudes and their success conditions.

In this section, I will add another (related but more subjective) condition. I assume that irrationality presupposes the *ability* to detect an error in one's attitudes. That is, you must be able to see that an irrational set of attitudes is defective. Suppose you hold two contradictory intentions. You must have sufficient logical and inferential abilities to see that you have arranged your attitudes inadequately.

It is crucial to note that I am only stipulating a counterfactual condition here. For a set of attitudes* to be irrational, there is no need to actually identify a necessary defect. That would be a blatant way of "over-intellectualizing" irrationality. Instead, the idea is that if you were to entertain an irrational set of attitudes, you would come to see that something was amiss.

Here is the general idea I have in mind. Suppose you intend to *p* and believe (*p* only if you intend *q*), yet you do not intend *q*. Suppose you call these attitudes* to your consciousness. So you say to yourself "I will *p*, but only if I intend *q*. But I do not intend *q*." If, upon employing your full logical and inferential capacities, you were to conclude that your attitudes are *not* incompatible, then

I suggest that you are indeed exempt from being irrational. In other words, you are irrational only if you have the capacity to infer that your attitudes are somehow defective or incompatible.[55]

The hardest thing in this context is to pin down exactly what kind of logical and inferential capacity irrationality presupposes. My suggestion is that the required capacity will coincide with the logical capacity to infer a contradiction from a set of attitudes and their success propositions. That is, suppose that a pattern of attitudes precludes those attitudes' joint success and that this is transparent to the degree of a valid inference. By adopting this pattern of attitudes, you are irrational only if you have the capacity to infer the contradiction that is entailed by your attitudes* and their success conditions. Put succinctly: you must be able to deduce a contradiction from $M \cup S_M$. This is the sense, I assume, in which irrationality presupposes the *ability* to detect a flaw in your attitudes*.

Before I add this condition to FISA, let me anticipate, albeit briefly, two possible critiques. On the one hand, you may think that this condition is too weak. As I envision it, the condition says that if you were to call a set of your attitudes* and their success propositions to your cognitive attention while entertaining your maximal inferential capacity, you would infer a contradiction. You may think that this counterfactual should be restricted further. Perhaps you think it should only include those attitudes to which you have conscious access.

The effect of that restriction would be clear. For a set of attitudes to be irrational, one needs to be able to consciously entertain them.[56] I believe that would be overly restrictive, however. For one, it would preclude modelling "unconscious or implicit biases" (i.e., unconscious or automatic prejudicial attitudes that predicate your social behavior) as structural irrationality. We should not deprive a theory of irrationality of that possibility. On the other hand, you may think that the condition is too strong. Suppose you intend *p* and believe that *p* only if you intend *q*, yet you do not intend *q*. Also, if you were to consider your attitudes* and their success propositions, you would infer a contradiction. Yet you explicitly deny that truth and implementation are success conditions for belief and intention, respectively. Should you not be exempt from being irrational?

My quick answer to this is no. The conditions for irrationality must be subjective, but not too subjective. In fact, Broome has already established this point

[55] Of course, there is an important caveat. The error you identify needs to relate to the constitutive aims of your attitudes. It is not enough to conclude, for example, that combining two attitudes is phenomenologically or aesthetically unpleasing or incompatible.

[56] Cf. Wallace, "Normativity, Commitment, and Instrumental Reason," 22.

convincingly.[57] Following this point and not repeating the argument here, I do not think that a theory of (ir)rationality should be predicated, by and large, on your subjective sensitivities toward success conditions. Moreover, in some implicit sense, I doubt that you can detach yourself from the success conditions of your attitudes. Success conditions are grounded in essential dispositions that come with your attitudes. If you competently judge that you ought to $p$, then, for as long as you are aware that it is not the case that you ought to $p$ or that $p$ is impossible, you will be disposed* to discard your judgment. This is the sense in which I suppose that the success of your attitudes plays a correcting or structuring role. You will normally discard an attitude when you become aware that its success conditions will not be satisfied. So at least in a dispositional or implicit sense, I take it that you will subscribe to the success conditions of your attitudes. It is this implicit adherence that makes it plausible to include success propositions as part of an inference that you need to be able to perform in order to count as irrational.

In order to make FISA subject to this condition, I shall qualify the account as follows. I will refer to the resulting account as FISA$^{+1}$:

> *FISA$^{+1}$ Formal*: $M$ represents an irrational set of attitudes* if and only if (1) $M \cup S_M$ entails$_{\text{fo}}$ a contradiction $C$ and (2) you can infer $C$ from $M \cup S_M$.

As mentioned above, it is important to emphasize the counterfactual character of "can infer." Irrationality does not presuppose that you must actually infer a contradiction by considering $M \cup S_M$. That would be a blatant case of over-intellectualizing one's account of irrationality. Rather, it presupposes that if you were made aware of the propositions in $M \cup S_M$ and you were to utilize your full logical capacities (i.e., making all the correct inferences you can make), you would infer a contradiction.

### 5.2.3. *Implicit Approval*

Finally, let us turn to the third condition of reflective accessibility: *implicit approval*. This condition says that if your attitudes* and success propositions entail$_{\text{fo}}$ a contradiction, you are irrational only if you *implicitly accept* the validity of the formal entailment and the falsity of the entailed contradiction. The acceptance must only be implicit insofar as it only requires the *absence* of a justified belief. I will explain this below.

[57] Broome, *Rationality through Reasoning*, sec. 6.2: "Your rationality cannot be judged entirely by your own standards" (93).

An example inspired by the so-called preface paradox illuminates the necessity of this condition. Suppose that, on the basis of good evidence, you believe that every single proposition in a book you have authored is true. Suppose, again on the basis of good evidence, that you also believe there is no book for which it holds that every statement is true. Clearly, the success propositions of your beliefs entail$_{\text{fo}}$ a contradiction. According to FISA$^{+1}$, this combination of beliefs turns out to be irrational (given that you can infer a contradiction from the two success propositions).

However, suppose you are also an accomplished dialetheist (I am thinking of someone like Graham Priest here). You have developed a precise view as to when a contradiction turns out to be true. You agree that (1) "Every single proposition in my book is true" and (2) "No book contains only true propositions" entails a contradiction, although you do not accept that the truth of 1 excludes the truth of 2 (and *vice versa*). From your considered standpoint, both of your beliefs can succeed at the same time. Are you irrational?

I would say no. You have constructed a refined argument from which it follows that both beliefs can succeed. You rely on an arsenal of grounds for why this is so. I assume you are *justified* in believing that your two beliefs can succeed simultaneously. So, if the world is as you justifiably take it to be, it is possible for your attitudes to succeed simultaneously. It is not credible to deem you irrational.[58]

Here is an actual case where this point becomes relevant. Suppose we treat the inference from "*a* is *F*er than *b*" and "*b* is *F*er than *c*" to "*a* is *F*er than *c*" as a formal entailment (I discussed an alternative to this view in section 5.2.1.) Suppose, however, that on the basis of your evidence, you justifiably deny the necessary transitivity of comparative "___ is *F*er than ___" relations (think of someone like Larry Temkin here). That is, you deny the validity of the inference from "*a* is *F*er than *b*" and "*b* is *F*er than *c*" to "*a* is *F*er than *c*." If your considered view is correct, then a set of cyclical preferences can succeed together. An ascription of irrationality would not be warranted.

In order to make FISA$^{+1}$ subject to these considerations, I suggest extending the account by adding a third condition (3) to it.

> *FISA$^{+2}$ Formal*: $M$ is irrational if and only if (1) $M \cup S_M$ entails$_{\text{fo}}$ a contradiction $C$, (2) you can infer $C$ from $M \cup S_M$, and (3) you have neither (a) a justified belief that the entailment from $M \cup S_M$ to $C$ is invalid nor (b) a justified belief that $C$ is true.

[58] Cf. Broome, *Rationality through Reasoning*, 91.

This added condition ensures that if you have reached a robust vantage point from which it is possible for a set of attitudes to succeed at the same time, it is no longer defensible to accuse you of irrationality, even if your view turns out to be incorrect. Recall that structural irrationality is tied up with subjective incoherence. Therefore, if a belief of yours that all attitudes* in a set *M* can succeed simultaneously is well founded and epistemically justified, we have no grounds to accuse you of being irrational in adopting *M*.

It is important not to conceive of this as an easy excuse. I will not adopt a particular position on epistemic justification here. However, this kind of justification must satisfy a few important constraints. First, justification must be an entirely internal matter and must supervene on the mind. Otherwise, my account will violate the rule that structural rationality supervenes on the mind.[59] Second, justification must be in some sense rigorous and demanding. Being justified in believing that a particular set of attitudes can be jointly successful is not something you can bootstrap into existence. So, an unfounded belief that you are justified will not suffice. The bar for justification must be set significantly higher. In fact, I imagine that only a handful of specialized philosophers and Buddhists (perhaps only Graham Priest!) satisfy this condition.

## 6. Conclusion

This paper has sought to establish the essence of structural irrationality. I aimed to determine what creates and unifies the domain of structurally irrational attitudes. In a nutshell, I argued that structural irrationality consists in the transparent suspension of the possibility of attitudinal success. The core principle of this account can be formulated as follows: a set of attitudes* is irrational if and only if holding those attitudes* precludes the joint success of the present attitudes in that set. I formalized this by assuming that every attitude comes with a set of success propositions. If that set, conjoined with the propositions that represent the present and absent attitudes, entails a contradiction, then we have identified an irrational set of attitudes.

Although the core principle alone comes with considerable explanatory power—in sections 3 and 4, I showed that it can unify six fundamental types of structural irrationality—I also argued that not every suspension of the possibility of attitudinal success instantiates a set of irrational attitudes. If, for example, you believe or intend a necessary falsity, then your attitudes cannot succeed, but you are not necessarily irrational. I argued that the suspension must be reflectively accessible. I have defined three conditions to guarantee this.

[59] Broome, *Rationality through Reasoning*, 89, 151–52; Wedgwood, “Internalism Explained,” 349.

First, the suspension of the possibility of success must be objectively transparent. That is, by looking at the attitudes* and their success propositions, it must be transparent to the degree of a formally valid inference that the attitudes in question cannot succeed jointly. Second, and related, you must have the logical ability to perform that inference. That is, by looking at the attitudes* and their success propositions and by mustering all of your logical and conceptual abilities, you must be able to discern that it is impossible for your attitudes to succeed. Third, you must not be justified in believing that the attitudes can succeed jointly or that the inference in fact lacks the property of being truth preserving. If these three conditions are met, then a set of attitudes turns out to be irrational.

As a consequence of this proposal, the following picture of irrationality emerges. Irrationality consists in a constitutive defect with regard to your attitudes. This defect stems entirely from how you have structured your attitudes*. In addition, you are able to discern the defect. Of course, whether this account can capture the full range of phenomena that are vulnerable to structural irrationality, which arguably includes certain combinations of graded belief/credence as well as combinations involving attitudes such as hopes, fears, and the like, is something future research will need to show.

As a final thought, suppose that the constitutive success of agency is invariably linked to the success of an agent's attitudes. Then the picture of rationality I have drawn in this paper is one in which structural rationality can be seen as the most foundational prerequisite for being a successful agent.[60]

*University of Bayreuth*
*julian.fink@uni-bayreuth.de*

## REFERENCES

Anscombe, G. E. M. *Intention*. Cambridge, MA: Harvard University Press, 1957.

Beall, Jc, and Greg Restall. "Logical Consequence." In *Stanford Encyclopedia of Philosophy* (Winter 2016). https://plato.stanford.edu/archives/win2016/entries/logical-consequence/.

Broome, John. "Enkrasia." *Organon F* 20, no. 4 (2013): 425–36.

———. *Ethics out of Economics*. Cambridge: Cambridge University Press, 1999.

———. *Rationality through Reasoning*. Chichester: Wiley Blackwell, 2013.

[60] I am grateful to John Broome, Benjamin Kiesewetter, Franziska Poprawe, Andrew Reisner, Olivier Roy, Vuko Andrić, and numerous anonymous referees for comments on earlier drafts of this paper.

———. *Weighing Lives*. Oxford: Oxford University Press, 2004.

Brunero, John. *Instrumental Rationality: The Normativity of Means-Ends Coherence*. Oxford: Oxford University Press, 2020.

Fink, Julian. "A Constitutive Account of 'Rationality Requires.'" *Erkenntnis* 79, no. 4 (August 2014): 909–41.

———. "Structural Irrationality Does Not Consist in Having Attitudes You Ought Not to Have: A New Dilemma for Reasons-Violating Structural Irrationality." *Australasian Journal of Philosophy* (forthcoming).

———. "What (In)Coherence Is Not." *Grazer Philosophische Studien* 99, no. 2 (October 2022): 125–34.

Hume, David. *A Treatise of Human Nature*. Oxford: Oxford University Press, 2000.

Katsafanas, Paul. *Agency and the Foundations of Ethics: Nietzschean Constitutivism*. Oxford: Oxford University Press, 2013.

Kiesewetter, Benjamin. *The Normativity of Rationality*. Oxford: Oxford University Press, 2017.

Kolodny, Niko. "Why Be Rational?" *Mind* 114, no. 455 (July 2005): 509–63.

Kripke, Saul A. "A Puzzle about Belief." In *Meaning and Use*, edited by Avishai Margalit, 239–83. Dordrecht: D. Reidel, 1979.

Lord, Errol. *The Importance of Being Rational*. Oxford: Oxford University Press, 2018.

———. "What You're Rationally Required to Do and What You Ought to Do." *Mind* 126, no. 504 (October 2017): 1109–54.

MacFarlane, John. "What Does It Mean to Say That Logic Is Formal?" PhD diss., University of Pittsburgh, 2000. http://johnmacfarlane.net/dissertation.pdf.

McKay, Thomas, and Michael Nelson. "Propositional Attitude Reports." In *Stanford Encyclopedia of Philosophy* (Spring 2014). https://plato.stanford.edu/archives/spr2014/entries/prop-attitude-reports/.

Rabinowicz, Wlodek. "Modeling Parity and Incomparability." In *Patterns of Value: Essays on Formal Axiology and Value Analysis*, vol. 2, edited by Wlodek Rabinowicz and Toni Rønnow-Rasmussen, 201–28. Lund: Lund University, 2004.

———. "Value Relations." *Theoria* 74, no. 1 (March 28, 2008): 18–49.

Reisner, Andrew. "Is the Enkratic Principle a Requirement of Rationality?" *Organon F* 20, no. 4 (2013): 436–62.

Scanlon, Thomas. "Structural Irrationality." In *Common Minds: Themes From the Philosophy of Philip Pettit*, edited by Geoffrey Brennan, Robert Goodin, Frank Jackson, and Michael Smith, 84–103. Oxford: Clarendon Press, 2007.

Shah, Nishi. "How Action Governs Intention." *Philosophers' Imprint* 8, no. 5

(July 2008): 1–19.

Temkin, Larry S. “Intransitivity and the Mere Addition Paradox.” *Philosophy and Public Affairs* 16, no. 2 (Spring 1987): 138–87.

Wallace, R. Jay. “Normativity, Commitment, and Instrumental Reason.” *Philosophers' Imprint* 1, no. 3 (December 2001): 1–26.

Wedgwood, Ralph. “Internalism Explained.” *Philosophy and Phenomenological Research* 65, no. 2 (September 2002): 349–69.

———. *The Nature of Normativity*. Oxford: Clarendon Press, 2007.

Worsnip, Alex. *Fitting Things Together: Coherence and the Demands of Structural Rationality*. Oxford: Oxford University Press, 2021.

———. Review of *The Normativity of Rationality*, by Benjamin Kiesewetter. *Notre Dame Philosophical Reviews*, July 2, 2018. https://ndpr.nd.edu/reviews/the-normativity-of-rationality.

———. “What Is (In)Coherence?” In *Oxford Studies in Metaethics*, vol. 13, edited by Russ Shafer-Landau, 184–206. Oxford: Clarendon Press, 2018.

*Journal of Ethics and Social Philosophy*
VOL. 26, NO. 2 · DECEMBER 2023

*https://doi.org/10.26556/jesp.v26i2.2317*

# THREE SHORTCOMINGS OF THE TROLLEY METHOD OF MORAL PHILOSOPHY

*Guy Crain*

Several authors have criticized what James O'Connor has called the "trolley method" of moral philosophy.[1] Named for trolley problems that typify cases used in the method, the method includes presenting cases where there are one or more parties, one of whom is the *agent*—the moral agent poised to make the relevant decision. There are *variables*—objects, settings, and other parties described in a scenario that only function as features either merely to be accepted by the agent or to be managed by the agent in some way (e.g., the trolley, a switch, five people on a track). There are *options*—the limited courses of action available to the agent to interact with some of the variables; importantly, each option presented seems morally problematic. And there is also a *respondent*—the person(s) to whom the trolley problem is presented and who is expected to identify the correct or best option available to the agent. Thus, while the trolley method includes the traditional trolley problem cases, it includes a far broader range of moral dilemma-esque thought experiments. For a more concrete idea of trolley method cases, consider the following examples:

> *Switch*: A man is standing by the side of a track when he sees a runaway train hurtling toward him: clearly, the brakes have failed. Ahead are five people tied to the track. If the man does nothing, the five will be run over and killed. Luckily he is next to a signal switch: turning this switch will send the out-of-control train down a side track, a spur, just ahead of him. Alas, there is a snag: on the spur he spots one person tied to the track: changing direction will inevitably result in this person being killed.[2]
>
> *Footbridge*: George is on a footbridge over the trolley tracks. He knows trolleys and can see that the one approaching the bridge is out of control. On the track further along from the bridge there are five people; the banks are so steep that they will not be able to get off the track in time.

[1] O'Connor, "The Trolley Method of Moral Philosophy."

[2] Edmonds, *Would You Kill the Fat Man?*, 8.

George knows that the only way to stop an out-of-control trolley is to drop a very heavy weight into its path. But the only available, sufficiently heavy weight is a fat man also watching the trolley from the footbridge. George can shove the fat man onto the tracks in the path of the trolley, killing the fat man, or he can refrain from doing this, letting the five die.[3]

*Ticking Time Bomb*: A terrorist has been captured, and you know that he has planted a small atomic bomb in a major city that is due to detonate in two hours. The terrorist will not tell you where the bomb is, and unless you use torture to obtain the information from him, thousands of people will die.[4]

*Jim the Botanist*: Jim is a botanist on an expedition in South America. He wanders into the central square of a remote village in which twenty people are restrained against a wall and being guarded by armed men in uniforms. Pedro, the officer in charge, questions Jim and comes to believe that his presence in the village is a mere coincidence. Pedro informs Jim that the captives are a randomly selected group of inhabitants that are about to be killed in order to put an end to recent acts of protest against the government. Pedro would like to honor Jim's presence by offering him the opportunity to kill one of the innocent villagers himself. If Jim accepts the offer, Pedro will release the surviving nineteen villagers. If Jim refuses, Pedro will kill Jim and the twenty prisoners. Violent resistance is not an option.[5]

*Harry the President*: Harry is the president and has just been told that the Russians have launched an atomic bomb toward New York. The only way the bomb can be prevented from reaching New York is by deflecting it, but the only deflection path available will take the bomb onto Worcester. Harry can do nothing, letting all of New York die, or he can press a button that deflects the bomb, killing all of Worcester.[6]

*Organ Transplant*: A surgeon knows of five seriously ill patients in a hospital who all urgently need organ transplants. Two require kidneys, two need lungs, and one needs a heart. An innocent, healthy, and young drifter with no family or attachments comes to the hospital for a routine

[3] Thomson, "Killing, Letting Die, and the Trolley Problem," 207–8

[4] Edmonds, *Would You Kill the Fat Man?*, 49.

[5] Williams, "A Critique of Utilitarianism." See also Moseley, "Revisiting Williams on Integrity."

[6] Thomson, "Killing, Letting Die, and the Trolley Problem," 208.

> checkup. If the surgeon chooses to sedate the drifter and harvest the drifter's organs, the five patients will live. If the surgeon does not do so, the five will die.[7]

I will use "trolley problems" to refer to any case relevantly similar to the cases above.

In the trolley method, respondents' answers to such cases are then argued to have one or more of a wide range of implications—from criticisms of major moral theories to how to behave in (allegedly) analogous real-world scenarios. The trolley method, then, employs what Laura Martena calls the probative and heuristic functions of trolley problems—where trolley problems are used either to test ethical theories and principles by whether they correspond to respondents' case-specific intuitions or to build ethical theories and formulate principles by inferences from respondents' case-specific intuitions.[8]

The trolley method has been widely criticized. The criticisms include:

- It detracts attention from the more important systematic or institutional factors that give rise to such bleak choices in the first place.[9]
- To the extent that the trolley method is treated as analogous to scientific methods, there are significant problems with both the internal and external validity of the thought experiments it uses.[10]
- It fails to test important ethical features of the agents forced to act—features such as resourcefulness.[11]
- It fails to be action-guiding.[12]
- It fails to predict what people would actually do in relevantly similar scenarios.[13]
- It often includes conceptions of agents who possess capacities not had by many or perhaps any actual human.[14]
- It is "outright harmful."[15]

[7] Edmonds, *Would You Kill the Fat Man?*, 33.

[8] Martena, "Thinking inside the Box," 385.

[9] Rennix and Robinson, "The Trolley Problem Will Tell You Nothing Useful about Morality."

[10] Wilson, "Internal and External Validity in Thought Experiments." See also Bauman et al., "Revisiting External Validity."

[11] O'Connor, "The Trolley Method of Moral Philosophy," 248.

[12] O'Connor, "The Trolley Method of Moral Philosophy," 250.

[13] Bostyn, Sevenhant, and Roets, "Of Mice, Men, and Trolleys."

[14] O'Neill, *Towards Justice and Virtue*, 41.

[15] Martena, "Thinking inside the Box," 385.

I find many of these criticisms apt. But there are three criticisms of the trolley method that have not, to my knowledge, been adequately articulated in the literature. In this paper, I will argue that trolley problems (and, therefore, the trolley method) have three significant shortcomings—namely, the foregrounding of high-stakes ethical choices, the *faux* anonymization of moral agents, and the mischaracterization of ethical decision-making.

## 1. Backgrounding the Foreground

Trolley problems tend to spotlight high-stakes moral decisions. They often involve the death of multiple parties, significant bodily injury, torture, acts of violence, the starvation of entire populations, and so on—decisions most people should never expect to face. I am not the first to point this out. Others have criticized the trolley method for its lack of realism. However, in my view, these critics commit the very error they identify in the trolley method and thus miss the depth of the shortcoming. For example, Christopher Bauman and colleagues point out that trolley problems lack "mundane realism"[16]—that is, how likely it is that the events in a study resemble those that participants confront in their everyday lives:

> Trolley problems also lack mundane realism because the catastrophes depicted in sacrificial dilemmas differ considerably from the type and scale of moral situations people typically face in real life. To illustrate this point, we measured how realistic our participants found trolley problems compared with short scenarios about contemporary social issues (viz. abortion and gay marriage . . .). . . . Few participants in psychology experiments have direct experience making quick decisions that determine who will live and who will die, and few would even expect to face anything even remotely similar.[17]

Notice that to illustrate the lack of "mundane realism" in trolley problems, Bauman and colleagues had subjects respond to an item about abortion. True, this issue reflects a greater realism than trolley problems. Still, it is hardly "mundane." It, too, as Bauman and colleagues put it, "[differs] considerably from the type and scale of moral situations people *typically* face in real life" (emphasis added). Replacing trolleys with abortion improves on the trolley method but not by much. The very test Bauman and colleagues use to correct for mundane realism itself lacks mundane realism.

[16] Bauman et al., "Revisiting External Validity."

[17] Bauman et al., "Revisiting External Validity," 542.

Barbara Fried comes the closest to the criticism I am raising when she writes that the trolley method

> has resulted in non-consequentialists' devoting the bulk of their attention to an oddball set of cases at the margins of human activity, while largely ignoring conduct that (outside the context of criminal activity and warfare) accounts for virtually all harm to others: conduct that is prima facie permissible (mowing a lawn, fixing your roof, driving a car down a city street) but carries some uncertain risk of accidental harm to generally unidentified others.[18]

From here, though, Fried proceeds to argue that the trolley method leads its authors and consumers to underestimate the frequency of tragic choices, which is a mistake because, so she claims, such choices are "ubiquitous" and, for the most part, "quotidian."

To illustrate, she considers the example of transportation infrastructure investment (specifically, trolley safety): "Suppose that if we invest $5 billion … in safety measures, we can reduce expected deaths or serious injuries from trolley accidents from one in every 10 million trolley trips to one in every 12 million trolley trips. Should we (must we?) make that investment?"[19] While almost no one will ever have to face a decision like the one described in Switch, many more persons will have to decide whether to invest a certain amount of money in safety protocols that will result in either a higher or lower number of expected deaths. However, in light of Fried's trolley safety example (which is intended as a stand-in for any relevantly analogous political policy decision), I take it that her use of "ubiquitous" and "quotidian" is intended to mean ubiquitous and quotidian only when compared to trolley problems and therefore that Fried's criticism is still different from mine.

It is easy to miss this, given Fried's misleading use of "we." "We" gives the sense that I, the reader, am included—I am a part of deciding what dollar amount to invest in trolley safety. But, in fact, I am not. Deciding what dollar amount to invest in this or other analogous collective projects is not a regular part of the ethical decisions I make in a day. I am not in a position politically to have the authority to make that decision in any direct sense. Perhaps if I were to vote on a ballot measure concerning transportation funding or for a candidate who might, in turn, vote on that issue, I would play a part in the decision; but even that act does not come up for me with any "quotidian" level of frequency. Even if I were politically situated to make that decision in a more direct

[18] Fried, "What Does Matter?," 506.

[19] Fried, "What Does Matter?," 512.

sense—for instance, if I happened to be a member of a city council whose job it was to propose or vote on some infrastructure budget—I would likely only make that decision no more than once a year. So, similar to Bauman and colleagues' use of abortion, the example Fried gives of a "ubiquitous" decision still does not represent the vast majority of ethical decisions faced by the vast majority of persons. While Fried rightly observes that the fact that the trolley method has led to a "lopsided allocation of attention" has itself not received adequate attention, neither has Fried, I think, paid it adequate attention. While Fried's aim is to spotlight tragic choices to show their commonness, the result is still, in practice, backgrounding the foreground.

In an attempt to pay this problem adequate attention, consider that among trolley problems, there is a glaring lack of attention paid to issues such as writing thank-you cards; empathizing with one's spouse; the frequency with which to call or visit one's elderly parents; a healthy relationship with one's neighbors; how to conduct oneself in a workplace breakroom; the language one uses when interacting with online acquaintances; the verbal or facial expressions one makes when dealing with customer service workers; appropriate amounts of sleep and rest; how to express volatile emotions; whether one should cover one's car with bumper stickers that have antagonistic slogans on them; how to help one's child transition from adolescence into adulthood; proper boundaries and communication practices with potential marital partners; whether to greet persons one is passing by; the importance of dietary health; putting money in fundraiser collection tins; drawing undue attention to oneself; validating the feelings of those with whom one is conversing; and things such as bragging, swearing, smiling, listening, joking, complaining, and gossiping. Some philosophers do give attention to some of these issues (though they do not employ the trolley method in doing so).[20] I suspect, though, that some might react to the above list with a sense that such small things are either not ethical issues or not worthy of the attention of ethicists. Surely, there is precedent for thinking this is not right. Consider that a leader of a world religion took the time to teach his followers about greeting others and that a saint of that same religion stressed the importance of smiling.[21]

In light of what I take to be the far more "mundane realist" list of ethical concerns I present above (the larger category of which I will refer to as the "ethics of the mundane"), both trolley method peddlers and their mundane realist critics contribute to the view that ethics is either rare or elite. They are

[20] Regarding interacting with online acquaintances, see Barney, "[Aristotle], *On Trolling*." Regarding swearing, see Roache "Naughty Words." Regarding volatile expressions, see Roache, "Honestly, It's Fine!" See also Olberding, "The Wrong of Rudeness."

[21] See Matthew 5:47; and Reilly, "10 Of Mother Teresa's Most Powerful Quotes."

not alone. Perusing tables of contents of applied ethics textbooks from the past few decades also gives the impression that the average person might only face a few ethical decisions in their lifetime (namely, concerning abortion, suicide and euthanasia, and reproductive technology) and that ethical decisions worth a significant amount of attention are made only by the few persons who occupy positions of significant political power (namely, concerning war and terrorism, criminal and capital punishment, environmental policy, affirmative action, and economic policy). Perhaps the closest such manuals come to quotidian issues are treatments of vegetarianism and drug use.[22]

That the trolley method foregrounds high-stakes moral decisions and pays no attention to the ethics of the mundane is a shortcoming for three reasons. First, these low-stakes issues make up the stuff of everyday moral life. Not only does most of daily life for most moral agents not concern the potential rogue doctor harvesting a healthy, unattached drifter's organs for the sake of saving five other patients, as in Organ Transplant, neither does *daily* life for *most* moral agents concern abortion (it is difficult even to imagine a world in which at least fifty-one percent of all moral agents decide whether to have an abortion *every day*) let alone allocating public funds for transportation safety. Granted, some moral agents face such issues some of the time. An event strikingly similar to Jim the Botanist took place in Colombia in 1987.[23] Self-driving-car designers do consider scenarios relevantly similar to Switch.[24] Masahiro Morioka argues that the trolley method resembles the rationale behind the United States's decision to drop atomic bombs in Japan.[25] Health care staff working triage with limited resources face utilitarian trade-offs not unlike trolley problems.[26] And, of course, a sizable number of agents do face the decision of whether to have an abortion.[27] Yes, these things happen, and they do deserve philosophical attention. But high-stakes moral decisions occupy the *daily* life of very few people. By foregrounding what either most people face rarely or what few people face at all, the trolley method, in effect, backgrounds most of moral life.

[22] See, Bonevac, *Today's Moral Issues*; Boss, *Analyzing Moral Issues* and *Ethics for Life*; Cahn, *Exploring Philosophy of Religion*; Hinman, *Contemporary Moral Issues*; MacKinnon, *Ethics*; Rachels, *Moral Problems*; Shafer-Landau, *The Ethical Life*; Soifer, *Ethical Issues*; and White, *Contemporary Moral Problems*.

[23] Lederach, *The Moral Imagination*, 13–16.

[24] Lin, "Robot Cars and Fake Ethical Dilemmas." However, some argue the scenarios are disanalogous in important ways: see Roff, *The Folly of Trolleys*.

[25] Morioka, "The Trolley Problem and the Dropping of Atomic Bombs."

[26] Kneer and Hannikainen, "Trolleys, Triage and COVID-19."

[27] Centers for Disease Control and Prevention, "CDC's Abortion Surveillance System FAQs."

Second, this foregrounding of high-stakes moral decisions can be precedent setting or norm setting. Practicing the trolley method or using it as a primary pedagogical tool—this pattern of action can build the impression that ethics just amounts to the rare decision faced by many or the normal decision faced by the few. It can create the impression that if one wants to do work in ethics, then one *ought* to focus on warfare, killing in self-defense, lethal self-driving-car accidents, healthcare funding, prison reform, climate change legislation, and so on. But if I, who will likely never be in a position to make any such decisions, am on my way to visit an incarcerated acquaintance for the first time, and I am wondering what I ought or ought not to say during the visit—from the perspective of the trolley method, well, that is not really the stuff of ethics. To illustrate the shortcoming, imagine Smith, who routinely pours time and energy into some matter of international conflict happening far away from where she lives and has few cognitive resources left over to recognize the importance of energetically applauding at the end of her son's band recital: the foreground of her moral life—the stuff most frequently proximate to her—gets backgrounded due to misdirected attention. By this practice of directed attention, the ethics of the mundane are out of sight, out of mind, and the rare or elite matters become the paradigm cases of doing ethics, let alone behaving ethically.

Third, the trolley method is, by design, a terrible tool for working on the ethics of the mundane. This is because trolley problems differ from the ethics of the mundane in at least three ways. First, trolley problems feature neatly quantified choices—this life lost or that life lost, one life versus five lives, *x* amount of time inflicting serious injury versus thousands of lives lost, and so on. But what does a neatly quantified choice about thank-you cards even look like? It is true that in choosing to spend *x* amount of time writing a thank-you card, I have made some trade-off—the opportunity cost of having spent that time doing something else that, perhaps, would have had a higher moral payoff. But worrying about spending five minutes writing a thank-you card versus spending four minutes and devoting that one minute to some other activity seems wrongheaded. True, people routinely make decisions about these trade-offs—when to tell the kids it is time to leave the park and go home, how much time to allot to visiting a friend, and so on. But aiming to make those choices quantifiably precise itself seems like ethically bad practice. Imagine meeting your friend Smith for lunch and saying, "Smith, to maximize the moral payoffs of our friendship or avoid a morally impermissible use of my time, you will stay here interacting with me for at most forty-three minutes; however, I have determined that even one minute more is morally subpar, given that it will mean one minute fewer than the morally optimal amount of time spent writing thank-you cards this afternoon." Second, trolley problems are, by design, urgent.

Grave consequences loom no matter what the agent in a trolley problem does. But, again, viewing or acting as though each mundane ethical choice made in a day is a crisis seems wrongheaded and itself like bad ethical practice (let alone mentally unhealthy). Third, trolley problems involve only bad options. Such dilemmas are not analogous to my options when, say, considering whether I ought to brag or complain about something.

In sum, not only does the trolley method ignore most of most people's moral life and, in effect, thereby consign most of life to the realm outside ethics, but it also lacks the tools necessary to remedy this.

## 2. The Faux Anonymization of Moral Agents

Trolley problems are presented in one of two ways. There are second-person trolley problems that are written such that the agent is the respondent (such as in Ticking Time Bomb). And there are third-person trolley problems that are written such that some nondescript third party (e.g., "Jim the botanist," "Harry the president," or "a surgeon"), not the agent, is the respondent. First, consider third-person trolley problems. When the agent is just "someone" or "a person" or even "a surgeon," this creates the impression that the respondent can imagine a nearly featureless agent navigating the case—"nearly featureless" because in some cases limited details are provided ("Jim" is a botanist and "Harry" is the president); but also "nearly featureless" because the respondent must imagine *some* minimum set of features so that the agent is capable of acting *as a moral agent* in the case. So, the respondent is surely thinking of an agent who is, for example, not in a coma. Whatever features the agent has beyond this minimal set, plus whatever features are provided in the description, are treated as operationally unimportant.

Others have pointed out that the parties mentioned in trolley problems are under-described and that the details left indeterminate are potentially morally relevant[28]—though, to my knowledge, not a great deal of attention has been paid specifically to the under-described nature of the agent.[29] Virginia Held, however, has criticized the "dominant moral theories" (i.e., Kantianism, utilitarianism, and Aristotelian virtue ethics) because they operate as though there is such a thing as a nondescript "agent as such."[30] This chimerical abstraction, according to Held, lacks thick interconnectedness to other agents as such, and

[28] See Bauman et al., "Revisiting External Validity," 542; Wilson, "Internal and External Validity in Thought Experiments"; and JafariNaimi, "Our Bodies in the Trolley's Path."

[29] O'Connor, "The Trolley Method of Moral Philosophy," 246–48.

[30] Held, *The Ethics of Care*, 13.

thus the dominant moral theories miss moral issues that arise in the contexts of family, friendship, and social groups.

The criticism of the trolley method I aim to develop here builds on but differs from Held's. Consider this progression of three trolley problems I routinely present to ethics students after they have read excerpts from Held's *Ethics of Care*:

> *Bystander*: Suppose Smith is an unarmed bystander who witnesses Jones initiate a violent attack against an innocent victim, Williams.
>
> *Grandfather*: Suppose you are taking a slow, leisurely stroll with your elderly grandfather, who is heavily dependent on a walking cane. You tell him you want to go into the store right behind a sidewalk bench to buy a couple of items. He can rest on the bench, and you will continue the walk after you are done in the store. While you are in the store, your grandfather notices that across the street, a muscular male youth brutally attacks a smaller mid-thirties male. Your grandfather is the only bystander witnessing this attack and is unarmed.
>
> *Marine's Wife*: Suppose a short, 100-pound female notices that her tall, muscular, 280-pound former marine husband is being violently attacked by a 200-pound, unarmed male who is less fit than her husband.

For each case, I ask students whether the agent is obligated to intervene (violently, if necessary) to try to stop the attacker and protect the victim. The majority of respondents do not give the same answer in all three cases. A considerable majority say that the agent is obligated to intervene in Bystander but not in Grandfather or Marine's Wife. I, then, ask if more detailed descriptions of the agent and variables in each case affected their answers. Again, a considerable majority say yes. Last, I ask them to give a brief description of "Smith"—the person they imagined as the agent in Bystander. Overwhelmingly, student responses indicate that they imagined an able-bodied male. Interestingly, many of these responses are worded such that the students do not even realize that they themselves assigned a gender to Smith—responses such as, "I imagined him as young and fit." I, then, point out to students that there is nothing in Bystander that indicates that Smith was *not* an elderly grandfather or a marine's wife, let alone that Smith was male. (I once had a student resist this point, insisting that she could "just tell he was a guy" without being able to provide any further explanation of how she was able to tell this.)

I offer this not as an experiment performed with any scientific rigor but at least as an illustration of Held's view of the shortcomings of using this "agent as such" in any method of doing ethics, and also of something more. The cases function in much the same way as the infamous father/son/surgeon riddle:

> A father and a son are in a terrible car crash. They are both rushed to the hospital. The father is pronounced dead on arrival, while the son is in a critical condition. The son needs emergency surgery and is rushed to the operating room. The surgeon looks at the boy and says, "I cannot perform this operation—that is my son!" Who is the surgeon?

Respondents come up with a variety of solutions but routinely fail to come up with the solution that the surgeon could be the boy's *mother.* The big "aha"—the reason that it is a riddle with the potential to stump people at all—is that the respondent will likely supply more information than is actually presented in the riddle itself, and that supplied information is precisely what excludes the mother solution. Yet another "aha" is that the failure to come up with the mother solution is not endemic to male respondents. Females, females with mothers who were doctors, and even female doctors also tend not to come up with the mother solution.[31]

According to James Wilson, thought experiments similar to trolley problems are presented as though it is expected that the respondent will fill in the details left indeterminate but "only add colour and detail that is morally irrelevant."[32] I disagree. I think their presentation is more insidious. Third-person trolley problems give the impression that but for the sparsest of features, the agent is *anonymized*; any other features are operationally unimportant, and therefore, analysis can proceed as though those details are left indeterminate. The problem revealed by the surgeon riddle and my Bystander, Grandfather, and Marine's Wife cases is that respondents do *not*, in fact, proceed as though those features are left indeterminate, even when they take themselves to be proceeding in that way. This sense of anonymization is merely a veneer. Blanks *are* filled in—likely in an unwitting way. And more importantly, those blanks tend to get filled in strikingly similar ways—my students tend to assume "Smith" is an able-bodied male without realizing they have done so, and both males and females tend to assume that the surgeon must somehow be male without realizing they have done so.

Thus, what is taken to be a method involving anonymized agents actually involves agents who are rather descript. Consequently, the features that are implicitly supplied become part of the model agent operating as a stand-in for just anyone—anyone who is a moral agent. The more this becomes part of the practice of the trolley method, the more these sneakily supplied features become integral parts of *the* stand-in moral agent. If those integral parts include anything like "young and fit male" or "surgeon that must somehow be a male,"

[31] Barlow, "BU Research." See also Gil, "The Fifth Floor—Riddle Me This."

[32] Wilson, "Internal and External Validity in Thought Experiments," 138.

then that stand-in for just anyone does not really represent just anyone. If trolley method users are operating in a way that excludes female surgeons, elderly grandfathers, or wives of marines from being the agent, then it is not hard to imagine many more well-described agents who are likely being operationally excluded. This is what strikes me as particularly problematic about this faux anonymization of the trolley method: underneath the veneer of operational inclusion is a practice that is rather exclusive. The trolley method is not ethics for everyone.

Do second-person trolley problems avoid this problem? I do not think so. The degree to which second-person trolley problems are operationally inclusive is an empirical matter concerning the demographic makeup of respondents. In one case of empirical research, trolley problems (specifically, Switch and Footbridge) were posed to seventy thousand participants in forty-two different countries.[33] Perhaps, then, moral psychology's use of trolley problems is becoming more diverse, but surely, the majority of respondents to the broad variety of trolley problems being used *as part of the trolley method* are either academics or college students. If that is true, then most respondents imagining themselves as the agent are likely imagining an agent who is at least WEIRD; this acronym, which stands for wealthy, educated, industrialized, rich, and democratic, was coined by social scientists raising the criticism that social scientists have poor grounds for taking their findings to be representative of human beings in general when the demographics of their sample base likely represent less than twelve percent of the world's population.[34]

I say "*at least* WEIRD" because, as revealed previously about third-person trolley problems, the imagined agents are likely even more particular—relatively young (to the exclusion of the elderly) but not excessively young (to the exclusion of younger adolescents and children); able (to the exclusion of not only those with severe physical limitations but likely even those with lesser limitations such as dependence on a walking cane); and, if considering academic respondents specifically, probably male. Similar to third-person trolley problems, if second-person trolley problems are presented to enough respondents, it can create the sense that the agent has been anonymized—the agent could be any one of the many respondents, say, seated in a large university classroom taking an ethics course or any given reader of an academic work. But on closer

[33] Awad et al., "Universals and Variations in Moral Decisions Made in 42 Countries by 70,000 Participants."

[34] Henrich, Heine, and Norenzayan, "Most People Are Not WEIRD," 29. See also Henrich, Heine, and Norenzayan, "The Weirdest People in the World?"; and Arnett, "The Neglected 95%."

inspection, the actual respondents used in the trolley method are a notably homogenous group, and the agent is therefore not so anonymized after all.

Because of this faux anonymization, the trolley method has an agent-inclusivity problem. This is not a gripe about excluding diverse intuitions. Others have pointed out this worry. In response to a meta-analysis suggesting that the bulk of trolley problem respondents have largely similar intuitions, Edouard Machery and Stephen Stich conducted a larger meta-analysis showing there are significant differences in trolley problem intuitions among different demographic groups.[35] According to Machery and Stich, it would be a "disaster" if philosophers, psychologists, and anthropologists became convinced that "OUR intuitions (i.e., the intuition of educated, white, wealthy, Western people) are human intuitions."[36] The criticism I am raising is distinct from but analogous to theirs—that the trolley method's faux anonymization of moral agents can insidiously create the impression that to be a moral agent at all is to be a WEIRD, young, and fit male.

Recall that a significant number of respondents to the surgeon riddle—whether male or female—are equally stumped because they associate "surgeon" with maleness. The point is that however similar people's intuitions are about trolley problems, they are likely based on a conception of the category of "moral agents" that is itself exclusionary. Trolley method ethics represents an ethics for the educated, healthy, young, and able. Telling someone the relevant agent is an elderly grandfather seems to change things even though the cases never explicitly excluded such agents from being the nondescript third party. Those second-person respondents engaged in trolley method ethics are largely relatively wealthy persons highly educated in Western university systems. There are differently abled persons, persons in eldercare facilities, persons at the margins of stature or age, persons who lack much formal education, and so on who have to make moral decisions most days of their lives. The trolley method, in effect, leaves them out.

## 3. THE MISCHARACTERIZATION OF MORAL DECISIONS

Trolley problems present a fairly univocal model of what ethical decision-making is like. To illustrate that model, consider what I will call the myth of the voting booth. Imagine being enclosed in a voting booth. There are two boxes in front of you with small slits through which tokens can be inserted. You hold one

[35] Knobe, "Philosophical Intuitions Are Surprisingly Robust across Demographic Differences."

[36] Quoted in Weinberg, "Philosophical Intuitions and Demographic Differences." See also Stich and Machery, "Demographic Differences in Philosophical Intuition."

token. Dropping the token into one box as opposed to the other will indicate that you have chosen in favor of one outcome as opposed to the other. Within the booth, you are aware of both outcomes, and your consideration of which outcome you ought to favor has occurred entirely during the time in which you occupy the voting booth. You reach out and drop the token into one of the boxes, and this action is the result of your conscious consideration and deliberation while in the voting booth.

Granted, this description of voting sounds cartoonish. But it illustrates a certain conception of what ethical decision-making is like. The voting booth conception is one of a historyless, radically uninfluenced moral decision. That decision is made between at least two options of which the agent is fully aware. The agent's phenomenological experience of that decision includes being fully mentally present and the act being entirely the result of the agent's in-the-moment considerations; and these phenomenological experiences are veridical. This is not the standard complaint that trolley problems routinely involve levels of knowledge and certainty that are disanalogous to real-world ethical decision-making, as others have pointed out.[37] Rather, I mean to highlight the isolated, free-of-historical-influence, and performed-with-extreme-awareness nature of this conception of ethical decision-making.

In contrast, consider what I will call the water-walking analogy. Growing up, I had a neighborhood friend with a swimming pool. In the summer, I would go to his house to swim. Often, while we were swimming, his mother and sister would get in the pool to exercise. They would set a kitchen timer and leave it on the edge of the deck. For fifteen minutes, they would walk the internal perimeter of the circular pool as fast as they could. At first, we stayed in the middle of the pool to stay out of their way. But around the ten-minute mark, it was fun to join them at the perimeter because they had built up a current. Their water walking eventually created a makeshift lazy river. The current made it easy to walk behind them. We could even lift our legs and let the current move us. At the fifteen-minute mark, the kitchen timer would buzz. That meant it was time for my friend's mom and sister to turn around and walk in the opposite direction for fifteen more minutes. That turnaround was an interesting experience. Suddenly, for an eight- or nine-year-old, it was nearly impossible to walk the perimeter of the pool. The current against us was so strong that it was a struggle to make any forward progress. We would move our feet in the right direction, but due to the current, we still traveled backward, despite taking forward-moving steps. Eventually, persevering with forward-moving steps would start to pay off, and we could not only move forward but slowly accelerate. By the end

[37] Lieberman, "Fight the Hypo," 14–15.

of the second fifteen-minute portion, we had reversed the current, though the second was not as strong as the first.

The water-walking analogy illustrates a conception of ethical decision-making different from the myth of the voting booth. Unlike the stationary, enclosed voting booth, the water-walking analogy includes motion through a medium. It thereby illustrates activity that is dynamic and influenced. There is causal interplay between internal and external factors, and the influence of those factors themselves is, in part, a product of the agent's past decisions and actions—an interplay that creates a behavioral momentum. Because of this, to attempt to understand the agent's actions as a series of isolable moments where the agent faced a clear and limited set of discrete options from which to choose (e.g., "so now at time *t* I have the option to plant my foot on the pool floor with amount of pressure *p*, and I have the option not to do that—which is best?")—in other words, to attempt to understand it in a voting-booth-like way—mischaracterizes those acts. The acts exhibit varying degrees of mental presence, some of which are performed as part of a pattern (where that attempted pattern might include mistakes), and the performance of those acts, whatever their degree of mental presence, also cannot be legitimately conceptually isolated from the influence of the medium through which the motion takes place. Rather, each act on the agent's part is better characterized as a part of the whole activity and its environment—as part of a flow.

The myth of the voting booth and the water-walking analogy illustrate two different conceptions of moral decision-making. Of the two, the myth of the voting booth best captures the conception common to trolley problems. There are real-world moral decisions that are more akin to the voting booth than to water walking. And I concede that when discussing difficult choices, people often take themselves to be in situations more akin to the voting booth. Nevertheless, this exclusive focus on voting-booth-like decisions is problematic for the trolley method because the water-walking analogy better represents most real-world ethical decision-making.

The trolley method fails to capture, as psychologist John Bargh describes it, the "automaticity of everyday life."[38] An overwhelming number of daily decisions (let alone ethical ones) are made on autopilot, and that autopilot sequence draws on internal, unreflective resources to respond to and navigate environmental features and cues. Consider the activity of driving home from work. For many people, dozens or hundreds of potentially injurious, if not nearly fatal, decisions are made during that activity. Many of those decisions are arguably ethical in nature—for example, whether to obey traffic laws. Yet, as I suspect

[38] Bargh, "The Automaticity of Everyday Life."

many readers can also attest, it is possible for me to complete that activity, arrive home, and have no lucid memory of taking part in the activity in any careful, deliberative way. And that is because I probably did not take part in that activity in a careful, deliberative way. For this reason, it seems strange to think of all the individual acts I commit while driving home as the product of discrete decisions. "I *decided* to turn left at that light" connotes far more mental presence than there was. Perhaps I have taken the same route home from work so many times that the various turns, speed changes, signals, stops, and starts are performed by something like muscle memory and neither require nor prime any careful, conscious reflection on my part. And even when I encounter new variables during that drive home—a different set of surrounding cars, lights turning red that are typically green, pedestrians in areas different from before—responding to variables of those types has likely also become second nature to me—different but sufficiently similar to many driving home "flows" I have previously experienced.

Similarly, there are a variety of internal and external factors subtly interacting to stealthily influence my behaviors—behaviors that could be more or less morally valuable than others. My prior development of values, prior experiences with having to accept trade-offs among those values, deliberative skills, the skill to notice morally relevant features of a situation, the development of my moral imagination—these parts of my history contribute to a "current" that will influence how I behave. I cannot perform acts I cannot conceive; my moral conceptual capacities are thus one part of a current that excludes certain options when I act as an agent trying to respond to variables in a given scenario. I will fail to take into account what I am unlikely to notice; thus, similarities among my previous attention-directing flows are also a part of the current likely to limit the options available to me. Further, my physical capacities play a role. It likely will not occur to me to attempt to pull a lever that would divert a trolley, as in Switch, if I have routinely lacked the physical strength necessary to move such objects. I am also unlikely to behave in ways that require cognitive energy that I lack; thus, my sleeping and eating patterns are a part of the current that might either enable or limit my ability to listen attentively to a friend's grievances. Clearly, such internal factors will greatly influence how I behave in non-voting-booth-like scenarios: the degree of ease with which I ignore a panhandler at a traffic stop, how big I smile when I make eye contact with my child who is performing on stage, the tone of voice with which I respond to my wife when she is explaining why she had a bad day, the amount and type of body language with which I convey that I am listening to someone who is talking to me, whether I freeze in situations of threat, and so on. Internally speaking, these behaviors are habituated or routinized—the result of rehearsals that have contributed to behavioral momentum.

A wealth of social science suggests that a great many external factors affect our ethical behaviors. Whether seminary students help someone in their path is affected by their sense of being in a hurry.[39] Whether someone will report the truth of what she sees is affected by whether the reports of other people near her differ from what she plainly sees for herself.[40] Whether a person will agree to be an organ donor is affected by whether becoming an organ donor is an opt-in or opt-out process.[41] The same is true for the amount of withholdings employees choose to divert into their retirement savings.[42] Whether someone will put money into an honor-system pay box is affected by whether a picture of someone's eyes is displayed near the box.[43] How a person evaluates the quality of a variety of objects can be greatly influenced by the order in which those objects are presented to the evaluator.[44] Whether students cheat on a test is affected by whether they take that test in front of a mirror.[45] Cheating is even affected by as little as telling the students "Don't be a cheater" as opposed to telling them "Don't cheat."[46] In other words, features of our environment perform a function similar to the current in the water-walking analogy—giving lazy-river-like assistance to some behaviors and against-the-current resistance to others, where the influence of those currents is likely unnoticed. In fact, many of these studies find that participants not only are unaware of these subtle influences but also even deny them.[47]

Other studies illustrate ways these internal and external factors interplay to sway our behaviors. Correll and colleagues found that while participants might take themselves to be trying to decide whether to shoot someone on the basis of whether that person is armed, the actual shots fired might be affected by the person's skin color and the degree to which participants associate "black" with "dangerous"—an association about which the participants likely lack introspective awareness.[48] Several studies find that whether we offer someone a job and the amount we offer them as an initial starting salary can be greatly affected by

[39] Darley and Batson, "'From Jerusalem to Jericho.'"

[40] Asch, "Studies of Independence and Conformity."

[41] Johnson and Goldstein, "Defaults and Donation Decisions."

[42] McKenzie, Liersch, and Finkelstein, "Recommendations Implicit in Policy Defaults."

[43] Ernest-Jones, Nettle, and Bateson, "Effects of Eye Images on Everyday Cooperative Behavior."

[44] Nisbett and Wilson, "Telling More Than We Can Know."

[45] Diener and Wallbom, "Effects of Self-Awareness on Antinormative Behavior."

[46] Bryan, Adams, and Monin, "When Cheating Would Make You a Cheater."

[47] Nisbett and Wilson, "Telling More Than We Can Know," 243–44.

[48] Correll et al., "The Police Officer's Dilemma."

their name and the degree to which we associate that name with a gender or ethnicity.[49]

To make the water-walking analogy more vivid, consider an anecdotal case. Author and martial artist Terry Dobson writes of an experience he had on a Tokyo train while studying martial arts in Japan.[50] A heavily intoxicated man entered a commuter train and began harassing and attempting to assault commuters. Dobson had been studying aikido eight hours a day for three years, and he was confident he could physically subdue the drunk. Aikido heavily discourages its students from engaging in physical conflicts; its philosophy is one of avoiding such conflicts and avoiding harm to opponents even during physical conflicts. But Dobson saw this as a clear case where violence was called for. He taunted the drunk to direct the drunk's attention away from other passengers. The drunk took the bait and started heading for Dobson. Before they met, an elderly gentleman seated on the train shouted, "Please come talk to me!" The drunk redirected his attention and said a few rude things to the old man. The old man persisted, eventually asking about the drunk's family. The drunk sank into a seat and began sobbing. He told the old man that his wife had died and that he was now unemployed and homeless. The old man listened and empathized. When Dobson got off the train, the two were still chatting. Dobson writes, "What I had wanted to do with muscle and meanness had been accomplished with a few kind words."

To analyze this case using the myth of the voting booth would be a mistake. Dobson did not experience the scenario as a voting booth in which he faced the scenario with a sense of static enclosure and carefully surveyed a variety of options, that is, one in which he noticed one of his clear options was to try to speak and empathize with the drunk but after careful deliberation among all his options "voted" against that option. Rather, Dobson's years of martial-arts training caused him to perceive the drunk as an aikido opponent. In the heat of a threatening situation, the current of Dobson's previous physical conflict rehearsals kicked in. In reaction to an environmental cue, he moved with a current the buildup of which he had been contributing to for three years. Yet another agent present in the scenario, the old man, had no such training and thus reacted from a different behavioral flow. Upon witnessing the old man's response, Dobson realized his training had precluded a morally superior response to the same environmental cue; he lacked a history of momentums

[49] See Steinpreis, Anders, and Ritzke, "The Impact of Gender on the Review of the Curricula Vitae of Job Applicants and Tenure Candidates"; Moss-Racusin et al., "Science Faculty's Subtle Gender Biases Favor Male Students"; and Carlsson and Rooth, "Evidence of Ethnic Discrimination in the Swedish Labor Market."

[50] Dobson, "A Soft Answer," 119.

that would have lazy-river-assisted him into easily reacting the way the old man had. Again, this case is simply unlike the myth of the voting booth and therefore unlike the conception of ethical decision-making common to trolley problems.

Our current best evidence suggests that most of our behavior is not calculated or deliberative but conditioned over time to become increasingly automated; it is not performed in an influence vacuum but facilitated, cued, and corralled. Thus, the ethical life is far more like the water-walking analogy than the myth of the voting booth. That does not make most human behavior any less moral or any less the legitimate subject of ethical scrutiny. But if the trolley method employs a conception of ethical decision-making unlike most ethical behavior, then in the case of most moral behavior it is not the right tool for the job. By presenting cases of decision-making exclusively like the myth of the voting booth, the trolley method does one of two things. Either it blatantly excludes the majority of ethical decisions from consideration (once again, backgrounding most of moral life), or it fails to represent the majority of ethical decisions, either by mischaracterizing them or by giving undue credit to the folk view of the degree to which our decisions and actions are performed uninfluenced and with self-awareness.

## 4. CONCLUSION

While I have not encountered the exact objections I have raised here elsewhere in the literature, I do not think what I am arguing is terribly original. Held warned against general domains of ethical investigation—domains such as the courtroom or the marketplace—leading to the development of tools that are not helpful for ethical decisions or practices outside those domains—such as the family and home life.[51] Confucius took the time to address what seem, in comparison to trolley problems, very mundane issues of customary practice and social interaction.[52] Aristotle warned against seeking undue precision in ethics.[53] And in criticizing the trolley method–esque conception of ethical decision-making, I draw on broader criticisms raised by virtue theorists and situationists alike that major moral theories such as utilitarianism and Kantianism miss important and basic parts of the ethical life. Pulling these resources together to shine a light on the trolley method reveals a practice that jarringly fails to include most of the lived moral experience of most moral agents.

[51] Held, *The Ethics of Care*, 24–25.

[52] *Analects*, 9.3.

[53] Aristotle, *Nicomachean Ethics*, bk. 1, ch. 3.

Despite these shortcomings, I am not arguing that trolley problems should be altogether discarded. I do think they are interesting tools that tell us something about moral psychology. For example, work involving participant responses to trolley problems has revealed interesting things about the role of emotion in moral judgment.[54] Further, case sets such as Bystander, Grandfather, and Marine's Wife seem informative about how some people cognitively process the content of case descriptions. And, like Martena, I think the deconstruction of trolley problems is a good educational use of trolley problems for ethics students.[55] But I disagree with Martena that trolley problems are poor educational tools when used to illustrate the differences between major ethical theories such as utilitarianism and Kantianism. While Martena is right to point out that the "solutions" to trolley problems can present an oversimplified view of such theories, I find that for some students, having that oversimplified, cartoonish understanding is an entry point; for some students, the oversimplified version is already a conceptual challenge. (I suspect that, due to, perhaps, the curse of knowledge, we philosophy teachers too often forget how difficult philosophy can be for many of the uninitiated.) Once students grasp the oversimplified version, they are then in a position to see and appreciate the oversimplifications—eventually developing a more nuanced view of such theories, a ladder that some students need to climb in order to understand the reasons why, once at the top, they should kick it away. This might also describe their worthwhile use for philosophers—a blunt instrument that in some instances can help us as long as we remain aware of its bluntness and do not expect it to do the work of sharper tools.

*Rose State College*
*gcrain@rose.edu*

## REFERENCES

Aristotle. *Nicomachean Ethics*. 2nd ed. Translated by Terence Irwin. Indianapolis: Hackett Publishing Company, 1999.

Arnett, Jeffrey J. "The Neglected 95%: Why American Psychology Needs to Become Less American." *American Psychologist* 63, no. 7 (October 2008): 602–14.

[54] See Haidt, "The Emotional Dog and Its Rational Tail"; and Greene, "An fMRI Investigation of Emotional Engagement in Moral Judgment."

[55] Martena, "Thinking inside the Box."

Asch, Solomon E. "Studies of Independence and Conformity: I. A Minority of One against a Unanimous Majority." *Psychological Monographs: General and Applied* 70, no. 9 (1956): 1–70.

Awad, Edmond, Sohan Dsouza, Azim Shariff, Iyad Rahwan, and Jean-François Bonnefon. "Universals and Variations in Moral Decisions Made in 42 Countries by 70,000 Participants." *Proceedings of the National Academy of Sciences* 117, no. 5 (February 2020): 2332–37.

Bargh, John A. "The Automaticity of Everyday Life." In *The Automaticity of Everyday Life*, edited by Robert S. Wyer, 1–61. New York: Psychology Press, 1997.

Barlow, Rich. "BU Research: A Riddle Reveals Depth of Gender Bias." *BU Today*, January 16, 2014. http://www.bu.edu/today/2014/bu-research-riddle-reveals-the-depth-of-gender-bias/.

Barney, Rachel. "[Aristotle], *On Trolling*." *Journal of the American Philosophical Association* 2, no. 2 (Summer 2016): 193–95.

Bauman, Christopher W., A. Peter McGraw, Daniel M. Bartels, and Caleb Warren. "Revisiting External Validity: Concerns about Trolley Problems and Other Sacrificial Dilemmas in Moral Psychology." *Social and Personality Psychology Compass* 8, no. 9 (September 2014): 536–54.

Bonevac, Daniel A. *Today's Moral Issues: Classic and Contemporary Perspectives*. 7th ed. New York: McGraw-Hill, 2013.

Boss, Judith A. *Analyzing Moral Issues*. 6th ed. New York: McGraw-Hill, 2013.

———. *Ethics for Life*. 6th ed. New York: McGraw Hill, 2014.

Bostyn, Dries H., Sybren Sevenhant, and Arne Roets. "Of Mice, Men, and Trolleys: Hypothetical Judgment versus Real-Life Behavior in Trolley-Style Moral Dilemmas." *Psychological Science* 29, no. 7 (July 2018): 1084–93.

Bryan, Christopher J., Gabrielle S. Adams, and Benoît Monin. "When Cheating Would Make You a Cheater: Implicating the Self Prevents Unethical Behavior." *Journal of Experimental Psychology: General* 142, no. 4 (November 2013): 1001–5.

Cahn, Steven M. *Exploring Philosophy of Religion: An Introductory Anthology*. 4th ed. New York: Oxford University Press, 2017.

Carlsson, Magnus, and Dan-Olof Rooth. "Evidence of Ethnic Discrimination in the Swedish Labor Market Using Experimental Data." *Labour Economics* 14, no. 4 (August 2007): 716–29.

Centers for Disease Control and Prevention. "CDC's Abortion Surveillance System FAQs." November 17, 2022. https://www.cdc.gov/reproductive-health/data_stats/abortion.htm.

Confucius. *The Analects*. Translated by D. C. Lau. London: Penguin Books, 1979.

Correll, Joshua, Sean M. Hudson, Steffanie Guillermo, and Debbie S. Ma.

"The Police Officer's Dilemma: A Decade of Research on Racial Bias in the Decision to Shoot." *Social and Personality Psychology Compass* 8, no. 5 (May 2014): 201–13.

Darley, John M., and C. Daniel Batson. "'From Jerusalem to Jericho': A Study of Situational and Dispositional Variables in Helping Behavior." *Journal of Personality and Social Psychology* 27, no. 1 (October 1973): 100–8.

Diener, Edward, and Mark Wallbom. "Effects of Self-Awareness on Antinormative Behavior." *Journal of Research in Personality* 10, no. 1 (March 1976): 107–11.

Dobson, Terry. "A Soft Answer." *Reader's Digest*, December 1981.

Edmonds, David. *Would You Kill the Fat Man? The Trolley Problem and What Your Answer Tells Us about Right and Wrong*. Princeton, NJ: Princeton University Press, 2015.

Ernest-Jones, Max, Daniel Nettle, and Melissa Bateson. "Effects of Eye Images on Everyday Cooperative Behavior: A Field Experiment." *Evolution and Human Behavior* 32, no. 3 (May 2011): 172–78.

Fried, Barbara H. "What Does Matter? The Case for Killing the Trolley Problem (or Letting It Die)." *The Philosophical Quarterly* 62, no. 248 (July 2012): 505–29.

Gil, Inma. "The Fifth Floor—Riddle Me This." *BBC World Service*, March 29, 2018. Video, 2:45. https://www.bbc.co.uk/programmes/p062qcoc.

Greene, Joshua D., R. Brian Sommerville, Leigh E. Nystrom, John M. Darley, and Jonathan D. Cohen. "An fMRI Investigation of Emotional Engagement in Moral Judgment." *Science* 293, no. 5537 (September 2001): 2105–8.

Haidt, Jonathan. "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment." *Psychological Review* 108, no. 4 (October 2001): 814–34.

Held, Virginia. *The Ethics of Care: Personal, Political, and Global*. Oxford: Oxford University Press, 2006.

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. "Most People Are Not WEIRD." *Nature* 466, no. 7302 (July 2010): 29.

———. "The Weirdest People in the World?" *Behavioral and Brain Sciences* 33, nos. 2–3 (June 2010): 61–83.

Hinman, Lawrence M. *Contemporary Moral Issues: Diversity and Consensus*. 3rd ed. Upper Saddle River, NJ: Pearson Education, 2006.

JafariNaimi, Nassim. "Our Bodies in the Trolley's Path, or Why Self-Driving Cars Must *Not* Be Programmed to Kill." *Science, Technology, and Human Values* 43, no. 2 (March 2018): 302–23.

Johnson, Eric J., and Daniel G. Goldstein. "Defaults and Donation Decisions." *Transplantation* 78, no. 12 (December 2004): 1713–16.

Kneer, Markus, and Ivar Rodríguez Hannikainen. "Trolleys, Triage and COVID-19: The Role of Psychological Realism in Sacrificial Dilemmas." *Cognition and Emotion* 36, no. 1 (2021): 137–53.

Knobe, Joshua. "Philosophical Intuitions Are Surprisingly Robust across Demographic Differences." *Epistemology and Philosophy of Science* 56, no. 2 (2019): 29–36.

Lederach, John Paul. *The Moral Imagination: The Art and Soul of Building Peace*. New York: Oxford University Press, 2005.

Lieberman, Jethro K. *Fight the Hypo: Fake Arguments, Trolleyology and the Limits of Hypotheticals.* New York: Tribeca Square Press, 2014.

Lin, Patrick. "Robot Cars and Fake Ethical Dilemmas." *Forbes Magazine*, April 3, 2017. https://www.forbes.com/sites/patricklin/2017/04/03/robot-cars-and-fake-ethical-dilemmas/.

MacKinnon, Barbara. *Ethics: Theory and Contemporary Issues*. 5th ed. Belmont, CA: Wadsworth Cengage Learning, 2007.

Martena, Laura. "Thinking inside the Box: Concerns about Trolley Problems in the Ethics Classroom." *Teaching Philosophy* 41, no. 4 (December 2018): 381–406.

McKenzie, Craig R. M., Michael J. Liersch, and Stacey R. Finkelstein. "Recommendations Implicit in Policy Defaults." *Psychological Science* 17, no. 5 (May 2006): 414–20.

Morioka, Masahiro. "The Trolley Problem and the Dropping of Atomic Bombs." *Journal of Philosophy of Life* 7, no. 2 (August 2017): 316–37.

Moseley, Daniel. "Revisiting Williams on Integrity." *Journal of Value Inquiry* 48, no. 1 (2014): 53–68.

Moss-Racusin, Corinne A., John F. Dovidio, Victoria L. Brescoll, Mark J. Graham, and Jo Handelsman. "Science Faculty's Subtle Gender Biases Favor Male Students." *Proceedings of the National Academy of Sciences* 109, no. 41 (October 2012): 16474–79.

Nisbett, Richard E., and Timothy D. Wilson. "Telling More than We Can Know: Verbal Reports on Mental Processes." *Psychological Review* 84, no. 3 (May 1977): 231–59.

O'Connor, James. "The Trolley Method of Moral Philosophy." *Essays in Philosophy* 13, no. 1 (January 2012): 243–56.

Olberding, Amy. *The Wrong of Rudeness: Learning Modern Civility from Ancient Chinese Philosophy*. New York: Oxford University Press, 2019.

O'Neill, Onora. *Towards Justice and Virtue*. Cambridge: Cambridge University Press, 1996.

Rachels, James, ed. *Moral Problems: A Collection of Philosophical Essays*. New York: Harper and Row, 1971.

Reilly, Katie. "10 of Mother Teresa's Most Powerful Quotes." *Time*, September 3, 2016. https://time.com/4478287/mother-teresa-saint-quotes/.

Rennix, Brianna, and Nathan J. Robinson. "The Trolley Problem Will Tell You Nothing Useful about Morality." *Current Affairs*, November 3, 2017. https://www.currentaffairs.org/2017/11/the-trolley-problem-will-tell-you-nothing-useful-about-morality.

Roache, Rebecca. "Honestly, It's Fine!" *Aeon*, December 1, 2016. https://aeon.co/essays/what-is-worse-a-passive-aggressive-silence-or-an-outright-shout.

———. "Naughty Words." *Aeon*, February 22, 2016. https://aeon.co/essays/where-does-swearing-get-its-power-and-how-should-we-use-it.

Roff, Heather M. *The Folly of Trolleys: Ethical Challenges and Autonomous Vehicles*. Brookings Institution, December 17, 2018. https://www.brookings.edu/research/the-folly-of-trolleys-ethical-challenges-and-autonomous-vehicles/.

Shafer-Landau, Russ. *The Ethical Life: Fundamental Readings in Ethics and Moral Problems*. New York: Oxford University Press, 2015.

Soifer, Eldon. *Ethical Issues: Perspectives for Canadians*. 2nd ed. Peterborough, ON: Broadview Press, 1997.

Steinpreis, Rhea E, Katie A Anders, and Dawn Ritzke. "The Impact of Gender on the Review of the Curricula Vitae of Job Applicants and Tenure Candidates: A National Empirical Study." *Sex Roles* 41 (October 1999): 509–28.

Stich, Stephen P., and Edouard Machery. "Demographic Differences in Philosophical Intuition: A Reply to Joshua Knobe." *Review of Philosophy and Psychology* 14, no. 2 (June 2023): 401–34.

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

Weinberg, Justin. "Philosophical Intuitions and Demographic Differences." *Daily Nous* (blog), December 2, 2019. https://dailynous.com/2019/12/02/philosophical-intuitions-demographic-differences/.

White, James E. *Contemporary Moral Problems*. 5th ed. St. Paul, MN: West Publishing Company, 1997.

Williams, Bernard. "A Critique of Utilitarianism." In *Utilitarianism: For and Against*, by J. J. C. Smart and Bernard Williams. Cambridge: Cambridge University Press, 2008.

Wilson, James. "Internal and External Validity in Thought Experiments." *Proceedings of the Aristotelian Society* 116, no. 2 (July 2016): 127–52.