JOURNAL *of* ETHICS & SOCIAL PHILOSOPHY

VOLUME XXIII · NUMBER 2
*November 2022*

ARTICLES



DISCUSSIONS

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

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

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

*Journal of Ethics and Social Philosophy*
VOL. 23, NO. 2 · NOVEMBER 2022
*https://doi.org/10.26556/jesp.v23i2.1797*


# THE RENT TAX IS TOO DAMN LOW

## JUSTICE, PRODUCTIVITY, AND THE TAX BASE

*Matthew T. Jeffers*

TAXATION is the means by which society finances its public initiatives. It is the area of public policy where the topic of distributive justice plays its largest role, for it determines who will bear the burdens of paying for society's collective aims. The current economic and public policy literature poses the tax debate as one involving trade-offs between economic considerations and moral values. In particular, optimal tax theorists are concerned with optimizing across competing considerations of equity and efficiency.[1] Such an approach assumes an inherent antagonism between these two normative aims. A similar antagonism also persists in the philosophical literature, where taxation itself is often perceived as being at odds with principles of distributive justice, especially the principle of labor ownership.[2]

My contention is that these antagonisms are not inherent to all forms of taxation, but rather they exist because tax theorists and policy makers tend to focus on a narrow range of options for the tax base. The tax base, both in theory and practice, is traditionally confined to certain kinds of economic gains, such as earned income, gains from capital, and consumption. When economists and philosophers evaluate taxes, they tend to do so in the context of thinking about this prevailing tax base.[3] I begin by discussing the two major drawbacks of the prevailing tax base: (1) the equity-efficiency tradeoff and (2) the diminishment of labor ownership. I then diagnose the common cause responsible for these moral and economic drawbacks. Next, I discuss an alternative tax base that

1 This has been true since Mirrlees's foundational paper. See Mirrlees, "An Exploration in the Theory of Optimum Income Taxation."

2 Such as seen in Locke, *Second Treatise of Government*, 12–18; George, *Progress and Poverty*, 122–24; Nozick, *Anarchy State and Utopia*, 175; Dworkin, "What Is Equality?"; Vallentyne, "Self-Ownership and Equality"; and Otsuka, "Self-Ownership and Equality." A more moderate expression of that antagonism can be found in Russell, "Self-Ownership, Labor, and Licensing," 194–95; and Brennan, "Striving for the Middle Ground," 5–14.

3 See Halliday "Justice and Taxation," 1114–16.

avoids this common cause and consequently does not possess these two major drawbacks. I conclude by offering some illustrative, though tentative, examples of the alternative tax base, knowing that further investigation is necessary.

This is not a comprehensive tax-and-spend proposal. I make no claims regarding the scope of government, the size of spending, or tax rates and schedules. My concern here is only on *what kinds of things* we tax, not the extent to which we should tax them.[4] Further, the proposal herein is general, not absolute. The proposed tax base may not be suitable in some circumstances. I only argue that, as a general matter, it will likely be better to replace the prevailing tax base with a proposed alternative tax base.

Before proceeding, it is worth noting that the tax base I endorse would in some sense be more familiar to the classical political economists and philosophers of the early modern period than to contemporary theorists. Thinkers of this period, notably Adam Smith, David Ricardo, and John Stuart Mill, often distinguished economic activities on the basis of their origin and normative legitimacy.[5] By contrast, the tendency in modern economics is to treat all economic gains the same, with the implicit assumption being "if there is money being paid, there is value being generated." Throughout this piece I will challenge this assumption and make distinctions between economic gains on the basis of wealth generation or lack thereof. One way of viewing this project is as a modernization and evolution of the classical conception of rental taxation in a way that is consistent with the standards of contemporary economic and philosophical thought.[6]

## 1. TWO OBJECTIONABLE FEATURES OF THE PREVAILING TAX-BASE REGIME

Countries in the developed world have tax-policy differences; however, their tax-base regimes are more alike than they are different. In most developed countries the majority of the tax burden falls on workers in the form of taxes on earned income such as wages or salaries, and also in the form of payroll taxes

4 Some scholars ask whether it "make[s] sense to evaluate the tax system independently of what the tax revenue is used for" (Brennan and Tsai, "Tax Ethics," 399). I am agnostic about such a claim, except to note that if it is permissible to raise general funds for some legitimate governmental purpose, then this paper is an investigation into the moral considerations of what the *sources* of such funds ought to look like.

5 See Smith, *The Wealth of Nations*, 1130–32; Ricardo, *Principles of Political Economy and Taxation*, 167–69; and for an overview of Mill's views on taxation see Halliday, *Inheritance of Wealth*, 47–57.

6 The contemporary assumptions I am referring to include using (i) standard marginalist analysis, (ii) a subjective (utility) theory of value, (iii) an emphasis on broadly egalitarian desiderata and, (iv) an updating of the concept of labor ownership.

or other social insurance levies.[7] The second-largest source of tax revenue in most developed countries is consumption taxes, either in the form of sales or value-added taxes.[8] The burden of these taxes falls predominantly on consumers. Typically, the third-largest source of tax revenue by country is corporate income and capital gains taxes. The burden of this tax is often thought to be shared among consumers, workers, and firms.[9] In most developed countries these three tax sources constitute the vast majority of tax revenue, with other sources of tax revenue being meager by comparison. These taxes pose serious drawbacks, the first of which is the equity-efficiency tradeoff.

### *1.1. The Equity-Efficiency Tradeoff*

The approach that dominates the economic literature on optimal taxation characterizes the primary tax problem as a trade-off between economic efficiency and distributive equality.[10] Recent practitioners are now using a broader array of normative considerations, but the central focus of optimal tax theory remains balancing equity and efficiency.[11] To solve this problem optimal tax theorists typically employ a social welfare optimization approach, where the aim is "keeping tax distortions to a minimum, subject to restrictions by the need to raise revenue and maintain an equitable tax burden."[12] I first describe the efficiency costs of taxation and then discuss how trying to minimize these costs comes at the expense of equity. I then highlight how optimal tax theorists have traditionally sought to curtail the effect of this equity-efficiency trade-off and then gesture toward an alternative way of avoiding the trade-off altogether.

All taxes have income effects, that is, they decrease the amount of money that the taxed individual has available to them. However, not all taxes are *distortionary*. The fundamental problem with most forms of taxation is that they artificially change the relative price of a given bundle of economic offerings (goods, services, labor, etc.) from its market baseline price. Under standard assumptions, the market baseline price is efficient because in a competitive

7 Enache "Sources of Government Revenue in the OECD."

8 Enache "Sources of Government Revenue in the OECD."

9 Arulampalam, Devereux, and Maffini, "The Direct Incidence of Corporate Income Tax on Wages."

10 Mirrlees, "An Exploration in the Theory of Optimum Income Taxation," 175; and Mankiw, Weinzierl, and Yagan, "Optimal Taxation in Theory and Practice," 5.

11 Notable papers that broaden the list of normative considerations include Saez and Stantcheva, "Generalized Social Marginal Welfare Weights for Optimal Tax Theory," 25; and Fleurbaey and Maniquet, "Optimal Income Taxation Theory and Principles of Fairness."

12 Auerbach and Hines, "Taxation, and Economic Efficiency," 1347.

environment the price of a given bundle approximates its marginal cost.[13] Taxes change the relative prices between various bundles and consequently agents choose new bundles that are inferior to the bundles that they would have chosen absent the tax.[14] To illustrate, imagine that, in some competitive market, the price of an apple is $1.00. If a $0.50 tax is levied on each apple, then those who would have bought an apple between $1.00 and $1.50 no longer make that purchase and those exchanges are foregone. Those consumers will instead use their $1.00 to purchase goods or services that are less valuable to them than the apple would have been. Economists call this a *distortion*, since the individuals are choosing less preferable alternative consumption bundles than they would choose absent the tax. The difference in surplus between the optimal bundle and the inferior bundle is the *deadweight loss* and represents the value that society loses as a result of the tax.[15] Taxes on consumption, capital, and earned income all distort people's economic decisions and incur deadweight loss.

The other major concern that looms large for tax theorists is equality. Optimal tax theorists are not preoccupied with ensuring that distributions are equal, but they do assume that more equal distributions of wealth are better than less equal distributions, subject to efficiency concerns. Egalitarian desiderata are typically justified with reference to diminishing marginal utility or Rawlsian concerns over the welfare of the least advantaged.[16] However, one does not need to be a Rawlsian to see how it would be a pyrrhic victory for tax justice if increasing the total size of the economic pie comes at the expense of diminishing the size of the slices available to most people. Most arguments against egalitarianism do not challenge the principle that more equal distributions are favorable to less equal distributions, *ceteris paribus*; rather, they challenge giving priority to egalitarian aims *at the expense* of other important criteria, such as efficiency, fairness, or desert. As I hope will become clear by the end of this paper, equality and these other desiderata tend to run together, if only we would select the correct tax base.

Since Mirrlees's 1971 paper, the fundamental problem of optimal tax theory has been the equity-efficiency tradeoff.[17] The basic notion is that marginal tax

13 Mankiw, Weinzierl, and Yagan, "Optimal Taxation in Theory and Practice," 4.

14 Slemrod, "Optimal Taxation and Optimal Tax Systems," 159.

15 Feldstein, "The Effect of Taxes on Efficiency and Growth," 4.

16 For the Rawlsian influence see Fleurbaey and Maniquet, "Optimal Income Taxation Theory and Principles of Fairness," as well as Stiglitz, "Pareto Efficient and Optimal Taxation and the New Welfare Economics." For diminishing marginal utility see Mirrlees, "An Exploration in the Theory of Optimum Income Taxation."

17 See Mirrlees, "An Exploration in the Theory of Optimum Income Taxation"; and Saez, "Using Elasticities to Derive Optimal Income Tax Rates," 205.

rates distort a person's production or consumption behavior and the size of the distortion increases as the rate of income or consumption rises.[18] To see why, compare a tax on a person's first earned dollar to a tax on a person's hundred-thousandth earned dollar. The tax on the first dollar earned is unlikely to disincentivize someone from working. This is because, at lower income levels, the income effect is much more operative than the substitution effect. People are less willing to substitute leisure for work when their basic needs and interests are not satisfied. By contrast, at higher levels of income, the substitution effect plays a larger role. People are more willing to substitute work for leisure when giving up work only comes at the expense of fewer luxury goods. The consequence is that taxes at lower levels of income result in less distortions than taxes at higher levels of income. The same reasoning applies to consumption. Hence, a tax on a person's first earned dollar has much less deadweight loss than a tax on their hundred-thousandth earned dollar. This is the equity-efficiency trade-off.

The paramount problem for optimal tax theory is how to minimize the equity-efficiency tradeoff. Imagine that making the economic pie larger also shrinks the size of the median pie slice. Optimal tax theorists are trying to keep the economic pie from shrinking while also making the size of the median pie slice as large as possible. Most inefficiency occurs when a tax discourages work or consumption at the margin. To solve this problem, tax theorists have tried to optimize tax rate schedules within the prevailing tax regime by trying to determine the shape of the ability distribution.[19] Distinguishing between high and low income ability means that it becomes possible to set tax rates such that "few individuals would be affected at the margin and many would be affected inframarginally."[20] The idea is to structure tax rates such that they are greatest below the earnings potential of high-ability individuals while also remaining *above* the earnings potential of low-ability individuals.[21] The motivation behind determining the distribution of ability is to extract tax revenue through the most inelastic portion of a person's earnings schedule.

Consistent with the example described above, the general strategy for optimal tax theorists has been to remain within the prevailing tax regime and try to either (a) exploit commodity inelasticities, or (b) exploit income inelasticities.[22] One problem with this approach is informational. It is extremely difficult

18 Stiglitz, "The Origins of Inequality and Policies to Contain It," 583–86.

19 Mankiw, Weinzierl, and Yagan, "Optimal Taxation in Theory and Practice," 6.

20 Mankiw, Weinzierl, and Yagan, "Optimal Taxation in Theory and Practice," 6.

21 Brewer, Saez, and Shephard, "Means-Testing and Tax Rates on Earnings," 91

22 See Saez, "The Desirability of Commodity Taxation under Non-linear Income Taxation and Heterogeneous Tastes," 228, and "Using Elasticities to Derive Optimal Income Tax Rates."

to determine where the inelasticities are, especially given the enormous heterogeneity of preferences within a population of workers or consumers.[23] Another problem with this approach is that it tries to optimize *given* the trade-off, instead of trying to find a way *out* of the trade-off altogether.

An alternative approach is to search for a tax base where the equity-efficiency trade-off is not present, or if it is sometimes present it is at least greatly diminished.[24] One way of doing this may involve shifting our thinking about *who* or *what* we tax. For example, the prevailing tax-base regime centers taxation on the *economic agent*, i.e., the person responsible for making production and consumption decisions. The fundamental challenge with taxing economic agents is that they are *agents*—they make decisions in response to changes in the environment. The source of economic distortion is the fact that taxes alter trade-offs and so the agent will *substitute away* from the optimal bundle to a less optimal bundle. But what if instead of taxing economic agents, we sought other forms of taxation that did not alter the trade-offs of economic agents?

What if instead of focusing on a tax base consisting of economic agents, we searched for a tax base that consisted more of economic *patients*. To borrow the agent-patient distinction from moral philosophy, an *economic patient* is not themselves an economic decision maker in a particular context, but is someone who nevertheless is affected by economic choices or circumstances.[25] Taxing an economic agent changes their relevant trade-offs, hence introducing the possibility of economic distortion. By contrast, taxing an economic patient *cannot* change their relevant trade-offs, because in the given context, they are not making any economic decisions. Since the equity-efficiency tradeoff is a consequence of a tax base replete with economic distortions, it may be a worthwhile strategy to search for a tax base less susceptible to distortions, or at least less susceptible to *bad distortions*. As a brief prelude to my general strategy, I recommend searching for a tax base that consists of either (a) economic agents whose decisions we *want* to distort or (b) economic patients who have no relevant decisions to distort. I will elucidate this strategy in sections 3 and 4.

### *1.2. Diminishing Labor Ownership*

The concept of labor ownership is firmly embedded within the prevailing economic and social structure. While the idea has ancient roots, it was first introduced

23 See Dahan and Strawczynski, "Optimal Income Taxation"; and Sandmo, "Optimal Redistribution when Tastes Differ."

24 "It is indeed worth emphasizing that ethical principles may be relevant not only to the design of the income tax, but also to the selection of the tax base" (Fleurbaey and Maniquet, "Optimal Income Taxation Theory and Principles of Fairness," 1032).

25 For the original agent-patient distinction, see McPherson, "The Moral Patient."

in the modern era by Locke's famous labor-mixing argument.[26] However, Locke's argument was the formal articulation of an already existing norm rather than the generation of an entirely new idea. The principle of labor ownership can be expressed in myriad ways, but essentially the idea is that *the person (or people) who create or produce a thing have some bundle of property rights in that thing that others do not possess.* This typically includes the right to use, gift, trade, sell, and, more controversially, destroy, their creation. The "thing" created may be the performance of certain activities, e.g., labor, or a product, service, or even string of words or set of ideas. The principle of labor ownership is not identical to the concept of property, but serves as the normative basis for many property claims in that it is a method of determining who has original title to property.

Many thinkers have sought to justify the principle of labor ownership in various ways. The preeminent economic justification is that the principle of labor ownership gives people enormous incentives to engage in productive activity, which makes both themselves and society wealthier.[27] Moral justifications for labor ownership have been more varied. One common justification is derived from individual sovereignty. An early example is found in Henry George: "Is it not primarily the right of a man to himself to the use of his own powers to the enjoyment of the fruits of his own exertions?… As a man belongs to himself, so his labor when put in concrete form belongs to him."[28] According to this argument, labor ownership is a logical extension of the sovereignty that one has over one's own body and activities. Modern variations of the individual sovereignty argument for labor ownership remain popular with prominent theorists, including Vallentyne, Otsuka, and Steiner.[29] Another recent conception of labor ownership deploys this familiar concept of self-sovereignty but constrains it within the limits of social reciprocity.[30] Daniel Russell contends that "property in one's labor … [should be] understood as a social institution for balancing two freedoms: freedom to act even if it interferes with someone else, and freedom from interference."[31] Under this conception, labor ownership ought to be protected on the condition that such protections create "reciprocal benefits" within the community.[32]

26 Locke, *Second Treatise of Government*, 12–18.

27 See Demsetz, "Toward a Theory of Property Rights," 348.

28 George, *Progress and Poverty*, 122.

29 See Vallentyne, "Self-Ownership and Equality"; Otsuka, "Self-Ownership and Equality"; Steiner, "Left Libertarianism and the Ownership of Natural Resources"; and Delmotte and Verplaetse, "What Is Wrong with Endowment Taxation?"

30 See Russell, "Self-Ownership, Labor, and Licensing."

31 Russell, "Self-Ownership, Labor, and Licensing," 174.

32 Russell, "Self-Ownership, Labor, and Licensing," 187.

Other moral arguments for labor ownership are derived from differential sacrifice between persons. For instance, Dworkin argues that a version of labor ownership can be justified by morally relevant differences in choice under certain conditions of equality:

> If he earns enough by working hard, or by working at work that no one else wants to do, to satisfy all his expensive tastes, then his choice for his own life costs the rest of the community no more than if his tastes were simpler and industry less. . . . The choice should be indifferent under equality of resources, so long as no one envies the total package of work plus consumption that he chooses. So long as no one envies, that is, his life as a whole.[33]

Dworkin is not alone in thinking that differential sacrifice in the form of hard work or other features can confer labor ownership. In public political culture, "hard work," "risk," and "sacrifice" are commonly offered reasons for justifying moral claims to one's property or position of differential advantage. Regardless of whether or not any of these arguments succeed, labor ownership is currently a widely accepted foundational economic norm that people follow, that market participants presume, and that courts uphold. This alone does not justify the practice, but makes the principle a plausible starting point.

Supposing temporarily that we accept the principle of labor ownership, what then is the implication for taxation? At first glance, it appears that labor ownership is completely at odds with any form of taxation assessed against returns on human labor, that is, *earned income*, which includes wages, salaries, bonuses, commissions, and profits. Proponents of this view, most notably Nozick, radically suggests that taxation on earned income is "on a par with forced labor."[34] Similarly, the political slogan "taxation is theft" reflects this absolutist conception of labor ownership, where any deviation amounts to a rights violation.[35] Yet, there are serious difficulties with this absolutist approach to labor ownership.[36] Commentators as diverse as Buchanan, Murphy, and Nagel have argued that upholding labor ownership claims in the first place likely requires a system of public infrastructure and hence some system of

33 Dworkin, "What Is Equality?" 306.

34 Nozick, *Anarchy State and Utopia*, 169.

35 See Kearl, "Do Entitlements Imply that Taxation Is Theft?"; and Mack, "Non-absolute Rights and Libertarian Taxation."

36 Although Nozick's conception of labor ownership is derived from Locke's, it is construed in absolute terms, whereas Locke's conception was conditional on obligations of assistance embedded in an original natural-law understanding of the provisos. See Lamb, *Property*, 58–62.

tax collection.[37] Consequently, it becomes problematic to claim that one has absolute entitlement to pretax earnings on the basis of labor ownership, since those pretax earnings require the joint production of the wider community.[38] This entanglement of productive causes has led some skeptics of labor ownership to discard the concept altogether, notably Rawls.[39] Subsequent theorists, such as Murphy, Nagel, and Lindsay, argue that property claims can only be upheld when they are the consequence of a legitimate social and legal structure. For example, Lindsay contends that "ownership is a social fact, and as such derives its legitimacy from the extent to which people living under it give it their uncoerced consent."[40] The upshot of these legal constructivist accounts is that there are no "pre-political" claims to property on the basis of labor ownership; hence pretax earnings are an illegitimate benchmark and have no moral relevance to the topic of tax justice whatsoever.[41] It is worth noting that some scholars have pointed out that the Rawlsian critique of a joint product does not necessarily entail a conventionalist view of pretax earnings or other property claims.[42] Regardless, the core of the dispute raised by Nagel and Murphy remains. Do labor ownership claims on pretax earnings have any independent normative significance given that they are reliant on the preexisting legal and social system?

I want to challenge the notion that the moral relevance of labor ownership claims on pretax earnings is predicated on being causally independent of the social-legal structure. There can be strong reasons for endorsing the independent normative significance of labor ownership claims, *even if* the existence of those claims is dependent on the social and legal structure. To see how, let us consider an alternative way of thinking about labor ownership and its connection to tax justice by considering a concept I call *effective control*.

People have more effective control over their lives the closer their own decisions are tightly connected to the outcomes they experience. The less that their

37 Buchanan, "The Ethical Limits of Taxation"; Murphy and Nagel, *The Myth of Ownership*.

38 Controversy over this idea played out in public political culture over the interpretation of President Obama's "you didn't build that" comment. See Blake, "Obama's 'You Didn't Build That' Problem."

39 See Rawls, *A Theory of Justice*, 10–11, 240–42.

40 Lindsay, "Ownership by Agreement," 935.

41 See Murphy and Nagel, *The Myth of Ownership*, 74.

42 Notably, Geoffrey Brennan argues that the moral relevance of pretax earnings is consistent with a Rawlsian constitutional approach, and Jorgen Pedersen argues that a Rawlsian commitment to independence and self-respect entails a thin conception of private property. See Brennan, "Striving for the Middle Ground," ch. 3; and Pedersen, *Distributive Justice and Taxation*, 151–52.

decisions are connected to those outcomes, the less effective control they possess. To illustrate, consider the following. A 1 percent tax on a person's earned income is likely to have little effect on a person's ability to control their own lives; for most people it is unnoticeable. By contrast, a fully enforced 100 percent tax on earned income will completely deprive someone of their ability to control their economic situation; without assistance, they will die. The upshot is that as the tax rate on earned income increases, the effective control that people have over their economic situation decreases. For example, consider an effective 25 percent tax rate on earned income. Someone earning $20 per hour and working forty hours per week would have pretax earnings of $800, with post-tax earnings of $600. To make up the $200 difference, an individual would have to work more than thirteen additional hours every week. Hence, taxes on earned income reduce the *effectiveness* of each labor hour and thus stymie the relationship between an individual's actions and their economic outcomes.[43] The effectiveness of economic decisions, such as working, saving, investing, obtaining additional skills, and changing careers, are all reduced by the extent to which earned income is taxed. Preserving labor ownership bolsters effective control by maintaining a tight relationship between what a person puts in (e.g., labor, skill learning, creativity, risk taking) and what a person gets out (e.g., wages, salaries, benefits, or other economic gains). Thus, preserving labor ownership by lowering the tax rate on earned income, or other economic gains from human labor, is desirable because it enables people to have more effective control over their economic situation. The upshot is that if we can find an alternative tax base that does not diminish labor ownership, then this would be a powerful reason that would count in favor of adopting that tax base.

I am *not* making the claim that preserving labor ownership is the only useful economic norm for giving people effective control over their lives, simply that preserving labor ownership significantly aids in this cause. The argument from effective control shows that we need not be wedded to either extremes regarding the principle of labor ownership. We do not need to choose between the Nozickean account where property claims are inviolable side constraints and the legal constructivist account where there are "no property rights antecedent to the tax structure."[44] Instead, we can hold that adhering to the principle of labor ownership enables individuals better effective control over their lives, without positing that the principle is absolute or that it requires recognizing some pre-political natural right. Furthermore, this approach is consistent with the more moderate expressions of labor ownership offered and defended by

43 This is why taxes on earned income have a disincentive effect.

44 Murphy and Nagel, *The Myth of Ownership*, 74.

contemporary theorists.[45] Notably, this conclusion about labor ownership does not imply that we should never tax earned income. The advantages from a particular spending program might outweigh the disadvantages from taxing earned income to fund such a program. However, the point is that taxing earned income comes with a moral cost—it diminishes the effective control that people have over their own economic lives.

The argument that I will make in the remaining sections does not rest upon accepting the principle of labor ownership. Some theorists may prefer different criteria altogether. To the extent that someone finds any version of labor ownership attractive, they will also find any tax base that does not diminish labor ownership attractive relative to one that does diminish it.

## 2. TYPES OF ECONOMIC GAINS

### *2.1. Productive vs. Unproductive Economic Gains*

The previous section described how the prevailing tax regime suffers from two drawbacks, but it did not describe their common cause. To explain the origin of the two problems, I distinguish between productive and unproductive economic gains.

An economic gain refers to money, property, or any other economic asset. Economic gains are *productive* in situations where the gain is instrumental in the creation of wealth, that is, any kind of valuable social surplus.[46] This occurs whenever the economic gain motivates an agent to engage in productive activity. Notably, the motive need not be selfish, as for example when individuals work in order to produce for their families or contribute to charitable causes. As an illustration, if I produce apples in exchange for money, the apples are the "wealth" and the money is the "economic gain." This is a productive economic gain, since the money was instrumental in the creation of the apples.

Recall that the prevailing tax regime includes taxes assessed on earned income, gains from capital, and consumption. These gains either constitute wealth itself, such as in the case of consumption, or are instrumental in the creation of wealth, such as in the case of earned income or gains from capital. Without the prospect of earning income or obtaining profits, the activities

45 See Russell, "Self-Ownership, Labor, and Licensing," 175; and Brennan, "Striving for the Middle Ground," 5.

46 Wealth is anything valuable or useful to human beings; it includes everything from automobiles and food catering to insurance.

responsible for generating wealth would vastly diminish.[47] Thus, the prevailing tax regime is composed of taxes levied on productive economic gains. What's more, the reason that taxing these economic gains is inefficient is *because* they are instrumental in the creation of wealth or social surplus. Taxing these gains affects the trade-offs of economic agents in a way that *distorts* their production and consumption decisions and consequently reduces the total supply of wealth. Taxing productive economic gains also diminishes labor ownership, because the person or people who produce a thing retain less of the benefit of that production. The fact that the prevailing tax regime possesses both of these drawbacks is not incidental. It is a consequence of most taxes in the prevailing tax-base regime being levied against productive economic gains.

So, if taxes on productive economic gains possess these two drawbacks, what other economic gains can be taxed that lack these drawbacks? Some economic gains *are not* instrumental in the creation of wealth—they are *unproductive* economic gains. Money obtained from theft and fraud are examples, as these are the result of the coercive taking of preexisting wealth. For obvious reasons, these cannot possibly serve as a tax base. However, there is a class of unproductive economic gains that are not intrinsically immoral, and hence may be legally permitted and could serve as the foundation of a tax base. Economists call these type of unproductive economic gains "rents." Rents are typically defined as "those benefits to an agent that are in excess of the minimum necessary for the agent to accept the transaction."[48] This definition is consistent with the original meaning, but is unnecessarily narrow. The use of the word "transaction" implies that rents are generated through exchange, but there is no reason to suppose that all rents accrue as a consequence of exchange. For example, when a landowner's property doubles in value because of the nearby development of parks and schools, it is not because she is party to some exchange or engages in any transaction. Yet, this example is consistent with what economists would generally regard as rents because the accrued gain does not contribute to supplying the land.[49] The fact that a person's land value is increased by proximate economic development is not instrumental to said development; it is a side effect. The landowner gains a value because of the productive activity of others. Thus, in order to allow for a wider consideration of the tax base, and to not unnecessarily prejudice ourselves against forms of

47 Earning money is not the only reason people engage in productive activities, but it is contributory, if not necessary, otherwise "working" would be indistinguishable from "volunteering."

48 Schwerhoff, Edenhofer, and Fleurbaey, "Taxation of Economic Rents," 400.

49 For the traditional definition, see Varian, *Intermediate Microeconomics*, 412.

rent that are "exchanges," I will continue using the broader account of unproductive economic gains.

## 3. THE ARGUMENT IN THE ABSTRACT

Before discussing specific nominees for rental taxation, I want to give the general form of the argument. My first goal is to demonstrate, in the abstract, how taxing rents avoids the two aforementioned drawbacks and hence makes for a promising tax base. I begin with the equity-efficiency tradeoff.

### *3.1. Does Rental Taxation Possess the Equity-Efficiency Tradeoff?*

As defined, rents are a form of unproductive economic gains. That is, the gains accrued from rents do not play a role in the creation of wealth because they do not induce any productive economic activity.[50] Consequently, a tax on an economic gain that does not produce wealth cannot create a disincentive to create wealth because no wealth is being created in the first place.[51] If the tax does not reduce the creation of wealth, then the tax would have no efficiency costs. That is, the same amount of wealth would exist before and after the tax. Consequently, a tax without efficiency costs would not pose an equity-efficiency tradeoff.[52]

To illustrate, suppose Bob gains $200 of economic rent. Whether the government taxes Bob's gain at a rate of 0, 50, or 100 percent, there is no equity-efficiency tradeoff, because the efficiency cost is always the same: zero. The tax rate determines how much Bob and the government split the $200, but the rate does not change the amount of wealth in existence; it just determines the distribution of the economic gain. It is important to clarify that the equity-efficiency tradeoff is *not* the same as tax progressivity, but it determines how progressive a tax can be, subject to the efficiency constraint. Economists look for taxes that do not have an equity-efficiency tradeoff because they can make the rates as progressive as desired without causing any efficiency loss.[53]

A tax on rents may not only have no efficiency cost, but may actually encourage the creation of wealth depending on the type of rent that is taxed. *Windfall rents* occur when someone receives a rent that they did not pursue, but still

50 Mulligan, "Do People Deserve Their Economic Rents?" 183.

51 Schwerhoff, Edenhofer, and Fleurbaey, "Taxation of Economic Rents," 410.

52 In a non-competitive market it is possible for a tax on rents to pose an efficiency cost at the margin. I discuss these noncompetitive market rents in section 6.

53 Schwerhoff, Edenhofer, and Fleurbaey, "Taxation of Economic Rents," 389.

accrue, simply because they are the beneficiary of an economic spillover.[54] In the context of windfall rents, the beneficiary is an economic patient, because no action or decision that they could undertake would affect the total supply of an economic good.[55] In the case of windfall rents, as for example when a person's property values increase because of nearby economic development, a tax on the windfall will not induce any behavioral change and so there will be no efficiency loss.[56] On the other hand, if a rent is generated because of *rent-seeking behavior*, then a tax on rent-seeking activity will alter the trade-offs of the rent seeker. Specifically, a tax on rents will increase the attractiveness of productive economic behavior at the margin relative to rent-seeking behavior. This is simply because a tax makes the opportunity cost of rent-seeking behavior higher relative to productive behavior. [57] Thus, a tax on rents of this kind would not only be costless, but would actually generate additional wealth.[58]

Whether a tax on rents simply has no efficiency loss or has a *negative efficiency loss*, i.e., it positively generates wealth, depends on whether the rent is levied on the windfalls of economic patients or on the rental gains made by active rent seekers. The upshot is that taxes on competitive market rents have no equity-efficiency trade-off and, in the case of taxing rent seekers, may actually by hyperefficient.[59]

There are situations where rents are generated by rent-seeking behavior, but the beneficiary of the rent is not the same as the rent seeker. In these cases, taxing the rental beneficiary will not result in efficient outcomes because those imposing the rent do not themselves face the altered trade-offs directly.[60] I discuss these cases of noncompetitive market rents in section 5.

### *3.2. Does Rental Taxation Diminish Labor Ownership?*

Recall that the principle of labor ownership says that the person (or people) who creates or produces a thing has some bundle of property rights in that

54 Alterman, "Land Use Regulations and Property Values," 1–5.

55 Medda, "Land Value Capture Finance for Transport Accessibility," 156.

56 Medda, "Land Value Capture Finance for Transport Accessibility," 157.

57 Sobel and Garret, "On the Measurement of Rent Seeking and Its Social Opportunity Cost," 117.

58 Stiglitz, "The Origins of Inequality and Policies to Contain It," 433.

59 "The tax shifts investment towards reproducible stocks, alleviating their undersupply and leading to higher output and aggregate consumption" (Edenhofer, Mattauch, and Siegmeier, "Hypergeorgism," 476).

60 This occurs in the case regulatory capture, where rule makers create, administer, or authorize rules that generate rents for other parties such as in the case of labor rents or monopoly rents. See Aidt, "Rent Seeking and the Economics of Corruption, 147–51.

thing that others do not possess. As previously mentioned, the "thing" created may be the performance of certain activities, e.g., labor, or a product, service, or even a string of words or set of ideas, over which the owner has the right to sell, trade, use, or otherwise dispense. A condition of having a labor-ownership claim seems to be that something was produced or instantiated (in the case of labor) to have a claim over. However, rents are economic gains that do not contribute to the creation of wealth. If no wealth, however broadly defined, was generated by the rentier, how could they have a labor-ownership claim to the economic gains accrued from wealth that others created? If the rentier engages in no productive activity, it is hard to see how their property claims could be grounded in the principle of labor ownership.

One response might be that perhaps the creation of wealth is not required to confer labor ownership, just that the rentier is engaged in some activity that can be construed as labor in which they reallocate wealth from others to themselves. Earlier, we distinguished between rents generated by windfalls and rents generated by rent-seeking behavior. In the case of windfall rents, the person receiving the rent performs no activity that can be construed as labor; rather, the economic gain they receive is merely a side effect of the actions of others. For this reason, windfall rents can be disqualified from having any grounding in the principle of labor ownership. By contrast, rent-seeking behavior arguably involves the performance of some kind of labor. However, the so-called labor under consideration is not useful, at least in the sense that it does not produce any additional wealth or social value. The labor of rent-seeking behavior merely reallocates already existing wealth from others to themselves.[61] Rent-seeking behavior has this feature in common with stealing, but nobody would justify burglary on grounds that the thief put in some work. Thus, it cannot be on grounds of effort alone that labor ownership is justified.

Traditional accounts of labor ownership, whether they be Lockean, Georgist, Nozickean, or Dworkinian, imply that it is the creation of something *new or valuable* that entitles a person to property, or at the very least that others not be made *worse off* in the process.[62] This also corresponds to the justificatory language people use in economic contexts where they reference the "product" or "fruits" of their labor.[63] People do not typically justify their holdings by referencing the "takings" or "appropriations" of their labor. Whether through some

61 What Stiglitz calls "exploitation rents" ("The Origins of Inequality and Policies to Contain It," 432–34).

62 See Locke, *Second Treatise of Government*, ch. 5; George, *Progress and Poverty*, 122; Nozick, *Anarchy State and Utopia*, 175; and Dworkin, "What Is Equality?" 287.

63 This language is also common in the literature, see Otsuk, "Self-Ownership and Equality," 74; and Steiner, "Left Libertarianism and the Ownership of Natural Resources," 5.

process of labor mixing, differential sacrifice, or self-sovereignty, labor-ownership claims are conferred when there is an act of creation or production, or some form of contribution generated by the labor owner. Yet, rent seeking is not act of creation; it adds nothing to the social surplus. On the contrary, rent-seeking behavior appropriates wealth from others without generating any value, thereby making others worse off. Further, rent-seeking behavior cannot be justified on grounds that it enables the individual to have greater effective control over their lives. This would imply that my ability to effectively control my life comes at your expense and vice versa. It is only on grounds that others not be made worse off that such a principle can be justified, and the gains from rent seeking fail to satisfy this common standard.[64] Consequently, property claims in rent cannot be justified on grounds of labor ownership, thus taxing rents does not diminish labor ownership.

It is important *not* to overstate the implications of this argument. Denying that economic rents are grounded in the principle of labor ownership does not imply that there are no legitimate property claims in rent. Property claims in rent could plausibly be justified on other grounds.[65] Enacting institutional rules of any kind will likely generate property claims in rents of one form or another, but that is acceptable if we have good reasons for adopting those rules. The upshot of my argument is not to deny that we can have property claims in rent. Rather, the upshot is that taxing rents does not diminish the principle of labor ownership. This is an advantage to taxing rents over and above taxing productive economic gains for reasons discussed in section 2. Namely, preserving labor ownership creates good incentives, promotes individual sovereignty, aligns with established economic norms, and allows individuals greater effective control over their economic lives.

## 4. POTENTIAL NOMINEES FOR RENTAL TAXATION

I have argued that taxing rents lacks the drawbacks that afflict taxing productive gains, but have not yet specified what rental taxation would look like in practice. It is beyond the scope of this paper to offer policy prescriptions, though it may by instructive to provide some sense of what a tax base composed of rents may look like. The subsequent examples are only intended to demonstrate some

64 This common standard includes the Pareto principle, Locke's proviso, Nozick's proviso, and George's principles of acquisition. Also see Stiglitz and Rosengard, *Economics of the Public Sector*, 63–65; Locke, *Second Treatise of Government*, 12–18; Nozick, *Anarchy State and Utopia*, 172–82; George, *Progress and Poverty*, 122–24; and Russell, "Self-Ownership, Labor, and Licensing, 186.

65 For example, "first possession" rules. See Schmidtz, "The Institution of Property," 5–9.

possible avenues of rental taxation. Whether these tax nominees are desirable bases of taxation requires further investigation and research. I offer these examples to be illustrative, knowing that they are not definitive.

*4.1. Land Rents*

Land offers the paradigmatic example of rental taxation. Indeed, the term rents, originally coined by the eighteenth-century physiocrats, refers to the rental value of land.[66] A given piece of property or real estate has two sources of value: the value of the land itself and the value of improvements made by the owner. The economic gains accrued from the land-ownership portion alone (distinct from property improvements made by the land owner) can be regarded as a rent because it has no origin in the landowner's productive activities but rather is a consequence of the land's natural features as well as surrounding improvements.[67] For instance, much of a property's value is a consequence of its proximity to desirable features, such as schools, parks, restaurants, commercial activity, or beautiful scenery. As development increases in an area, landowners gain a significant windfall based on the improvements made by others, i.e., proximate improvers.[68]

A land value tax is designed only to tax the value derived from the land-holding portion of the property and *not* to tax the value derived from physical structures or other human improvements. As the land's value increases due to proximate development, and the landowner receives a windfall, a tax is levied on that windfall and returned to the community.[69] The tax is generally regarded as being efficient because it does not distort the behavior of the landowner.[70] The landowner is, in our terminology, an economic patient, because they do not supply the land or its proximate development; they merely capture positive spillovers from neighbors.[71] The landowner's behavior is not distorted by the tax because the supply of land is fixed, i.e., more land cannot be produced.[72] The extreme inelasticity of the supply of land means that a tax will not distort the behavior of the landowner and hence a tax on land's rental value generally has no efficiency costs; it can only redistribute the gains from land ownership. Hence,

66 Lackman, "The Classical Base of Modem Rent Theory," 287.

67 Schwerhoff, Edenhofer, and Fleurbaey, "Taxation of Economic Rents," 411–12.

68 Medda, "Land Value Capture Finance for Transport Accessibility."

69 For a survey of the different instruments used for achieving this, see Alterman, "Land Use Regulations and Property Values."

70 Mattauch, *Rent and Redistribution,* 11–13.

71 Mclean, "The Politics of Land Value Taxation," 11.

72 Oates and Schwab, "The Impact of Urban Land Taxation," 17–19.

a tax on land rents has no equity-efficiency tradeoff.[73] This desirable feature of land as a source of taxation has been recognized by many economists, notably Henry George, William Vickrey, Milton Friedman, and Joseph Stiglitz.[74]

Taxing the land's rental value is efficient, but does it diminish labor ownership? Since the land itself has been supplied by nature and increases to the land's marginal value are supplied by the activities of proximate improvers, the principle of labor ownership cannot apply to the landowner.[75] Labor ownership arguments on behalf of the landowner would only apply to improvements on the land or of its use, but not the land itself.[76] This point was first made prominent by Thomas Paine, who held that "the earth … [is] the common property of the human race" and "that it is the value of improvement only, and not the earth itself, that is individual property."[77] Indeed, if anyone has a claim on grounds of labor ownership to the rental value of the land it is the members of the local community responsible for the surrounding improvements that increase the land's marginal value, not the landowner. Labor ownership may imply that the proximate improvers would have claim to the portion of the surplus of the land's rental value that their improvements generated. Knowledge problems would likely make it impossible to disentangle whose contribution was responsible for each marginal improvement in the value of the land. From a labor-ownership standpoint, returning the surplus value of these positive spillovers back to the community in the form of tax revenue is better than allowing it to be captured by landowners, especially if these levies replace taxes on other economic gains accrued through labor ownership.

Some have raised concerns regarding implementation of a land value tax, most notably related to the problem of "unrealized value."[78] Specifically, while the landowner benefits from surrounding improvements, they may not mon-

73 This is not meant to imply that a land tax will always be progressive—the distributional benefits will depend on the particular land value tax policy and the distribution of land holdings—rather, it indicates that the tax rate can be made progressive without creating distortions.

74 See George, *Progress and Poverty*, 156; Vickrey, "Site Value Taxes and the Optimal Pricing of Public Services"; Stiglitz, "The Theory of Local Public Goods Twenty-Five Years after Tiebout," 14; "An Interview with Milton Friedman."

75 See Vallentyne, Steiner, and Otsuka, "Why Left-Libertarianism Is Not Incoherent, Indeterminate, or Irrelevant," 202.

76 "While the finiteness of land makes all claims to perpetual possession inconsistent with Locke's proviso, some claims to the use of land are consistent with it" (Tideman, "Takings, Moral Evolution, and Justice," 1724).

77 Paine, *Agrarian Justice*, 4.

78 For problems of unrealized value in connection to wealth taxation, see Fleischer, "Not So Fast," 265–66 and 288.

etize these benefits until they sell the land. A natural question then becomes whether the tax should be assessed as a point-of-sale tax on the land's value or as an annualized land-rental tax. A point-of-sale land value tax would consist of a single large tax payment assessed against the total rental value accrued from the point of purchase to the point of sale, while the annualized version consists of a continuous stream of annual payments assessed against each year's portion of the land rent. Determining the proper implementation of the tax depends on a number of empirical and conceptual considerations. The point-of-sale route allows greater flexibility to the taxpayer because it foregoes any liquidity concerns for those who are "land rich and cash poor."[79] However, from a public revenue standpoint, a series of smaller payments made by all landowners every year is preferable to large but infrequent payments where some landowners may only pay the tax every several decades. The annualized version of the tax is more consistent with conceptualizing the land tax as an annual fee that the owner pays in compensation for removing the land from "the commons" and thus giving the landowner exclusive right of use for as long as they continue to pay the rental portion back to the community.[80] In general, the annualized version is likely a more preferable tax instrument, but the details of land-tax policy ought to be construed in a way that accommodates liquidity concerns related to unrealized value, such as allowing payment deferrals or implementing modest tax-exemption thresholds on low-value land plots.[81]

*4.2. Incidental Inheritance*

The economic literature discusses many types of inheritance, but for our purposes these can be broken down into two types: deliberate bequests and incidental bequests.[82] A deliberate bequest is one in which the donor changes (or would have changed) their productive activity to ensure a specific recipient or group of recipients receives an economic gain following their death. For example, a person may want to help their adult children get a solid start or ensure that a disabled child has enough saved in trust for their entire life span,

79 Mclean, "The Politics of Land Value Taxation," 14.

80 George, *Progress and Poverty*, 158.

81 For an outline of proposals, see Mclean, "The Politics of Land Value Taxation," 11–14.

82 Roughly speaking, what is known in the literature as "accidental bequests" and "capitalist bequests" (or "wealth-loving bequests") qualifies as incidental bequests, while "voluntary" or "planned" bequests qualify as deliberate bequests. I adopt new terminology in this context because it better captures the difference between productive gains and economic rent. For the traditional terminology listed above, see Masson and Pierre, "Bequests Motives and Models of Inheritance," 54–88; and Piketty and Saez, "A Theory of Optimal Inheritance Taxation," 1866.

or they may have a specific charitable purpose in mind. Deliberate bequests are characterized by the fact that a donor had a goal for their bequest *and as a consequence* their economic activity was in part formed around that purpose.[83] By contrast, an incidental bequest occurs when the bequest was not instrumental in motivating the donor to engage in productive economic activity.[84] For example, this might occur when the donor's retirement savings exceeds their life span due to uncertainty regarding the donor's longevity or when the productive decisions of the donor end up being more profitable than their original expectations.[85] This would include the gains that the children of the ultra-wealthy (e.g., mega-millionaires and billionaires), receive upon their parents' death. The greater portion of these bequests are unproductive economic gains since it is unlikely that the billionth inherited dollar was instrumental in motivating the productive activities of the donor in the first place.[86]

The implication of this distinction is that deliberate inheritances are productive economic gains while incidental inheritances are rents. Taxing deliberate inheritance will alter the trade-offs of the donor because knowing that the bequest will be taxed will change their economic decisions. By contrast, incidental inheritances were not produced *for* the recipient, though the donor may have wanted the recipient to receive them in the event of the donor's death.[87] For example, we may want our children to have our excess accruals, but this does not mean we produced or saved those gains *for* that purpose. Taxing incidental inheritance will not alter the production and consumption decisions of the donor and is therefore efficient.[88]

As it concerns labor ownership, an inheritance tax cannot diminish the labor ownership of the inheritor, because the inheritor did not produce the inheritance.[89] A common right-libertarian argument against inheritance taxation argues that it diminishes the labor ownership of the donor, since it reduces

83 Deliberate bequests encompass "voluntary bequests," "planned bequests," and "altruistic bequests." See Masson and Pierre, "Bequests Motives and Models of Inheritance," 57; and Batchelder, "Leveling the Playing Field between Inherited Income and Income from Work through an Inheritance Tax," 70–71.

84 Piketty and Saez, "A Theory of Optimal Inheritance Taxation," 1866–67.

85 For accidental bequests in the case of life-span uncertainty see Hurd, "Mortality Risk and Bequests." For capitalist bequests see Masson and Pierre, "Bequests Motives and Models of Inheritance," 71.

86 See Batchelder, "Leveling the Playing Field between Inherited Income and Income from Work through an Inheritance Tax," 50–51; and Francis, "Wealth and the Capitalist Spirit."

87 Kopczuk, "Taxation of Intergenerational Transfers and Wealth," 10.

88 Batchelder, "Leveling the Playing Field between Inherited Income and Income from Work through an Inheritance Tax," 50–51.

89 A point recognized first by John Stuart Mill. See Pedersen, "Just Inheritance Taxation," 5.

their effectiveness in transferring their wealth to others.[90] Daniel Halliday challenges this argument by pointing out that the inheritance tax is considerably less coercive than other forms of taxation because it tends to pose fewer opportunity costs to the labor owner.[91] Essentially, the labor owner can fully make use of their wealth while they are alive, but only incurs a tax at the time of transfer—when they are dead—whereas all other taxes diminish the wealth of the labor owner while they are alive.[92] Further, by targeting incidental inheritance for taxation as opposed to deliberate inheritance, the tax base would be drawn from sources that do not interfere with the purposes for which donors generated these economic gains in the first place. Compared to either a generic inheritance tax or other forms of taxation, taxing incidental inheritance better preserves labor ownership and is non-distortionary.

Drawing a conceptual distinction between deliberate and incidental inheritance means that in theory it is possible to tax inheritance rents, but this does not itself assist the tax administrator in determining how to tax inheritance, since the actions and motivations of bequesters are unknown to them. Traditional explanations for what motivates inheritance include altruism toward progeny, precautionary savings as a kind of lifelong insurance, and promissory payments to adult children in exchange for elder care.[93] However, identifying a common bequest motive across large heterogeneous populations is notoriously elusive.[94] Given the mixed motivations of bequesters and the inability for any tax administrator to know the circumstances in each case, how could it be possible to create a tax that distinguishes the rental portions of inheritance from its productive portion? Of course, no administrable tax system can perfectly target a given tax base, but nonetheless a sound tax system should have a reasonable degree of accuracy.[95] There are at least two possibilities for practically separating incidental inheritance from deliberate inheritance. I briefly

90 For a brief overview of this type of argument see Halliday, "Is Inheritance Morally Distinctive?" 621–24.

91 Halliday, "Is Inheritance Morally Distinctive?" 627–31.

92 "Other taxes generally aren't like this. Taxes imposed before death generally have a greater impact on the value of the various options open to an individual at the time the tax is imposed" (Halliday, "Is Inheritance Morally Distinctive?" 632).

93 See Fried, "Who Gets Utility from Bequests?" 646–56; Blumkin and Sadka, "Estate Taxation with Intended and Accidental Bequests," 2–3.

94 "Any attempt to explain intergenerational transfers by a single motive seems hopelessly oversimplified" (Fried, "Who Gets Utility from Bequests?" 653).

95 Halliday remarks, "It should be conceded that most tax schemes are going to be heuristics that will generate a number of false positives and negatives" (*Inheritance of Wealth*, 187). Also see Fried, "Compared to What?" 385.

discuss two such possibilities in turn: (a) the "Rignano scheme" and (b) the pre-commitment strategy.[96]

The presence of inheritance can be a consequence of a single generation or the result of a continuous chain of multigenerational wealth. One proposal that sees periodic consideration in the philosophical literature involves treating single-generation versus multigeneration inheritance differently. [97] The Rignano scheme involves treating the presence of newly created inherited wealth at a fairly low tax rate while levying a high tax rate on the presence of wealth that is passed on to subsequent generations.[98] Under a Rignano scheme, if Alice earns and bequeaths a million dollars to her son, Bill, the tax on that inheritance is extremely minimal. Suppose in the next generation Bill passes on one-and-a-half million to his daughter, Karen. The first million would now be taxed at a very high rate of taxation since this wealth is carried over from the first generation, but the subsequent half million that Bill earns and passes on would be taxed at a low rate. The idea is consistent with maintaining incentives for donors to be productive across successive generations, as well as the idea that the originators of wealth should have the ability to transfer their wealth as they see fit, but that subsequent rentiers have less of a claim to transfer wealth free from taxation. The Rignano scheme's treatment of multigenerational transfers tends to capture a strong portion of incidental inheritance, since the taxed wealth was not even produced by its donor. However, the Rignano scheme's treatment of first-generation transfers is overly generous, as a significant portion of these transfers are likely to be incidental inheritance as a consequence of precautionary savings or unanticipated economic success. A Rignano scheme with perhaps a steeper initial progressive tax curve on first-generation transfers could be a decent approximating tool for taxing incidental inheritance while minimizing deliberate inheritance taxation.

Another possibility for taxing inheritance rents is to create a tax system with a precommitment device that incentivizes donors to reveal their preferences and thus create a separating equilibrium between incidental and deliberate inheritance. For example, governments could allow for tax-advantaged accounts up to some limit, but the donor would not be permitted to withdraw funds from the account and instead those funds would be placed in a trust for the recipient upon the donor's death. Alternatively, withdrawal from the tax-advantaged trust could be allowed, but only after paying an extremely high rate to serve as a penalty for failing to commit. Any inheritance or gift transferred

96 For the Rignano scheme, see Halliday, *Inheritance of Wealth*, 54–72 and 195–96.

97 See Halliday, *Inheritance of Wealth*, 54–72; Pedersen, "Just Inheritance Taxation," 6.

98 Halliday, *Inheritance of Wealth*, 54–72, 195–96.

outside this system would be subject to taxation at progressive a rate as desired. The consequence is that a properly designed combination of tax advantaged trusts with a significant withdrawal penalty could create a separating equilibrium between deliberate and incidental inheritance.

*4.3. Zero-Sum Financial Transactions*

Market transactions generally create value because participants receive something more valuable to them in exchange for giving up something less valuable to them. Some transactions initially appear to be zero sum, meaning that the amount of wealth that exists at the beginning of the transaction is the same that exists at the end of the transaction. Insurance is a notable example of being zero sum on paper, but actually increases the expected value for participants since it secures them against the possibility of greater losses. However, there is increasing skepticism that many zero-sum financial transactions have a socially beneficial function. For example, Posner and Weyl make the case that many of these transactions are merely a form of "financial gambling" and are "welfare reducing and contribute to systemic risk."[99] I make no claims regarding particular financial transactions but seek to demonstrate the plausibility of this hypothesis with a simple example.

Poker is a zero-sum game on paper. If it were not for the enjoyment of the participants in playing the game, poker would not produce any social surplus. Imagine that the participants in the poker game were not human beings enjoying themselves, but computer algorithms programmed by their creators to try and maximize their earnings from online poker games. If there are six poker programmers, then we know that five of them will be sorely disappointed with the outcome of the game. The consequence of the joyless poker game will be that five players will be worse off and one player will be significantly better off. The gains accrued by the victor of our joyless poker game produce no additional wealth; they just redistribute already existing holdings.

Taxing the gains from joyless poker games will not reduce efficiency, but curiously it would alter the trade-offs of the participants. This is because, unlike land rents and incidental inheritance, these rents are not windfalls but a consequence of rent-seeking activity. Taxing rent-seeking activity will alter the trade-offs of our poker programmers, but in a socially beneficial way. As discussed in section 3.1, taxing the gains from rent seeking will not only have zero efficiency costs, but will also create an incentive at the margin to engage in more productive activity. As the tax rate increases on these gains, the poker programmers have greater incentive to abandon rent seeking in pursuit of more profitable

99 Posner and Weyl, "An FDA for Financial Innovation," 1317.

and, likely, more productive activities. Furthermore, justifications from labor ownership are absent; no good was created nor any service rendered. It is on grounds of consent to participate in a game, not on grounds of labor ownership, that the gains from the joyless poker game are justified.

Is the joyless poker game merely an interesting contrivance or an accurate approximation of some of the zero-sum securities markets, including options, derivatives, futures, foreign currency, cryptocurrency, and sports gambling?[100] Such an answer is likely to be nuanced and requires further investigation, but many well-informed commentators question whether many of these zero-sum financial transactions actually create wealth.[101] If there are types of financial transactions that are akin to a joyless poker game, then these transactions are ripe possibilities for rental taxation.

### *4.4. Negative Externalities*

One of the primary justifications for a competitive market system is that an offering's price tends to approximate its marginal social cost. The assumption of this justification is that the marginal cost endured by the firm is equivalent to the social cost endured by society. Negative externalities describe the situation when this condition not met—when the price of an offering underrepresents its marginal cost. Negative externalities are peculiar because most activities that produce negative externalities *also* generate wealth. For example, burning fossil fuels creates pollution but also powers modern industry. The problem is that because the supplier or the demander of the productive activity does not bear its full costs, the activity is oversupplied. Hence, for negative externalities it is the *level* at which the productive activity is supplied that is the problem, as opposed to the activity itself being intrinsically unproductive. This means that for some *portion* of the activity (and *not* the activity considered on net), more wealth is being destroyed than being created. Thus, there are economic gains being accrued on a portion of the activity that is not only unproductive, but actively destructive. The gains accrued from portion of the activity that generates negative externalities can properly be considered rents since they are

100 The justification offered here for taxes assessed on gambling is on the basis of their possibly constituting economic rents, and not on paternalistic grounds. Using taxation for paternalistic reasons is an intriguing idea since raising cost of bad habits may reduce their prevalence. However, such a strategy can also be self-defeating. If the tax does not alter the person's behavior, all that has been accomplished is to make their bad habit even more expensive, which makes them worse off.

101 See Posner and Weyl, "An FDA for Financial Innovation"; Stiglitz, "The Origins of Inequality and Policies to Contain It," 432–34; Mazzucato, "Financing Innovation," 858; and Buiter, "Useless Finance, Harmful Finance, and Useful Finance."

gains that are not instrumental in the creation of wealth—in fact they destroy wealth, *at the margin*.[102]

Since we often do not want to prohibit the activity as a whole, but rather temper it, a commonly proposed solution to negative externalities is the "Pigouvian tax," a levy designed to internalize all costs of production so that firms produce the optimal amount of the offering.[103] The idea is that such a tax internalizes social costs and pushes the price toward the equilibrium point so that the price of the offering either approximates, or nears approximating, its marginal cost. While taxation on negative externalities is attractive for efficiency reasons, it operates differently than the other taxes on rents we have discussed. In the case of land, incidental inheritance, and zero-sum transfers, these rents could theoretically be taxed at 100 percent and remain efficient. This is because there is no portion of these rents that generates wealth. By contrast, in the case of activities that generate negative externalities, there is a specific price point (or range) at which the production of the offering is efficient. This means that, while Pigouvian taxation can be efficient, it does not necessarily avoid the equity-efficiency tradeoff in a way that is characteristic of the other forms of rental taxation. Pigouvian taxation may still be very attractive because it can increase efficiency, but because the tax is constrained by the goal of finding the optimal price range, it lacks the rate flexibility that the other forms of rental taxation possess. Still, Pigouvian taxation is likely to be progressive, since those who bear negative externalities tend to be poorer than those who benefit from not paying the costs of those externalities. However, Pigouvian taxation is considerably less flexible than the other rental taxes discussed.

Do levies on negative externalities diminish labor ownership? A Pigouvian tax that only discourages the destructive portion of the activity would not diminish labor ownership, so long as we assume that labor ownership is reasonably constrained by some account of harm or individual rights. This is a widely accepted condition placed upon freedom more generally and labor ownership in particular. Such a stipulation is reflected in Mill's harm principle and Locke's proviso.[104] However, the constraint on labor ownership only applies when the Pigouvian tax is properly designed and does not overly restrict the activity in question so as to diminish its wealth-generating portion.

102 Schwerhoff, Edenhofer, and Fleurbaey, "Taxation of Economic Rents," 401.

103 Mankiw, "Smart Taxes."

104 See Locke, *Second Treatise of Government*, 12–14; and Robson, *Collected Works of John Stuart Mill*, 223

## 5. COMPETITIVE MARKET RENTS VS. NONCOMPETITIVE MARKET RENTS

The aforementioned examples are potentially promising nominees for rental taxation. Yet, not all rents may be suitable targets for taxation and some rents may be better handled by adjusting the system of rules. The contemporary economic literature tends to focus on rents that are a consequence of regulatory capture.[105] Occupational licensing rules may protect consumers in situations of low information and intellectual property law can create incentives for research and discovery. Yet, the presence of these rule systems can also be leveraged or co-opted by rent-seeking firms, professional associations, or political lobbyists. When the rules governing economic suppliers (either firms or workers) are abused or co-opted in a way that artificially restricts some suppliers in favor of other suppliers, the subsequent gains are economic rent. Prime examples of these include monopoly/oligopoly rents and labor rents. There are many causes for an artificial scarcity of suppliers, including guilds, cartels, overly restrictive intellectual property law, excessively demanding licensing rules, or principal-agent problems.[106] The economic gains that accrue to permitted suppliers are artificially inflated because other legitimate suppliers are being prevented from participating, such as competing firms in the cases of monopoly/oligopoly rents and competing workers in the case of labor rents. The portion of gains attributable to decreased competition is unproductive because it does not stimulate wealth generation but merely transfers consumer surplus to producer surplus in the form of higher prices to consumers and higher wages/profits to permitted suppliers.[107]

In addition to being inefficient, these artificial restrictions also violate the labor ownership of those suppliers who are unfairly excluded. Suppose for example that obtaining a medical license is so onerous and challenging that it not only excludes unqualified medical practitioners, but also excludes a large portion of qualified medical practitioners. Qualified candidates that are excluded from medical practice are unfairly not being allowed to pursue promising careers and thus have a diminished potential to exercise their abilities in a way that allows them to control their own economic situation. Dan Russell argues that restrictive licensing constitutes "takings of property in labor" and

105 See Dal Bó, "Regulatory Capture"; and McChesney, "Rent Extraction and Rent Creation in the Economic Theory of Regulation."

106 See Aidt, "Rent Seeking and the Economics of Corruption," 147–51.

107 Labor rents may also occur as a result of principal-agent problems. See Kräkel and Anja, "Internal Labor Markets and Worker Rents." For example, CEOs and other executives likely receive labor rents as a consequence of their partial associations with their boards of directors. See Moriarty, "Do CEOs Get Paid Too Much?" 260–62.

therefore faces a justificatory burden.[108] That justificatory burden is rooted in a conception of reciprocity.[109] Specifically, he contends that "the justification for taking is that even those from whom property is taken are better off in the greater scheme of things for living in the sort of community that such a power to take property makes possible."[110] Notably, this conception of labor ownership is not absolutist. It would justify restrictions on labor where doing so would make those subject to those restrictions better off in the greater scheme of things, such as in the case of prohibiting unqualified doctors from practicing medicine. Yet, a restriction on labor does not satisfy this justificatory burden when it "does not allow positive sum transfers," such as in the case of prohibiting market entry of qualified medical applicants.[111] Accordingly, the rules that create supplier rents also diminish labor ownership by unfairly restricting liberty in cases where it is reciprocally beneficial.

Are supplier rents a promising target for taxation? Supplier rents are notably different from our previous examples of economic rent. In the other four cases we have discussed, the rents exist within, or alongside, competitive markets. None of the previous cases involved a market where competition was artificially restricted. By contrast, within the context of noncompetitive markets, the ensuing supplier rents are not a result of windfalls, nor of distortions, created by the market participants acting within the rules of the game, but rather as a result of *rules that distort the game*. Unlike our previous examples, it is the rule makers, *not* the market participants that are the relevant decision makers, i.e., economic agents, when it comes to the existence of these rents. Thus, it is the trade-offs that face rule makers that are most relevant in these cases. For example, an overly restrictive system of medical licensing that creates an artificial scarcity of doctors is a consequence of political lobbying on the part of medical associations or affiliated interest groups; it is not a result of practicing medicine. Thus, taxing doctors for their labor rents would not affect the trade-offs of the relevant economic agent in the right way, but rather would exacerbate the existing artificial scarcity of medical care. A proper rental tax in this case would involve discouraging behavior at the level of political lobbying, not at the level of medical supply. The problem with these labor rents is not the fact that some doctors are practicing medicine, but, instead, that other qualified would-be doctors are *prevented* from practicing medicine. Generalized, supplier rents accrue to *some* suppliers but are not the consequence of the activity of supplying. Instead,

108 Russell, "Self-Ownership, Labor, and Licensing," 174.

109 Russell, "Self-Ownership, Labor, and Licensing," 179.

110 Russell, "Self-Ownership, Labor, and Licensing," 180.

111 Russell, "Self-Ownership, Labor, and Licensing," 186.

supplier rents are the result of rule makers creating, administering, or enforcing rules that diminish the labor ownership of other economic suppliers.

Rental taxation could theoretically be effective in the case of supplier rents if applied at the proper level. The rental tax would have to be levied on rent-seeking behavior at the level of political decision making. While such a tax on rule-maker rents is intriguing, what would it look like in practice? Imaginations could run wild regarding the details of such a tax. It could be applied in the form of a large fixed fee for all political lobbying efforts, or as a penalty to firms that lose lawsuits that judges deem to be frivolous attempts to corner the market by leveraging the courts. The problems with such proposals are myriad. Systems of taxation virtually always apply to economic participants (including economic agents and economic patients), and I can recall no instance of taxation being applied to rule makers or at the level of political decision making itself. There are likely a host of good reasons for this, including deep knowledge problems concerning the optimal amount, context, and application of the tax; agency problems; and enormous constitutional concerns. Future proposals of rental-tax design might reflect on whether such a "rule-maker tax" is possible and/or desirable. However, given the size and scope of these challenges, rents that exist because of noncompetitive markets are likely better resolved by properly updating the rule systems than by trying to tax these rents. Rents that exist in the context of competitive markets are more promising targets of taxation, since there is either (a) no altered trade-offs, as in the case of windfall rents, or (b) in the case of rents from rent seeking, there is no disambiguation between the rental beneficiary and the rent seeker.

## 6. CONCLUDING REMARKS ON TAX-BASE SELECTION

There are different levels of disagreement that one could have with this project, not all of which are antithetical to that project's purpose. Someone might disagree that the tax nominees I have proposed constitute proper rents and may think that there are alternative rental nominees better suited for taxation. This level of disagreement is completely consistent with the greater aims of the project. Distinguishing which economic gains are rents from those that are genuinely productive requires further research. Yet, someone voicing this disagreement still accepts the central claim that we ought to replace taxes on productive economic gains with taxes on rents.

There are other levels of disagreement antithetical to the core project. Someone might deny that economic rents exist at all. Given the preponderance of evidence presented and absent a strong argument that all economic gains are productive, this standpoint is not credible. A more plausible objection contends

that it will be difficult to find a general class of taxable objects that always count as rents and not as productive economic gains. I have offered some potential nominees to illustrate that there are such general classes. One rebuttal might be to find exceptions with these nominees themselves and to argue that even these are not always rents. However, the appropriate standard does not lie in finding the platonic ideal of a tax base; instead it is a matter of choosing the best tax base among all relevant alternatives. Thus, if there are exceptions within a class of objects that generally constitute rents, we need only ask, "What are the alternatives?" For those skeptical that pure rents exist as a class, the thesis can be modified: "It is better to tax things that are more rent-like than things that are more consistently productive economic gains." Unless we think all classes of taxable objects are equivalently productive, an implausible proposition, we should want to tax classes of objects that are more rent-like instead of the class of objects that tends to be more productive.

Finally, the largest fundamental disagreement one could have with this project would be to claim that it is better to tax productive economic gains than rents. Given the drawbacks we have discussed, it is hard to imagine how someone might make this argument. Perhaps they would point to alternative normative features, ones that we did not consider within the scope of this paper. Yet, this avenue of objection looks increasingly dubious once we actually consider the other plausible normative features. Consider economic growth. It is widely established that taxing productive economic gains, especially gains from capital, have deleterious effects on growth.[112] By contrast, rental taxation is thought to have minimal effects on growth, and may even *increase* economic growth.[113] What about other moral features, such as desert? For those holding that desert should play a justificatory role in economic outcomes, it is typically on grounds of social contribution or productive activity that desert claims to economic gains are justified. Yet, as discussed, these are precisely the kinds of characteristics that rental economic gains lack and that productive economic gains possess.[114] Indeed, people generally object to individuals accruing large economic gains when they are "unconnected to underlying productive capacity," such as occurs in the case of rents.[115] By contrast, desert claims to economic gains are most justifiable when the gain is a consequence of creating social or economic value to society, such as occurs with productive economic gains.[116] Looking at

112 Mankiw, Weinzierl, and Yagan. "Optimal Taxation in Theory and Practice," 20–22.

113 Murphy, Shleifer, and Vishny, "Why Is Rent-Seeking so Costly to Growth?"

114 See Lamont, "Incentive Income, Deserved Income, and Economic Rents," 45.

115 Mulligan, "Do People Deserve Their Economic Rents?" 183.

116 Mulligan, "Do People Deserve Their Economic Rents?" 184.

other plausible normative considerations, such as incentives, social reciprocity, or differentiating gains from luck and choice, none of these seem likely to overturn the attractiveness of rental taxation compared to the alternative.[117] If anything, these features seem to strengthen and not weaken the case for rental taxation. There does not seem to be any plausible normative criteria that might lead us to favor taxing productive gains instead of rents, let alone one that would tilt the holistic balance of reasons in that direction.

The upshot of my central argument is as follows: to the largest extent possible, we ought to replace levies on productive economic gains with taxes on economic rents. Yet, given the amount of funding that modern governments require, is such a tax base large enough? It is an empirical matter whether full replacement of the tax base is possible, but even a degree of partial replacement would create enormous improvements. Since "deadweight loss increases with the square of the tax rate," even small reductions on the tax rates of productive gains can have outsized economic effects.[118] A parallel argument applies for labor ownership. Pushing the tax rate down a few percentage points can have a significant impact on the effective control that people have over their lives. This is especially true at lower rates of income, where even marginal tax reductions can be the difference between retaining control over one's economic situation or being at the mercy of external circumstances.

The fundamental ethos of this approach to taxation is that we should try to avoid taxing people's propensity to create, produce, or consume wealth, but rather tax their propensity to co-opt, exploit, or diminish it. Future research in this rental-taxation program will likely revolve around two major topics. First, what are the best nominees for rental taxation and how do we design tax rules and mechanisms so as to separate the rental portion of economic gains from the productive portion of those gains? We have already discussed several plausible nominees but considerably more investigation into alternative rental tax bases is needed. Second, what are the political and public choice barriers that have so far prevented the widespread use of rental taxation and how do we overcome those barriers? The widely cited Mirrlees Review comments "the economic case for a land value tax is simple, and almost undeniable. Why, then, do we not have one already? Why, indeed, is the possibility of such a tax barely part of the mainstream political debate, with proponents considered marginal and unconventional?"[119] Given its obvious normative benefits, the same question

117 For rental taxation in the context of luck egalitarianism, see Vallentyne, "Self-Ownership and Equality," 331.

118 Stiglitz, *Economics of the Public Sector*, 584.

119 Mirrlees and Adam, *Dimensions of Tax Design*, 373.

can be applied to rental taxation at large: Why has it gained so little traction? The answer likely stems from political influence and opposition by embedded interest groups. Political philosophy can play a further role in the tax debate by emphasizing the normative difference between justifications rooted in private interest from those justifications based on publicly recognizable moral considerations.[120]

*jeffers.matt7@gmail.com*

## REFERENCES

Aidt, Toke S. "Rent Seeking and the Economics of Corruption." *Constitutional Political Economy* 27, no. 2 (June 2016): 142–57.

Alterman, Rachelle. "Land Use Regulations and Property Values: The 'Windfalls Capture' Idea Revisited." In *The Oxford Handbook of Urban Economics and Planning,* edited by Nancy Brooks, Kieran Donaghy, and Gerrit-Jan Knaap, 755–86. Oxford: Oxford University Press, 2011.

Arulampalam, Wiji, Michael P. Devereux, and Giorgia Maffini. "The Direct Incidence of Corporate Income Tax on Wages." *European Economic Review* 56, no. 6 (August 2012): 1038–54.

Auerbach, Alan J., and James R. Hines Jr. "Taxation and Economic Efficiency." In *Handbook of Public Economics,* vol. 3, edited by Alan J. Auerbach and Martin Feldstein, 1347–21. Amsterdam: Elsevier, 2002.

Batchelder, Lily L. "Leveling the Playing Field between Inherited Income and Income from Work through an Inheritance Tax." In *Tackling the Tax Code: Efficient and Equitable Ways to Raise Revenue,* edited by Jay Shambaugh and Ryan Nunn, 48–88. Washington, DC: Brookings, 2020.

Blake, Aaron. "Obama's 'You Didn't Build that' Problem." *Washington Post*, July 18, 2012. https://www.washingtonpost.com/blogs/the-fix/post/obamas-you-didnt-build-that-problem/2012/07/18/gJQAJxyotW_blog.html.

Blumkin, Tomer, and Efraim Sadka. "Estate Taxation with Intended and Accidental Bequests." *Journal of Public Economics* 88, nos. 1–2 (January 2004): 1–21.

Brennan, Geoffrey. "Striving for the Middle Ground: Taxation, Justice, and the Status of Private Rights." In *Taxation: Philosophical Perspectives*, edited by

120 For very helpful comments and discussion, I am grateful to Alexander Schaefer, Andrew Jason Cohen, Charles Delmotte, Bill Glod, Peter Lindsay, Jeff Carroll, David Bebeau, Rick Jeffers, Jeremy Wong, and two anonymous referees from the *Journal of Ethics and Social Philosophy*.

Martin O'Neill and Shepley Orr, 60–80. Oxford: Oxford University Press, 2018.

Brennan, Geoffrey, and George Tsai. "Tax Ethics: Political and Individual." In *A Companion to Applied Philosophy,* edited by Kasper Lippert-Rasmussen, Kimberley Brownlee, and David Coady, 397–410, Hoboken, NJ: Wiley Blackwell, 2016.

Brewer, Mike, Emmanuel Saez, and Andrew Shephard. "Means-Testing and Tax Rates on Earnings." In Mirrlees and Adam, *Dimensions of Tax Design: The Mirrlees Review*, 90–201.

Buchanan, James M. "The Ethical Limits of Taxation." In *Limits and Problems of Taxation,* edited by Finn R. Førsund and Seppo Honkapohja, 4–16, London: Palgrave Macmillan, 1985.

Buiter, Willem. "Useless Finance, Harmful Finance, and Useful Finance." April 12, 2009. Originally published at http://blogs.ft.com/. Now available at https://static1.squarespace.com/static/54c161ffe4b063fc8ab03446/t/54ce9538e4b08cdce8e0b767/1422824760004/Useless+Finance%2C+Harmful+Finance%2C+and+Useful+Finance.pdf.

Dal Bó, Ernesto. "Regulatory Capture: A Review." *Oxford Review of Economic Policy* 22, no. 2 (Summer 2006): 203–25.

Dahan, Momi, and Michel Strawczynski. "Optimal Income Taxation: An Example with a U-Shaped Pattern of Optimal Marginal Tax Rates: Comment." *American Economic Review* 90, no. 3 (June 2000): 681–86.

Delmotte, Charles, and Jan Verplaetse. "What Is Wrong with Endowment Taxation? Self-Usership as a Prerequisite for Legitimate Taxation." In *Building Trust in Taxation,* edited by Bruno Peeters, Hans Gribnau, and Jo Badisco, 95–114. Cambridge: Intersentia, 2017.

Demsetz, Harold. "Toward a Theory of Property Rights." In *Classic Papers in Natural Resource Economics,* edited by Chennat Gopalakrishnan, 163–77. London: Palgrave Macmillan, 1974.

Dworkin, Ronald. "What Is Equality? Part 2: Equality of Resources." *Philosophy and Public Affairs* 10, no. 4 (Autumn 1981): 283–345.

Edenhofer, Ottmar, Linus Mattauch, and Jan Siegmeier. "Hypergeorgism: When Rent Taxation Is Socially Optimal." *FinanzArchiv: Public Finance Analysis* 71, no. 4 (December 2015): 474–505.

Enache, Christina. "Sources of Government Revenue in the OECD." Tax Foundation, February 19, 2020. https://taxfoundation.org/publications/sources-of-government-revenue-in-the-oecd.

Feldstein, Martin. "The Effect of Taxes on Efficiency and Growth." National Bureau of Economic Research, no. w12201 (May 2006): 1–27.

Fleischer, Miranda Perry. "Not So Fast: The Hidden Difficulties of Taxing Wealth." In *Wealth*, edited by Jack Knight and Melissa Schwartzberg, 261–308. New York: New York University Press, 2017.

Fleurbaey, Marc, and François Maniquet. "Optimal Income Taxation Theory and Principles of Fairness." *Journal of Economic Literature* 56, no. 3 (September 2018): 1029–79.

Francis, Johanna L. "Wealth and the Capitalist Spirit." *Journal of Macroeconomics* 31, no. 3 (September 2009): 394–408.

Fried, Barbara H. "Compared to What? Taxing Brute Luck and Other Second-Best Problems." *Tax Law Review* 53 (1999): 377.

———. "Who Gets Utility from Bequests? The Distributive and Welfare Implications for a Consumption Tax." *Stanford Law Review* 51, no. 4 (April 1999): 641–81.

George, Henry. *Progress and Poverty*. New York: E. P. Dutton and Co., 1879.

Halliday, Daniel. *Inheritance of Wealth: Justice, Equality, and the Right to Bequeath*. Oxford: Oxford University Press, 2018.

———. "Is Inheritance Morally Distinctive?" *Law and Philosophy* 32, no. 5 (September 2013): 619–44.

———. "Justice and Taxation." *Philosophy Compass* 8, no. 12 (December 2013): 1111–22.

Hurd, Michael D. "Mortality Risk and Bequests." *Econometrica* 57, no. 4 (July 1989): 779–813.

"An Interview with Milton Friedman," *Human Events* 38, no. 46 (November 18, 1978): 14.

Kearl, J. R. "Do Entitlements Imply that Taxation Is Theft?" *Philosophy and Public Affairs* 7, no. 1 (Autumn 1977): 74–81.

Kopczuk, Wojciech. "Taxation of Intergenerational Transfers and Wealth." In *Handbook of Public Economics*, edited by Alan Auerbach, Raj Chetty, Martin Feldstein, and Emmanuel Saez, 329–90. Boston: Elsevier, 2013.

Kräkel, Matthias, and Anja Schöttner. "Internal Labor Markets and Worker Rents." *Journal of Economic Behavior and Organization* 84, no. 2 (November 2012): 491–509.

Lackman, Conway L. "The Classical Base of Modem Rent Theory." *American Journal of Economics and Sociology* 35, no. 3 (July 1976): 287–300.

Lamb, Robert. *Property.* Cambridge: Polity Press. 2021

Lamont, Julian. "Incentive Income, Deserved Income, and Economic Rents." *Journal of Political Philosophy* 5, no. 1 (March 1997): 26–46.

Lindsay, Peter. "Ownership by Agreement." *Political Studies* 63, no. 4 (October 2015): 935–50.

Locke, John. *Second Treatise of Government: An Essay concerning the True Original, Extent and End of Civil Government*. 1690. Hoboken, NJ: John Wiley and Sons, 2014.

Mack, Eric. "Non-absolute Rights and Libertarian Taxation." *Social Philosophy and Policy* 23, no. 2 (July 2006): 109–41.

Mankiw, N. Gregory. "Smart Taxes: An Open Invitation to Join the Pigou Club." *Eastern Economic Journal* 35, no. 1 (January 2009): 14–23.

Mankiw, N. Gregory, Matthew Weinzierl, and Danny Yagan. "Optimal Taxation in Theory and Practice." *Journal of Economic Perspectives* 23, no. 4 (Fall 2009): 147–74.

Masson, André, and Pierre Pestieau. "Bequests Motives and Models of Inheritance: A Survey of the Literature." In *Is Inheritance Legitimate?* edited by Guido Erreygers and Toon Vandevelde, 54–88. Heidelberg: Springer, 1997.

Mattauch, Linus. *Rent and Redistribution: The Welfare Implications of Financing Low-Carbon Public Investment*. PhD diss. Technische Universitaet, 2015.

Mazzucato, Mariana. "Financing Innovation: Creative Destruction vs. Destructive Creation." *Industrial and Corporate Change* 22, no. 4 (August 2013): 851–67.

McChesney, Fred S. "Rent Extraction and Rent Creation in the Economic Theory of Regulation." *Journal of Legal Studies* 16, no. 1 (January 1987): 101–18.

McLean, Iain. "The Politics of Land Value Taxation." In *Taxation: Philosophical Perspectives*, edited by Martin O'Neill and Shepley Orr, 185–202. Oxford: Oxford University Press, 2018.

McPherson, Thomas. "The Moral Patient." *Philosophy* 59, no. 228 (April 1984): 171–83.

Medda, Francesca. "Land Value Capture Finance for Transport Accessibility: A Review." *Journal of Transport Geography* 25 (November 2012): 154–61.

Mirrlees, James A. "An Exploration in the Theory of Optimum Income Taxation." *The Review of Economic Studies* 38, no. 2 (April 1971): 175–208.

Mirrlees, James A., and Stuart Adam. *Dimensions of Tax Design: The Mirrlees Review*. Oxford: Oxford University Press, 2010.

Moriarty, Jeffrey. "Do CEOs Get Paid Too Much?" *Business Ethics Quarterly* 15, no. 2 (April 2005): 257–81.

Mulligan, Thomas. "Do People Deserve Their Economic Rents?" *Erasmus Journal for Philosophy and Economics* 11, no 2 (Fall 2018): 163–90.

Murphy, Kevin M., Andrei Shleifer, and Robert W. Vishny. "Why Is Rent-Seeking so Costly to Growth?" *American Economic Review* 83, no. 2 (May 1993): 409–14.

Murphy, Liam, and Thomas Nagel. *The Myth of Ownership: Taxes and Justice*. Oxford: Oxford University Press, 2002.

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

Oates, Wallace E., and Robert M. Schwab. "The Impact of Urban Land Taxation: The Pittsburgh Experience." *National Tax Journal* 50, no. 1 (March 1997): 1–21.

Otsuka, Michael. "Self-Ownership and Equality: A Lockean Reconciliation." *Philosophy and Public Affairs* 27, no. 1 (January 1998): 65–92.

Paine, Thomas. *Agrarian Justice.* London: University of London, 1903.

Pedersen, Jørgen. *Distributive Justice and Taxation*. London: Routledge, 2020.

———. "Just Inheritance Taxation." *Philosophy Compass* 13, no. 4 (April 2018): e12491.

Piketty, Thomas, and Emmanuel Saez. "A Theory of Optimal Inheritance Taxation." *Econometrica* 81, no. 5 (September 2013): 1851–86.

Posner, Eric A., and E. Glen Weyl. "An FDA for Financial Innovation: Applying the Insurable Interest Doctrine to Twenty-First-Century Financial Markets." *Northwestern University Law Review* 107, no. 3 (2013): 1307–57.

Ricardo, David. *Principles of Political Economy and Taxation*. London: Electric Book Company, 2000.

Russell, Daniel C. "Self-Ownership, Labor, and Licensing." *Social Philosophy and Policy* 36, no. 2 (Winter 2019): 174–95.

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

Robson, John M., ed. *Collected Works of John Stuart Mill*. Vol. 1, *Autobiography and Literary Essays*. London: Routledge, 2013.

Sandmo, Agnar. "Optimal Redistribution when Tastes Differ." *FinanzArchiv/Public Finance Analysis* (1993): 149–63.

Saez, Emmanuel. "The Desirability of Commodity Taxation under Non-linear Income Taxation and Heterogeneous Tastes." *Journal of Public Economics* 83, no. 2 (February 2002): 217–30.

———. "Using Elasticities to Derive Optimal Income Tax Rates." *Review of Economic Studies* 68, no. 1 (January 2001): 205–29.

Saez, Emmanuel, and Stefanie Stantcheva. "Generalized Social Marginal Welfare Weights for Optimal Tax Theory." *American Economic Review* 106, no. 1 (January 2016): 24–45.

Schmidtz, David. "The Institution of Property." *Social Philosophy and Policy* 11, no. 2 (Summer 1994): 42–62.

Schwerhoff, Gregor, Ottmar Edenhofer, and Marc Fleurbaey. "Taxation of Economic Rents." *Journal of Economic Surveys* 34, no. 2 (April 2020): 398–423.

Sobel, Russell S., and Thomas A. Garrett. "On the Measurement of Rent Seeking and Its Social Opportunity Cost." *Public Choice* 112, nos. 1–2 (July 2002): 115–36.

Steiner, Hillel. "Left Libertarianism and the Ownership of Natural Resources." *Public Reason* 1, no. 1 (2009): 1–8.

Stiglitz, Joseph E. "The Origins of Inequality, and Policies to Contain It." *National Tax Journal* 68, no. 2 (June 2015): 425–48.

———. "Pareto Efficient and Optimal Taxation and the New Welfare Economics." National Bureau of Economic Research, no. w2189 (March 1987).

———. "The Theory of Local Public Goods Twenty-Five Years after Tiebout: A Perspective." National Bureau of Economic Research, no. w0954 (August 1982).

Stiglitz, Joseph E., and Jay K. Rosengard. *Economics of the Public Sector: Fourth International Student Edition*. New York: W. W. Norton and Company, 2015.

Slemrod, Joel. "Optimal Taxation and Optimal Tax Systems." *Journal of Economic Perspectives* 4, no. 1 (Winter 1990): 157–78.

Smith, Adam. *Wealth of Nations*. London: Electric Book Company, 2000.

Tideman, T. Nicolaus. "Takings, Moral Evolution, and Justice." *Columbia Law Review* 88, no. 8 (December 1988): 1714–30.

Vallentyne, Peter. "Self-Ownership and Equality: Brute Luck, Gifts, Universal Dominance, and Leximin." *Ethics* 107 no. 2 (January 1997): 321–43.

Vallentyne, Peter, Hillel Steiner, and Michael Otsuka. "Why Left-Libertarianism Is Not Incoherent, Indeterminate, or Irrelevant: A Reply to Fried." *Philosophy and Public Affairs* 33, no. 2 (Spring 2005): 201–15.

Varian, Hal R. *Intermediate Microeconomics: A Modern Approach*. 7th ed. New York: W. W. Norton and Company, 2006.

Vickrey, William S. "Site Value Taxes and the Optimal Pricing of Public Services." *American Journal of Economics and Sociology* 60, no. 5 (November 2001): 85–96.

*Journal of Ethics and Social Philosophy*
VOL. 23, NO. 2 · NOVEMBER 2022
*https://doi.org/10.26556/jesp.v23i2.1534*


# GENDER AS NAME

*Graham Bex-Priestley*

In 2018 Theresa May, then prime minister of the United Kingdom, launched a consultation on reforming the Gender Recognition Act and moving to a process of self-identification. Instead of the lengthy and medicalized two-year process we have now, people would be able to legally change their gender by an act of mere declaration. This is already the case in other countries, such as Ireland, Portugal, and Belgium. The subsequent UK prime minister, Boris Johnson, canceled the proposed change despite a clear majority of respondents to the consultation expressing support for it, with only 36 percent in favor of requiring a diagnosis of gender dysphoria and less than 20 percent in favor of requiring a medical report.[1] Champions of the proposal focus on the practical benefits of moving to self-identification, but some believe it would also reflect the metaphysical truth that people genuinely *are* the gender they identify as. In the course of this paper, we will see that most gender theories straightforwardly rule this out. My goal is to find a theory of gender that supports it.

In section 1, I will consider what kind of project I am engaged in, suggesting that it is probably best seen as an ameliorative one, and I will clarify its aim. I will then examine Talia Bettcher's position that we should understand "First-Person Authority" (FPA) as *ethical* rather than epistemic or metaphysical in section 2, and I will argue that anything less than metaphysical FPA would leave too much room for dissent.[2] In section 3, I will show why existing gender theories, including Bettcher's and Katharine Jenkins's theories, do not adequately secure FPA.[3] I will present my own theory in section 4. I propose to understand gender in a similar manner to names. Genders, like names, have no common meanings, but they do have significance. Most importantly, our genders, like our names,

1 Jamie Wareham, "Boris Johnson 'Scraps Plans' to Make Changing Gender Easier in Blow for Trans Rights," *Forbes*, June 14, 2020, https://www.forbes.com/sites/jamiewareham/2020/06/14/boris-johnson-to-make-changing-gender-harder-in-blow-for-trans-rights/; King, Paechter, and Ridgway, *Gender Recognition Act*, 41, 47.

2 Bettcher, "Trans Identities and First-Person Authority."

3 Bettcher, "Trans Women and the Meaning of 'Woman'"; Jenkins, "Amelioration and Inclusion."

are up to us. This raises several issues that I will address in section 5, such as the worry that my theory cheapens gender, the concern that it does not support transition-related healthcare, and the interesting choice point of what to say about authority over one's *past* gender. I will conclude in section 6.

## 1. THE PROJECT

Suppose that Sam has the biological sex characteristics of a typical cis woman, such as ovaries, XX chromosomes, and relatively high estrogen. Sam has none of the biological sex characteristics of a typical cis man, such as testes, XY chromosomes, and relatively high testosterone. When it comes to other characteristics, all of the stereotypes Sam fits into belong to the category of men: Sam has short head hair and long leg hair; wears trousers; is attracted to women; is socially dominant, ambitious, egotistical, and highly sexed (Sam is a white middle-class Brit); and loves fire, science, prog rock, philosophy, and violence. Sam wonders, *What gender am I?* One of Sam's friends tells Sam, "Biology be damned. Your traits are evidence you are a man." "Pish! Archaic stereotypes be damned," a second friend replies. "You're an atypical woman." "Damn biology *and* stereotyping," says a third friend, "and damn anyone telling you what your gender is. You are whatever *you* identify as."

I am interested in the third friend's response. Underlying their position is an endorsement of the idea that there is no "golden nugget of womanness"—no shared intrinsic qualities that all and only women (or other genders) have.[4] Beyond that, there is the idea that individuals have a kind of *authority* over their own gender, and this can seem rather mysterious. Other properties, even very personal ones like sexuality, are not like this. It is not the case that I am ginger if and only if I identify as ginger. I could be homosexual without identifying as such. What could gender be for it to yield to our own authority in this way? As I will explain in section 4, I believe that we should treat genders in the same way as names. There is not much of a mystery why, when Sam says (with sincerity) "My name is 'Sam,'" we grant full authority to the individual while being under no impression that Sam must share an intrinsic quality unique to all Sams.

Am I doing metaphysics? I initially thought so. I would have described my project as a proposal for what it is to *be* a particular gender. However, recent papers by Robin Dembroff and Elizabeth Barnes have called this into question, distinguishing metaphysical theories of gender and theories that give the extensions of gender terms. Dembroff argues against "the idea that gender classifications should track the gender kind membership facts," and Barnes

4 Spelman, *Inessential Woman*, 159.

argues that "giving a metaphysics of gender should be understood as the project of theorizing what it is—if anything—about the social world that ultimately explains gender. But that project might come apart from the project of defining or giving application conditions for our natural language gender terms like 'woman.'"[5] I am not sure whether the two can come apart in the required way, but if they can, this paper is on the semantic side. Here, I am interested in what might determine the correct extension of gender terms rather than explaining why gender exists at all. Consequently, many readers may view this paper as compatible with several metaphysical theories of gender.[6] For example, Ásta's theory that genders are socially conferred upon us from context to context, imposing "constraints and enablements" on us, might correctly theorize what it is about the social world that explains gender, while my theory explains in parallel how we can correctly continue to call a trans man a "man" despite being in a context that constrains him from, say, entering the men's bathroom.[7] It would be a happy result if what I say in this paper is consistent with the excellent work being done on the social realities of gender.

My project, then, is to find a theory of gender *terms* that vindicates Sam's third friend as speaking truthfully. I think it is clear that in doing so I am not describing what most people have in mind when they use gender terms. Does this mean I am not trying to figure out the public meaning of words like "woman"? Well, it could be the case that most people are completely wrong about the meaning of the words they use, but it would take some impressive metasemantic gymnastics to arrive at the conclusion that meanings are so detached from people's conceptions and patterns of usage. Given the diversity of usage of gender terms among different communities—say, among very socially conservative groups as compared to among trans rights activists—some philosophers have taken gender terms like "woman" to be context dependent or to have multiple meanings.[8] The pertinent question is which meaning(s) we *should* adopt at any given time. However, I will not restrict myself by only looking for *existing* meanings of gender terms.

One possible project I might be engaged in is that of describing whatever properties gender terms track. "Water" does not mean $H_2O$, but our word "water" tracks what we now know is a liquid with that atomic composition. Maybe what I am doing, then, is articulating the kind of thing that people track

5 Dembroff, "Beyond Binary," 22; Barnes, "Gender and Gender Terms," 704.

6 Thanks to an anonymous reviewer for raising this point.

7 Ásta, *Categories We Live By*, 74–75.

8 Saul, "Politically Significant Terms and Philosophy of Language"; Bettcher, "Trans Women and the Meaning of 'Woman'"; Laskowski, "Moral Constraints on Gender Concepts."

with gender terms in communities subscribing to self-identification. While I am open to this idea, I think that where I end up in this paper is more akin to prescribing a meaning. Like Sally Haslanger and Jenkins, I feel I am best understood as engaging in an *ameliorative* project.[9] We might describe the project as one of conceptual engineering rather than standard analysis: "Those pursuing an ameliorative approach might reasonably represent themselves as providing an account of our concept—or perhaps the concept we are reaching for—by enhancing our conceptual resources to serve our (critically examined) purposes."[10] The gist is that our concepts are malleable and we should shape them to work for us. What, then, are my purposes?

My primary purpose is to forge gender concepts that guarantee FPA. Some may and do argue that this is the wrong goal. I will not address their arguments in this paper, save one: while opponents of self-identification worry about the supposed harms of people being able to determine their own gender, some also think that the very idea of "identifying into" a given gender makes no sense. In Bettcher's words, it is "just not obvious how trans people are going to understand the term 'woman' when they self-identify (or do not self-identify) with that term."[11] I will be arguing that my theory of gender as name *does* make sense of self-identification. Otherwise, I will be assuming that the benefits of FPA to the wellbeing of trans people and society as a whole outweigh any potential harms.[12]

It is possible that different situations call for different goals and different operative concepts. For instance, you might agree that FPA is the right goal to have in interpersonal contexts but think we should use Haslanger's account of gender as class when analyzing discrimination against women in the workplace; after all, if your boss classes you as a woman, they are likely to treat you in a certain way regardless of your hidden biology or gender identity. This paper can therefore be read as an answer to the following question: If we accept that it is at least *sometimes* correct or desirable to recognize FPA, how ought we to conceive of gender in those situations?

9 Haslanger, *Resisting Reality*; Jenkins, "Amelioration and Inclusion."

10 Haslanger, *Resisting Reality: Social Construction and Social Critique*, 386.

11 Bettcher, "Trans Women and the Meaning of 'Woman,'" 246.

12 Valentine and Shipherd examine twenty years of research about what significantly impacts the mental health of transgender and gender nonconforming people, among whom depressive symptoms and suicidality are elevated. Alleviating factors shown to be highly beneficial include access to medical intervention, employment protection, and "the central importance of a social and community support network (information and formal) that affirms one's gender identity" ("A Systematic Review of Social Stress and Mental Health among Transgender and Gender Non-Conforming People in the United States," 26).

## 2. FIRST-PERSON AUTHORITY

It is sometimes thought that we are in a privileged position of authority when it comes to our own mental states. The idea that we should think similarly about gender comes from Bettcher: "Claims about self-identity in (some) trans subcultures have the form of first-person, present-tense avowals of mental attitudes (e.g. 'I am angry at you')."[13] Yet there are different ways to understand FPA. In the case of mental attitudes, the *metaphysical* (or *ontological* or *constitutive*) thesis that identification *determines* one's attitude is highly implausible.[14] Individuals cannot simply identify themselves into a particular mental state such as anger. We might prefer instead to consider an *epistemic* FPA according to which individuals are *best placed* to know their own minds. Bettcher, correctly in my view, argues that this will not do since we are often *not* best placed to know our own minds, owing to phenomena like self-deception. Instead, Bettcher opts for an *ethical* FPA according to which we *morally ought* to treat first-person avowals as decisive. One of her examples is someone proclaiming that they want to go home.[15] To fail to take this avowal as decisive would undermine their autonomy and erode their self-confidence. According to Bettcher, we have this ethical FPA over our own genders.

I worry that this is too weak. As Bettcher acknowledges, we are not always in the best epistemic position to know our own minds. Consider a case in which our friend, visibly fuming, avows that they are not angry with us. We do not believe them. Already we are in a place many trans rights activists do not want to be when it comes to gender; ideally, we would take a friend at their word when they avow that they are genderqueer. Returning to our angry comrade, we can dispute ethical FPA. It may well be morally permissible to say to them, "I don't believe you. I can tell you're angry with me, and you're right to be after what I did to your rabbit." Of course, sometimes it will be inappropriate to deny someone's sincere avowal too. So, if gender really is analogous to mental states, ethical FPA only holds on a case-by-case basis.

13 Bettcher, "Trans Women and the Meaning of 'Woman,'" 246–47.

14 The term "metaphysical" here looks to be in tension with how it was used in the discussion of Dembroff and Barnes in the previous section. To square things up, we should take the "metaphysical" in this instance to pertain to the extensions of terms, or whatever it is that makes sentences containing the relevant terms true or false, which is still importantly different from the "epistemic" and "ethical" to be discussed in a moment. Metaphysical FPA for gender, then, need not be a matter of theorizing what it is about the social world that ultimately explains gender; it is theorizing what it is that determines the correct application of gender terms.

15 Bettcher, "Trans Identities and First-Person Authority," 99.

Furthermore, without metaphysical FPA, ethical FPA is simply not going to be convincing to anyone who gives speaking the truth greater moral weight than harmful consequences. Even if a social conservative admits that depression and suicide may follow from describing someone as a gender that person disavows, the social conservative may insist that these bad consequences do not trump the truth. Indeed, I have spoken to anti-trans activists who say that pushing this line of argument is itself immoral because it is an attempt to guilt-trip people into saying untruthful things. While it may be *polite* to treat people's first-person avowals as decisive, perhaps we are under no obligation to do so when they are false.

In light of this, I am skeptical that Bettcher has "shown that the basis for such [first-person] authority resides in the ultimate priority of ethical considerations over metaphysical and epistemological ones."[16] I think defenders of FPA over gender need to tackle the issue head-on and endorse it as a metaphysical thesis: sincere avowals of one's own gender *guarantee their own truth*. This undercuts the social conservative's position I outlined and yields an iron footing for epistemic and ethical FPA too. However, metaphysical FPA is *prima facie* mysterious.[17] How can we have such authority? Where does this power come from? Do any existing theories of gender guarantee that self-identifications are always true?

## 3. OTHER THEORIES

Most theories of gender straightforwardly contravene FPA. I will briefly consider four of these before discussing two other theories in more detail. According to purely biological sex-as-gender views, women are adult human females (where "female" is a biological sex term).[18] Someone born male cannot be a woman in virtue of a mere act of self-identification. Sam, whom we met in section 1, would be classed as a woman, and Sam's second friend would be vindicated. This biologically reductive view rules out FPA and intentionally so.

What about other, more trans-friendly theories? Consider Haslanger's view:

> *S is a woman* iff$_{df}$ *S* is systematically subordinated along some dimension (economic, political, legal, social, etc.), and *S* is "marked" as a target for this treatment by observed or imagined bodily features presumed to be evidence of a female's biological role in reproduction.

16 Bettcher, "Trans Identities and First-Person Authority," 115.

17 Henceforth, an unqualified "FPA" is to be understood metaphysically.

18 Byrne, "Are Women Adult Human Females?"; Bogardus, "Evaluating Arguments for the Sex/Gender Distinction."

> *S is a man* iff$_{df}$ *S* is systematically privileged along some dimension (economic, political, legal, social, etc.), and *S* is "marked" as a target for this treatment by observed or imagined bodily features presumed to be evidence of a male's biological role in reproduction.[19]

This view makes room for the existence of trans people if they "pass" as a target for privilege or subordination based on a mistake in perception. A looming worry about this way of thinking about gender is how it may play into the trope, which fuels so much violence against trans folk, that they are "deceivers" about their true biological role in reproduction.[20] In any case, it is clear that Haslanger's theory does not secure FPA.[21] Self-identity simply does not come into the picture and nor do nonbinary genders. If someone identifies as genderqueer but is systematically subordinated in virtue of being perceived as having biologically female features, they will be classed as a woman.

Family resemblance views do a little better.[22] There are no necessary and sufficient intrinsic features that guarantee membership of a gender category; there is no golden nugget of womanness. Instead, we could take exemplars of paradigmatic women, men, or any other genders, and then see which of them we sufficiently resemble. Resemblance is vague and there will be borderline cases, and that is a feature rather than a bug: gender is messy. Does Sam share any features with paradigmatic women such as Beyoncé and Queen Elizabeth II? Yes, biological features, but that is all. There is unlikely to be sufficient resemblance to categorize Sam as a woman, given everything else. Sam's first friend would probably be vindicated. Does self-identification come into it? Perhaps! It could be the case that paradigms of genders tend to self-identify as those genders, and so self-identification is one possible shared feature.[23] However, it is certainly insufficient by itself. In short, while family resemblance theories may *give weight* to self-identification, they will not yield FPA: people may identify as genders they do not sufficiently resemble.

Theories that view gender as a performance are very trans friendly. Judith Butler tells us it is an illusion that we have a core, inner gender that we manifest

19 Haslanger, *Resisting Reality*, 230.

20 Bettcher, "Evil Deceivers and Make-Believers."

21 Jenkins, "Amelioration and Inclusion," 398–402.

22 Fileva, "The Gender Puzzles"; Heyes, *Line Drawings*; Munro, "Resemblances of Identity"; Stoljar, "Essence, Identity, and the Concept of Woman."

23 Fileva has a two-tiered account. The first tier is procedural and is a kind of idealized self-identification view: "Under suitably idealized conditions, the person who has that gender will identify with said gender" ("The Gender Puzzles," 189–90). (It is the second, substantive tier that invokes family resemblance.) Yet we are looking for FPA here and now, under nonidealized conditions.

(or hide) through our gender expression; there is only the expression.[24] We are not bound by biology or an innately gendered mind. *Anyone* can perform, for example, womanhood, and anyone can (and should) subvert gender norms. According to Butler, gender is not really something you *are*, but something you *do*. While this does allow for great scope in the genders people can correctly identify as, it does not quite give us FPA. What do we say of someone who is pressured to perform womanhood but identifies as genderqueer "underneath"? Presumably the answer is that there *is* no underneath. Defenders of these theories of gender might say that it is very sad that this person is pressured into performing a gender they do not wish to perform, but to think that gender is anything more substantial than this performance is a mistake. Dembroff, who proposes a theory of genderqueerness as necessarily involving active resistance to the dominance of the binary gender system, echoes this sentiment: "I diverge from standard interpretations of this situation, which say that this person is 'truly' genderqueer, and unjustly prevented from self-expression. In contrast, I read the situation as one in which someone is unjustly prevented from being genderqueer."[25] Pressure and injustice need not even be part of the story. Many people who identify as (say) women *choose* to appear and behave in ways coded as other genders. Thus, self-identification is once again insufficient.

The final two theories I will contemplate here are ones that take subjective identity very seriously and consider it integral to gender categorization, which means they cannot be ruled out as straightforwardly as the previous four. First up is Jenkins's norm-relevancy theory.[26] Jenkins argues for a twin concept where we begin with the concept of gender as class—Haslanger's theory—and use it to come to the concept of gender as identity. There are two steps to the process, one objective and one subjective. The first involves identifying the social norms for people classed as men and women (which will vary depending on culture). This is the objective part since it must be based on social realities; we may believe there is a norm that women ought to regularly remove their leg hair, but we cannot pretend there is a norm that women ought to hop to work. The second step is to identify which norms you feel apply to you. These norms form an internal "map" with which to navigate the world. Importantly, you do not need to *follow* the norms you feel are applicable. People who identify as women

24 Butler, *Gender Trouble*; Butler, *Bodies That Matter*; Butler *Undoing Gender*.

25 Dembroff, "Beyond Binary," 25. Ásta writes something similar about situations where an individual's resistance to the gender that has been conferred upon them fails to secure the appropriate social recognition, leaving them stuck with an unwanted set of constraints and enablements. She compares it to the phenomenon of being silenced—an attempt to be a particular gender may "receive no uptake and remain futile" (*Categories We Live By*, 76).

26 Jenkins, "Amelioration and Inclusion"; "Toward an Account of Gender Identity."

might violate these norms, and indeed may be fully motivated to do so. For a woman, growing out her leg hair can be an intentional act of resistance, "but her experience of having hairy legs is not the same as it would be if she identified as a man: if she identified as a man she would not be conscious of violating a norm of feminine appearance, since she would not see those norms as applying to her."[27]

This view gets us much closer to FPA than before. On Jenkins's account, someone born with a typically male biology who strongly resembles paradigmatic men and who routinely performs masculinity may have a female gender identity. What of nonbinary genders? There are no nonbinary classes in Haslanger's theory, so step 1 leaves us at a loss when it comes to identifying nonbinary gender norms. Jenkins believes that people identifying outside the man/woman binary can still be explained by reference to just the two sets of norms. Here are two of her illustrative examples:

> *S* has a **genderfluid** gender identity iff *S*'s internal "map" is at times formed so as to guide someone classed as a woman through the social or material realities that are, in that context, characteristic of women as a class, and at other times formed to guide someone classed as a man through the social or material realities that are, in that context, characteristic of men as a class.
>
> *S* has an **agender** gender identity [or: *S* lacks a gender identity] iff *S* does not have an internal "map" that functions to guide them through the social or material realities that are, in that context, characteristic of any gender class.[28]

It is worth mentioning that many nonbinary people do not wish to have their genders defined only in relation to norms associated with men and women. They may take umbrage with the idea that "there are just two channels: the 'woman' channel, and the 'man' channel" on their gender "radio" from which they can choose to compose themselves.[29] From conversations with nonbinary people, I know that some are passionate about rejecting the idea that their gender exists on any kind of spectrum between, or is any function of, man and woman. Regardless, let us run with the assumption that nonbinary folks are not forging their internal maps from scratch but rejecting or riffing off what is already out there, that they are not creating new norms but mixing and matching existing ones.

Jenkins faces a trilemma with how to understand the norms in play and what it means to feel they apply to oneself: Are they expectations that others

27 Jenkins, "Amelioration and Inclusion," 411–12.

28 Jenkins, "Toward an Account of Gender Identity," 735–36.

29 Jenkins, "Toward an Account of Gender Identity," 735.

will apply the norms *to* you, are they conscious endorsements of the norms, or are they subconscious acceptances? None of the options look very nice. The first horn is a nonstarter. Trans folk fully expect to be sanctioned by society for breaking the norms of the gender on their birth certificates, but Jenkins does not indicate that expectations of external pressure determine one's gender identity, nor is that idea in the spirit of her theory.

On the second horn, we have the problem that many liberally minded folks consciously reject gender norms. People who identify as women, for instance, may say that the norm that women ought to shave their legs is totally irrelevant to them and how they choose to live. It is a bad norm, and we should pay no heed to it. We could say that all people who consciously reject gender norms are agender, but this flies in the face of FPA, which I will return to in the paragraph after next. A different response to this is to insist that these gender rebels *do* think the norms they are violating are relevant in the sense that the norms apply to them, whether they like it or not.[30] The suggestion is that the second horn should be expanded to include conscious, *nonendorsed* acceptances of application. This raises the initial question again. In what sense do these women accept that the norms apply to them, despite their lack of endorsement? They know that they are likely to incur social penalties for exposing their hairy legs—from looks of disgust to verbal abuse or worse—but that is just the first horn of the trilemma. Another idea is that they believe the norms apply to them because they identify as women, but this is circular. Perhaps instead it is because they believe society has created this set of norms for women *as class* and that the norms apply to them simply because they are members of the target group. This will not be acceptable either, first because this would exclude some trans women who are not in that target group, and second because it risks bringing gender as class and gender as identity too close together. Opting for the second horn is unsatisfactory.

We are left with the third option, which is the best interpretation of the theory.[31] Jenkins undoubtedly draws our attention to an important psychological phenomenon. Typically, a woman who shaves her legs will not notice any norm breaking, whereas a man who shaves his legs will notice he is doing something that goes against the grain. It is likely to feel jarring for him. So, while he may consciously reject the norm that men should not remove their leg hair, it may be very difficult for him to exterminate the inner red light telling him to stop shaving his legs. On the third horn, these subconscious red (and green) lights are what constitute your inner map, and if they match (a sufficient subset of)

30 Thanks to an anonymous reviewer for this point.

31 Jenkins, "Toward an Account of Gender Identity," 730–31.

the norms for the class of men and none of the norms for the class of women, then you have a male gender identity.

The trouble with this third horn is the extreme difficulty of resisting one's conditioning. Bettcher makes a similar point about how a recently transitioned trans woman might not "have much of a map to guide them through the social and material realities of being classed as a woman" while retaining much of their acquired male map.[32] Wearing "women's" clothes for the first time in public is likely to be experienced as a breaking of norms. I would go further than this: for some, the "wrong" internal maps will *never* go away. One possible response is that even if some of the "wrong" map is retained in this person's psychology, the majority of their map is that of a woman.[33] This is an important point to make, but two problems remain. First, such "mixed" maps—even when skewed unequally—would count as nonbinary on Jenkins's theory, so we still violate FPA. Second, we continue to face the fundamental issue that our maps are not *up to us*, and so we would be letting our gender identities be determined by something beyond our direct control.

Jenkins herself responds to Bettcher's worry by admitting that "the norm-relevancy account does *not* entail that everyone is always right about their own gender identity" and reiterating that she only wishes to secure *ethical* FPA.[34] But does she manage to do this? Perhaps it would be wrong to tell someone their self-identification is incorrect because that would, in effect, be telling them they do not know their own mind and what their own internal map looks like. However, the workings of the person's mind may be common ground. Somebody who identifies as a man may openly admit that much of his internal map is that of a typical woman, and he may be skeptical that he will ever be able to undo his social conditioning: the norms of femininity he now consciously rejects were drilled too deeply into him throughout his childhood. If the only reason it would be unethical to tell him that *this means he is not really a man* is that it would be rude or lead to harmful consequences, we are left with the worries I raised in section 2. At bottom, the problem with the third horn is the same as the first: it leaves the facts of the matter largely imposed on us from outside instead of determined by ourselves. Agency about our genders is undermined. If we are interested in vindicating FPA, we should look elsewhere.

The final account I will consider in this section is Bettcher's existential self-identity theory. According to Bettcher, gender is not about *what* you are but rather *who* you are: "For example, the claim 'I am a trans woman' may be an

32 Bettcher, "Through the Looking Glass," 396.

33 Thanks to an anonymous reviewer for this point.

34 Jenkins, "Toward an Account of Gender Identity," 733.

avowal of a deep sense of 'who one is' (that is, of one's deepest values and commitments). And as such, this is the prerogative of the first person alone where defensible avowals of gender are presumptively taken as authoritative."[35] Such claims do not merely describe oneself but also communicate one's "reasons for acting."[36] This is meant to secure an ethical FPA because one takes up responsibility for such avowals, and it would be an affront to challenge someone's self-interpretation and to deny that they know what they "stand for."[37]

What values, commitments, and reasons for acting are communicated by saying "I am a woman"? A commitment to speaking less during meetings and an ethics of care over justice? Of course not. Bettcher is fully aware that there are no universally shared values among women. Throughout her work, she makes clear that people have very different views and it is up to the individual to decide what their gender identification means to them: "In general, one does not know in advance what a person's reasons are for self-identifying and gender presenting."[38] The worry is that this means one is no longer communicating anything at all. I can understand how a person saying "I am a socialist" communicates their values. Even if it is a little vague whether they stand for public ownership, worker ownership, or union power, the self-identified socialist clearly is not communicating that they want to squash workers' rights. Contrastingly, there are no values, commitments, or reasons for acting we can rule out when someone identifies that they are a woman. It is unclear, then, why we ought to link gender identities with these things. Indeed, it seems *prima facie* undesirable to make these associations at all.

Importantly, the theory still does not secure FPA. We can begin by noticing that someone's identification as a socialist can be false, and it can even be ethically justified to say to a self-identified socialist that they are not really a socialist, perhaps with a nod to their voting record or their endorsement of campaigns to weaken workers' rights. Returning to Bettcher's theory, we only need to add an extra step: when someone identifies as (say) a man, we ask them which values, commitments, and reasons for acting they have associated with manhood. Suppose that to this person, being a man is a matter of being committed to war and violence. Clearly, metaphysical FPA can be violated, and we may question ethical FPA too: it might not be bad to say to this person that their commitments are far less violent than they think they are. It could be said that Bettcher's theory is only intended to secure FPA for people who *sincerely and wholeheartedly* avow

35 Bettcher, "Trans Women and the Meaning of 'Woman,'" 247.

36 Bettcher, "Trans Identities and First-Person Authority," 111.

37 Bettcher, "Trans Identities and First-Person Authority," 110.

38 Bettcher, "Trans Identities and First-Person Authority," 110.

their genders, rather than for people who lie or are simply unsure.[39] However, the not-so-violent person previously mentioned could be entirely sincere and wholehearted in their belief about who they are but misguided about their own qualities. While Bettcher secures FPA for people who know their own minds, in section 2 we saw that Bettcher herself allows that we can be mistaken about ourselves. This opens up the space for rejecting sincere avowals of people's own genders. In short, even if we like the idea that genders are extremely individualistic codes for our existential self-identity, we still do not reach the desired result of having full authority over our own genders. So, let us look at my theory.

## 4. GENDER AS NAME

I propose that we conceive of genders as we do names. I am not calling for us to identify gender terms *with* names because, grammatically, they are different parts of speech: "Josie" is a proper noun and "woman" is a common noun.[40] The theory is only that genders are *determined in the same way* as names and they mean just as much. On this account, learning there are three women in the room gives us very little information about these individuals, just as learning there are three Michaels in the room would tell us nothing more than how to refer to them. The important feature is that the bearer has the appropriate authority. Your name is, in a very real sense, up to you. FPA should be easy. If our friend makes a sincere avowal that their name is now "Raphael," we do not merely defer in virtue of the fact that it would be ethically bad not to, and we do not simply believe them because they are better placed to know their name than we are—we defer because "Raphael" is genuinely their name. No biological or psychological inspections could reveal anything non-Raphael-ish. There is nothing further to question. Likewise, on my theory, if somebody tells us he is a man with he/his/him pronouns, we refer to him accordingly. We learn nothing for certain about his biology, character traits, or values. All we learn is how to address him. Genders and names are words we use to refer to people, and we get to choose our own.

I am sidestepping the debate about whether proper names are homonymous definite descriptions, meaningless referrers, or something else. This is not because doing so avoids difficult philosophy of language (although that is nice too) but because the debate is not relevant to the social conventions about names I am alluding to. While Saul Kripke and friends are trying to figure out how on earth names manage to refer to things (their work applies equally to "Sheffield" and "Socrates"), I am interested in what makes a name *mine* or *yours*.

39 Thanks to an anonymous reviewer for this point.

40 Thanks to Matt Cull for this linguistic point of order.

Suppose Alice is bullied at school because she looks like Batman's butler. Her bullies call her "Alfred," and she hates it. "That's not my name," Alice protests, and she is right. The name "Alfred" can refer to her, but it does not *belong* to her. Kripke has a story about how "Alfred" refers to Alice, but he does not distinguish between names we are merely called and names that are our own. This distinction allows us to say that "Confucius" is not Confucius's real name.[41]

It is an interesting question how *our* names—names that belong to us—are determined. As far as I am aware, there is no philosophical literature about this. I have two hypotheses: the *endorsement* account and the *declaration* account.[42] The former comes from thinking that our real names are the ones we *want* to be called. Alice wants to be called "Alice" and not "Alfred." Not just any old desire will do, though. Desires are cheap. We might ponder several names and think "I'd quite like to be called *X*, *Y*, and *Z*," but never adopt them for ourselves. Indeed, many of us dislike our names while still accepting them. According to the endorsement account, it is a sort of stamp of approval we give to a selected name that makes it our own. The declaration account requires something a little more public. It is not enough to endorse a name privately; we must declare it in some way, whether by announcing it verbally or writing it—or perhaps less explicit means would work too, such as answering to it regularly. There is an analogy to be made here with the nuances of giving consent—it can be done in different ways, but it must be *communicated*.

Cases where an individual privately endorses one name yet publicly declares another might tease out which of the two hypotheses is preferable. I can imagine one such individual later in life saying "My name back then was 'Harold' even though I wanted to be called 'Humphrey,'" while another says "I never told anyone until I was an adult, but my real name has been 'Ria' ever since I was seven years old." I do not know which hypothesis is correct. My own intuitions are murky, and discussions with others have revealed mixed hunches. There may be better hypotheses I have not thought of too. Yet whatever the details turn out to be, I think it will be widely accepted that we have the power to choose for ourselves what name(s) we answer to. I am proposing that genders work in the same way: we have the authority to decide which genders and pronouns are ours.

For an ameliorative project to have a chance at being successful, the intended concept should not be too far removed from the current one(s). Any attempt to make "man" mean bicycle is unlikely to work. Fortunately, there are many similarities between names and genders already. We are assigned a

41 Thanks to Stephen Ingram for this example.

42 Thanks to friends, colleagues, and an anonymous reviewer for rightly pressing me to come up with proposals.

gender and a name at birth. It is possible to change them legally, yet it is not necessary for interpersonal purposes: we do not need to check someone's birth certificate or change-of-name deed before we call them what they wish to be called. There are no common intrinsic features that all and only Michaels have; there is no golden nugget of Michaelhood. Yet there are stereotypes. Even if you cannot learn anything for sure from a name or a gender, you can make educated guesses. Anyone called "Sixtus Dominic Boniface Christopher Rees-Mogg" probably did not grow up on a council estate. Likewise, you might guess that your genderqueer colleague does not vote Conservative and your male friend does not have ovaries, but you cannot rule it out. I do not think it is an unreasonably large jump from common practices to conceive of genders as names.

Some trans people tell us they have always felt like a different gender to the one they were assigned.[43] Many anti-trans activists reply that being a woman is not a feeling.[44] When Shania Twain sang "Man! I feel like a woman!" what did she mean?[45] Submissive? Oppressed? Sexy? Empowered? Gassy? Any substantive answer will be open to counterexample. In this regard, my theory seems to side with the anti-trans activists. The idea that there is a particular way that women feel makes as much sense as the idea that there is a way it feels to be a person named "George." Yet there is an interpretation that *does* make sense. People can say "I feel like a 'Gaia' more than a 'Greta'" without committing themselves to the existence of universal Gaia feelings. Some people just think one name fits them better than another. "Gaia" can *feel right*, whereas "Greta" might feel wrong or uncomfortable. (This is the same sort of language we hear from trans folk about their gender: "The category 'trans woman' might be avowed or disavowed because … it does not fit or feel right.")[46] I do not know why this is the case. Perhaps it has to do with associations we have made throughout our lives; perhaps it is an aesthetic preference, or maybe for some people it is not a feeling at all but a conscious, even political, choice. The reason does not matter. What matters is that your name is up to you.

Two problems do arise, though.[47] There is a worry that the theory of gender as name is too intermediary. For example, for anyone who feels a sense of fit

43 Almost half the respondents in a survey of trans people "cited the congruency between their inner feelings and outer appearances as a positive aspect of claiming a transgender identity" (Riggle et al., "The Positive Aspects of a Transgender Self-Identification," 150).

44 Amy Eileen Hamm, "On Feeling Like a Woman," *Feminist Current*, July 7, 2018, https://www.feministcurrent.com/2018/07/07/feeling-like-woman/.

45 "Man! I Feel Like a Woman!" track 1 on Twain, *Come on Over*.

46 Bettcher, "Trans Women and the Meaning of 'Woman,'" 247.

47 Thanks to an anonymous reviewer for raising both of these issues. I address them here instead of in the following section because they relate directly to the previous paragraph.

with a gender/name, is *that* not the important thing? Is that not the determining feature? My answer is no for two reasons. First of all, it is possible to reject a gender/name even if there is a sense of fit, and in those cases, we should respect the individual's rejection. Second, given the diversity of reasons for adopting a gender/name, it would make the theory too disjunctive. In my view, it is not the reason behind the adoption of a gender/name but the very adoption itself that is the unifying and determining feature. The other worry that arises is what we might call the *wrong-reasons* issue. Some people might feel they fit a gender for what are intuitively bad reasons, such as associations they have made that are grounded in pernicious stereotypes or unjust societal forces. Somebody may think they are a man simply because they are ambitious and attracted to women, for example. While this is *a* problem, I do not think it is *my* problem. The ameliorative project is to respect FPA. If there is trouble here, it is trouble with the goal itself. Since this paper's aim is to find a way to vindicate FPA, I think the correct thing to say is that this person *is* a man even if he ought not to be. People can be politicians for the wrong reasons, and people can be men for the wrong reasons too.

An advantage of my view is that it can easily make sense of two often overlooked classes of people. The first are those who identify as more than one gender. Dembroff cites several real examples, such as a genderqueer woman.[48] While most theories of gender would struggle to take these folks at their word, for me it is straightforward. People can have more than one gender just as they can have more than one name. This makes it possible to truly say in a room of three people that there are two women and two nonbinary folks here, and there are two Sophies and two Smiths here. The other class of people are those who have one gender but have pronouns that do not "match" it. On my view, there is nothing odd about a woman with he/his/him pronouns.[49] They are just words used to refer to people and do not indicate anything about the individual, so such mixing is no problem at all.

The biggest advantage of this theory of gender is how it undercuts the opposition to ethical FPA. On other theories, there is room for people to argue that they are under no obligation to address someone in the way that person wishes to be addressed because they simply do not believe the person is the gender that person claims to be. If genders are names, then the force of this kind of opposition is restricted only to cases of insincerity. When our friend jokingly says his name is "Rumpelstiltskin," we do not have to address him as such because

48 Dembroff, "Beyond Binary," 11–12.

49 See, for example, Jules Ryan, "Why You Should Respect He/Him Lesbians," *Medium*, November 30, 2020, https://radiantbutch.medium.com/why-you-should-respect-he-him-lesbians-85dca31a5b4f.

we know he does not really want to be called "Rumpelstiltskin." When the speaker is sincere, though, it is a rather basic rule of respect and decency that we address people as they tell us to.[50] Families need not be torn apart by parents and siblings refusing to refer to a loved one by their chosen pronouns. If we all saw genders as names, the authority of the individual over their gender would be more easily recognized.

## 5. OBJECTIONS AND REPLIES

Some may worry that my theory means gender is not real. They may insist that when we identify as a gender, we (try to) latch onto something genuine about ourselves. A quick reply is that names are real: I am really called "Graham." Yes, on my theory, genders are socially constructed—true in virtue of social practices of naming—rather than something biological or psychological. But that is the point. If you try to identify a shared golden nugget of a particular gender, you will fail to secure FPA. A helpful commenter raised the pertinent worry that if genders are equivalent to reference numbers, it feels as though we lose something important. This is true, but names are not *merely* reference numbers. It can be offensive to refer to someone using numbers instead of their name; the prisoner numbers used in Nazi concentration camps are a quintessential example of dehumanization. Self-endorsed names are far more than mere tools of reference. Conceiving of genders as names, then, does not mean conceiving of genders as inconsequential.

Jenkins raises a similar objection to a theory Bettcher mentions and that we could call the *mere self-identification* account.[51] I did not focus on this theory earlier because, as Jenkins writes, "it is not quite clear whether [Bettcher] fully endorses it."[52] The theory is minimal and could be constructed as follows: S is gender *X* iff S identifies as *X*. To identify as *X* is to have the relevant dispositions, such as being disposed to answer yes when asked if one is an *X*. Jenkins levels two objections against this theory that could also be charged against mine. The first is that "it fares very poorly at showing that gender identity is important and deserves respect. . . . Why should we care about dispositions to utter certain

50 I appreciate that there is a reasonableness constraint. Just as it may be permissible to refuse to call someone "King Underpants III," it may also be permissible to refuse to use "dumfulumfulumfelophegus" as someone's pronoun. The line between what is and is not acceptable will be fuzzy, but it is clear that existing names and pronouns in common circulation such as "Rachael" and "they/their/them" are perfectly reasonable.

51 Bettcher, "Through the Looking Glass."

52 Jenkins, "Toward an Account of Gender Identity," 727.

sentences?"[53] My theory can be seen as a fleshing out of mere self-identification in order to answer precisely this question. Names *are* important and deserve respect. People care about their names, and like with genders, some care more than others. "There is a vigorous protest when our names are mispronounced or misspelt."[54] This vigorous protest comes even though names do not have common uniform meanings. (They have *etymology*; e.g., "Graham" comes from Old English meaning gray home or gravelly homestead, but "I live in a Graham" does not mean I live in a gray home.) To quote Bruce Willis's character in *Pulp Fiction*, "I'm American, honey. Our names don't mean shit."[55] And yet they do have a quality I will call *significance*.

Consider Neo in *The Matrix*. Neo cares about being referred to as such. Agent Smith calls him "Mr. Anderson" instead, a clear analogy to the phenomenon of deadnaming.[56] If names do not have common meanings, why does Neo care? To Neo, "Mr. Anderson" *signifies* conformity. Yet he would not say that every Mr. Anderson must be a conformist. To take a more serious example, Muhammad Ali was named "Cassius Clay" at birth. His name was very important to him in part due to his change of religion. ("Cassius Clay is a slave name. I didn't choose it and I don't want it. I am Muhammad Ali, a free name—it means beloved of God, and I insist people use it when people speak to me."[57] Note that he would not say you are a Muslim man if and only if you are named "Muhammad.") People would deadname him, including the media, his parents, and even Martin Luther King Jr., leading up to his famous "What's my name?" title fight.[58] We can see very plainly that many people attach great significance to

53 Jenkins, "Toward an Account of Gender Identity," 727–28.

54 "What's in a Name?" *Guardian* (Nigeria), October 10, 2019, https://guardian.ng/features/whats-in-a-name-2/.

55 Tarantino, *Pulp Fiction*, 1:12:56–59.

56 Lilly Wachowski of the Wachowski sisters, both now out as trans women, has confirmed their film *The Matrix* is an allegory for trans issues. Adam White, "The Matrix Was a Metaphor for Transgender Identity, Director Lilly Wachowski Confirms," *Independent*, August 21, 2020, https://www.independent.co.uk/arts-entertainment/films/news/matrix-trans-metaphor-lana-lilly-wachowski-red-pill-switch-sequels-a9654956.html.

57 Qtd. in Joshua Casper, "Cassius Marcellus Clay and Muhammed Ali: What's in a Name?" *History News Network*, May 14, 2019, https://historynewsnetwork.org/article/171955.

58 Before the fight,

> Ali complained: "Why do you call me Clay? You know my right name is Muhammad Ali."
>
> Terrell didn't understand why Ali was upset. He answered plainly. "I met you as Cassius Clay. I'll leave you as Cassius Clay."
>
> "It takes an Uncle Tom Negro to keep calling me by my slave name," Ali said. "You're an Uncle Tom." (Jonathan Eig, "What's My Name? The Title Fight in

their names. If we treat genders as names, then unlike under the mere self-identification account, we can show that genders are important and deserve respect.

It is important to stress that there is a big difference between intentionally using the wrong name and intentionally misgendering someone. We afford greater protections against the latter for good reason.[59] While Ali was hurt by incidents such as the one above, the harms perpetuated against trans people are far more severe and systematic. This difference is consistent with the proposal that names and genders are determined in the same way because the harms of misgendering include not merely the violation of FPA but also a great number of other things, not least of which is the increased likelihood of experiencing violence. I do not in any way suggest that society currently divides and oppresses on the basis of proper names as it does with (real or perceived) gender. The purpose of the previous paragraphs was to show that treating genders as names does not undercut the subjective importance some people give to their own genders, nor does it undercut the respect we give to other people's gender identities.

Some may insist that the idea that genders are determined like names undermines their lived realities. "My body dysmorphia is real and is what makes me a woman," someone may tell me. My reply, if it would not be too hurtful at the time, would be that body dysmorphia is neither necessary nor sufficient to be a woman. It sounds pedantic (because it is), but perhaps this person's body dysmorphia *caused* her to identify as a woman. The relationship is causal, not constitutive. And this does not make it any less important. A possible retort is that the causal relationship is sometimes the other way around. For some, it is not that they desire (say) top surgery and this desire causes them to identify as a man. Instead, they already identify as a man and desire gender-confirmation surgery as a result. I agree that this happens too. The "initial" reason to identify as a given gender may be dysmorphia, or it may be something else entirely. All of this is compatible with my theory.

From here we arrive at Jenkins's second objection to Bettcher's mere self-identification account, which also applies to my view, with respect to the need for trans-related healthcare (TRH). Just as "it is difficult to perceive any relationship at all between a linguistic disposition and the sort of felt need for one's body to be different that would prompt the desire to access transition-related healthcare," it is difficult to perceive a relationship between names and access to such

---

Which Muhammad Ali Asserted His Identity," *Undefeated*, June 4, 2016, https://theundefeated.com/features/whats-my-name/)

Note that the insult "Uncle Tom" is "just" a name too, but it has extreme social significance. Other names with (less extreme) shared public significance are "Becky," "Chad," and "Karen."

59 Thanks to an anonymous reviewer for this point.

healthcare.[60] My response is twofold, looking first at purely moral justifications and second at pragmatic issues of persuasion.[61] What morally justifies access to TRH? Fundamentally, it is the great benefit to the patient and the great harm it helps to prevent. That trans people are typically the people who require this healthcare is only contingent (on societal attitudes, pressures, and associations made between genders and bodily appearances/functions). This fact commits me to the view that if a woman has body dysmorphia and wishes to remove her breasts as a result, then she is just as entitled to healthcare access as a trans man who desires the same surgery. There are complicating factors, of course, since for many trans folk it is not *only* distress and discomfort with one's body that matters but other things too, such as fitting in with the rest of one's gender. Still, in principle I am happy to bite the bullet: if two people would genuinely experience a similar benefit from surgery and they would face a similar amount of harm in being denied it, their entitlement is the same regardless of their genders.

Real life is different. If a person requested surgery on the basis of affirming their name, they would not be treated seriously. Consequently, if the medical profession shifted toward thinking of gender as name, perhaps this would lead to practitioners taking gender-affirming healthcare (even) less seriously than now. This would be awful. If gender is name, how could we continue to press the case for access to TRH? First of all, we can fall back on what I take to be the genuine moral justification: it is beneficial to the patients. TRH saves lives. If that is not enough, there are two other things to try. One is to utilize the distinction discussed in section 1 between metaphysics and terminology. Gender *terms* are determined like names, but perhaps this is consistent with other philosophers' metaphysical theories of gender. If so, the case for TRH can be made on the basis of those social realities instead of on the basis of application conditions for gender terms. The other way to press for TRH involves acknowledging the difference between gender itself and the way society operates. An analogy here is with racism and race, where academic theory often comes apart from public attitudes. Even if we discover that race does not exist at all, this does not imply racism does not exist. Likewise, if gender is "only" a name, this does not mean people will automatically be accepted as the genders they are, and TRH can help in this regard.[62] In general, racial and gendered (and other) oppression and

60 Jenkins, "Toward an Account of Gender Identity," 728.

61 Thanks to an anonymous reviewer for pressing me on this and for suggesting what is now the second part of my response.

62 For example, Britons tend to support transgender people using facilities for their gender *unless* it is specified that they have not "undergone gender reassignment surgery," in which case Britons tend to oppose (Matthew Smith, "Where Does the British Public Stand

hardships go on, even when (and often especially when) based on race-related or gender-related falsehoods.

The final objection is an interesting one, and my reply will take us on a brief journey through the metaphysics of time and the possibility of backward causation about social facts. The objection is that my theory cannot accommodate a common phenomenon: some trans people do not merely identify as a gender from the present onwards, but avow that they *always were* that gender, even before identifying as such. The objection comes from correspondence with Will Gamester, and the phenomenon is described by Bettcher when discussing the mere self-identification account: "Admittedly, this means trans women who don't yet self-identify as women aren't yet women (in this sense). That said, once she does self-identify as a woman, she may well re-assess her entire life by saying she's always been a woman (something we should respect too)."[63] This is a curious thing. Names do not quite work this way, but a similar social norm is in play when it comes to backward reference. It is a standard rule of etiquette that we refer to, say, Muhammad Ali's early life in the way I do in this sentence—by using the name "Muhammad Ali" even though he was named "Cassius Clay" in his early life. It is easy to see why we do the same (unless instructed otherwise) with gendered pronouns. However, it is natural to say "Muhammad Ali's name as a youngster was 'Cassius Clay,'" which does not map onto the gender case as many would desire. We now come to a choice point, and I will end by describing the three routes we could go down. Since I do not know which route is best, I will leave the choice to you.

1. *Embrace it*. Like the defenders of performance theory in section 3, we could take a hard line toward the recently transitioned trans woman who wishes to reassess her life as always having that gender. "I'm sorry but you weren't a girl," we might say, "and if you think you were, you're reading too much into gender. It's just a name, and you've changed it." Again like the defenders of performance theory, we can add that it is very sad and unjust that the trans woman did not get to change her gender earlier. This hard-line option still gets us *present* FPA. However, if we want to respect *retroactive* FPA, we will have to abandon the analogy with names when it comes to how gender is determined across the board. We will have to make some new rules about how genders work when determining one's past. I suggest another analogy: annulment. On one way of looking at it, having a marriage annulled does not simply add a new social fact; it erases an old one. After an annulment, the marriage is *null and void*, meaning it did not take

on Transgender Rights?" YouGov, July 16, 2020, https://yougov.co.uk/topics/politics/articles-reports/2020/07/16/where-does-british-public-stand-transgender-rights).

63 Bettcher, "Through the Looking Glass," 396.

place. This view of annulment is that the act engenders backward causation, preventing a legitimate marriage from occurring in the past. Perhaps genders could work in the same way. If you now decide you are genderqueer, you may (or may not) choose to "annul" your previous gender and say you were never anything but genderqueer all along. And there are two different ways to interpret this.

2. *Future realism*. If the future exists and is "set," then any future annulments are already out there, preventing their respective marriages from occurring. Thus, if Baldrick in the year 1534 said "Henry VIII is married to Anne Boleyn," he was speaking falsely: it was never true that a legitimate marriage took place. Likewise, if a trans woman decides as an adult that she was always a girl growing up, then when she was growing up, she was wrong to refer to herself as a boy. This may sit very well with many trans folks. One potential downside is that it allows people to correctly disagree with a person's avowal. Young people who say they are transgender are often met with dismissal. "It's just a phase," their parents say. "They're not really transgender. They'll grow out of it." And indeed, for some it is just a phase (not that phases are bad). According to *future realism*, the parents may well be correct, since the future child could declare that they were never transgender at all. There is always the epistemic possibility that our future self will reinterpret our current gender. The final authority, on this picture, comes from the individual at their oldest, when they have the ability to determine the gender of all their past time slices. This is still a version of FPA, and we might call it *deathbed FPA*.

3. *Complete FPA*. We may subscribe to a view of backward causation according to which Baldrick spoke truly in 1534 when he said that Henry and Anne were married, but anybody who *now* says that Henry and Anne were married is speaking falsely.[64] Baldrick spoke truly because in 1534 the annulment had not happened, and so, back then, nothing was preventing the legitimacy of the marriage. Once the annulment had occurred, though, it was no longer the case that their marriage was legitimate. This does lead to some strange sentences, such as "Baldrick spoke truly when he said that Henry and Anne were married, but Henry and Anne were not married." Nevertheless, perhaps by using crafty subscripts, we can make sense of it.[65] Applying this to gender, where does this view of retroactive social facts take us? It yields the ultimate version of FPA: at any point in time, you are in complete authority over your current and past gender. A child sincerely avowing "I am a boy" is correct at the time, and their older self

64 This view of the changing past is endorsed by Barlassina and Del Prete, who argue that the proposition that *Lance Armstrong won the year 2000's Tour de France* used to be true but no longer is ("The Puzzle of the Changing Past").

65 The easiest way may be to make truth time relative: the proposition that *Henry and Anne were married in 1533* is true-at-1534, and it is false-at-1537.

who declares "I was never a boy" is correct also. This option gives us the benefits of the two previous options. Like the first option, you have *present* FPA. You are not held hostage to your future self; others cannot dismiss your present gender identity on the basis of a correct prediction that you will retroactively change it in the future. And like the second option, you have the power to reinterpret your earlier life as you see fit. Perhaps you have always been a particular gender all along, or perhaps your gender was not always fixed. It is up to you. It is up to you right now, it will be up to you tomorrow, and it was up to you in the past. On this picture, at every point in your life, your authority is absolute.

## 6. CONCLUSION

I have argued that we should treat genders as names in the sense that they are up to us, indicate nothing for certain about the bearers beyond how to refer to them, and yet often have strong personal significance. My argument rests on two big starting points: that ameliorative projects are feasible and that we ought to respect an individual's authority over their own gender, at least in interpersonal contexts. I hope to have shown that there is a way in which self-identification makes sense, is true, and yet does not make gender entirely empty. If I am right, then the debate really comes down to those starting points. If we have the power to choose what concept of gender to use going forward, what work do we want it to do? As I mentioned in section 1, we might have different priorities in different contexts. Structural oppression may best be analyzed using wholly different concepts; some people may need a safe space away from folks who resemble paradigmatic men; and we certainly should never lose sight of biology-based issues such as abortion access and tampon taxes. Yet when it comes to interpersonal contexts, I feel strongly that the balance of reasons weighs in favor of calling people what they wish to be called. Conceptual stubbornness will be an obstacle, but a great deal of hardship can be avoided if we can learn to be flexible and think of gender as name.[66]

*University of Leeds*
*g.bex-priestley@leeds.ac.uk*

66 This paper could not have been written without the help of friends, colleagues, and people from the trans and nonbinary community. Many thanks to Novenka Bex-Priestley, Gabriela Arriagada Bruneau, Matt Cull, Will Gamester, Stephen Ingram, Andy Kirton, Natasha McKeever, Christina Nick, Rich Rowland, Rebecca Schorsch, two anonymous reviewers (one of whom has since revealed themselves to be Nick Laskowski), and an audience at the University of Leeds for extensive discussions and feedback on previous drafts.

## REFERENCES

Ásta. *Categories We Live By: The Construction of Sex, Gender, Race, and Other Social Categories*. New York: Oxford University Press, 2018.

Barnes, Elizabeth. "Gender and Gender Terms." *Noûs* 54, no. 3 (September 2020): 704–30.

Barlassina, Luca, and Fabio Del Prete. "The Puzzle of the Changing Past." *Analysis* 75, no. 1 (January 2015): 59–67.

Bettcher, Talia Mae. "Evil Deceivers and Make-Believers: On Transphobic Violence and the Politics of Illusion." *Hypatia* 22, no. 33 (Summer 2007): 43–65.

———. "Through the Looking Glass: Trans Theory Meets Feminist Philosophy." In *The Routledge Companion to Feminist Philosophy*, edited by Ann Garry, Serene Khader and Alison Stone, 393–404. New York: Routledge, 2017.

———. "Trans Identities and First-Person Authority." In *You've Changed: Sex Reassignment and Personal Identity*, edited by Laurie Shrage, 98–120. Oxford: Oxford University Press, 2009.

———. "Trans Women and the Meaning of 'Woman.'" In *Philosophy of Sex: Contemporary Readings*, 6th ed., edited by Nicholas Power, Raja Halwani, and Alan Soble, 233–50. Lanham, MD: Rowan and Littlefield, 2013.

Bogardus, Tomas. "Evaluating Arguments for the Sex/Gender Distinction." *Philosophia* 48, no. 3 (July 2020): 873–92.

Butler, Judith. *Bodies That Matter: On the Discursive Limits of Sex*. New York: Routledge, 1993.

———. *Gender Trouble: Feminism and the Subversion of Identity*. New York: Routledge, 1990.

———. *Undoing Gender*. New York: Routledge, 2004.

Byrne, Alex. "Are Women Adult Human Females?" *Philosophical Studies* 177, no. 12 (December 2020): 3783–803.

Dembroff, Robin. "Beyond Binary: Genderqueer as Critical Gender Kind." *Philosophers' Imprint* 20, no. 9 (April 2020): 1–31.

———. "Real Talk on the Metaphysics of Gender." *Philosophical Topics* 46, no. 2 (Autumn 2018): 21–50.

Fileva, Iskra. "The Gender Puzzles." *European Journal of Philosophy*, 29, no. 1 (March 2021): 182–98.

Haslanger, Sally. *Resisting Reality: Social Construction and Social Critique*. Oxford: Oxford University Press. 2012.

Heyes, Cressida. *Line Drawings*. Ithaca: Cornell University Press, 2000.

Jenkins, Katharine. "Amelioration and Inclusion: Gender Identity and the Concept of Woman." *Ethics* 126, no. 2 (January 2016): 394–421.

———. "Toward an Account of Gender Identity." *Ergo* 5, no. 27 (September 2018): 713–44.

King, Daniel, Carrie Paechter, and Maranda Ridgway. *Gender Recognition Act: Analysis of Consultation Responses*. Government Equalities Office, 2020. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/919890/Analysis_of_responses_Gender_Recognition_Act.pdf/.

Laskowski, Nicholas. "Moral Constraints on Gender Concepts." *Ethical Theory and Moral Practice* 23, no. 1 (February 2020): 39–51.

Munro, Vanessa. "Resemblances of Identity: Ludwig Wittgenstein and Contemporary Feminist Legal Theory." *Res Publica* 12, no. 2 (June 2006): 137–62.

Riggle, Ellen, Sharon Rostosky, LaWanda McCants, and David Pascale-Hague. "The Positive Aspects of a Transgender Self-Identification." *Psychology and Sexuality* 2, no. 2 (February 2011): 147–58.

Saul, Jennifer. "Politically Significant Terms and Philosophy of Language." In *Out from the Shadows: Analytical Feminist Contributions to Traditional Philosophy*, edited by Sharon Crasnow and Anita Superson, 195–216. Oxford: Oxford University Press, 2012.

Spelman, Elizabeth. *Inessential Woman: Problems of Exclusion in Feminist Thought*. Boston: Beacon Press, 1988.

Stoljar, Natalie. "Essence, Identity, and the Concept of Woman." *Philosophical Topics* 23, no.2 (Autumn 1995): 261–93.

Tarantino, Quentin. *Pulp Fiction*. A Band Apart and Jersey Films, 1994.

Twain, Shania. *Come On Over*. Mercury Records, 1997.

Wachowski, Lana, and Lilly Wachowski. *The Matrix*. Warner Bros. and Village Roadshow Pictures, 1999.

Valentine, Sarah, and Jillian Shipherd. "A Systematic Review of Social Stress and Mental Health among Transgender and Gender Non-Conforming People in the United States." *Clinical Psychology Review* 66 (December 2018): 24–38.

*Journal of Ethics and Social Philosophy*
VOL. 23, NO. 2 · NOVEMBER 2022
*https://doi.org/10.26556/jesp.v23i2.1602*


# MORAL VAGUENESS AND EPISTEMICISM

*John Hawthorne*

EPISTEMICISM is one of the main approaches to the phenomenon of vagueness. But how does it fare in its treatment of *moral* vagueness? This paper has two goals. First, I shall explain why various recent arguments against an epistemicist approach to moral vagueness are unsuccessful. Second, I shall explain how, in my view, reflection on the sorites can inform normative ethics in powerful and interesting ways. In this connection, I shall be putting the epistemicist treatment to work, engaging with a family of somewhat neglected issues concerning continuity that lie at the interface of metaphysics and ethics.

Section 1 introduces epistemicism as well as a competing view—"classical magnets"—that will be helpful for later discussion. Section 2 address a variety of arguments against epistemicist treatments of vagueness in ethics, including Miriam Schoenfield's appeal to the irrelevance of linguistic anthropology to ethics, Tom Dougherty and Cristin Constantinescu's concerns about unknowable moral truths, and a proportionality argument leveled by Constantinescu.[1] Section 3 precisifies an interesting but underexplored cluster of continuity issues in the vicinity of the proportionality idea, and examines them through an epistemicist lens.

## 1. EPISTEMICISM AND CLASSIC MAGNETS

### *1.1. Epistemicism*

I will begin with a brief sketch of epistemicism. I shall present the version articulated by Timothy Williamson, who has done the most to popularize the view.[2] A somewhat barebones version of Williamsonian epistemicism will be adequate to the dialectical purposes of this paper.

1 See Schoenfield, "Moral Vagueness Is Ontic Vagueness"; Dougherty, "Vague Value"; and Constantinescu, "Moral Vagueness."

2 See Williamson, *Vagueness*. Interestingly, different versions of epistemicism are presented by Sorensen, *Vagueness and Contradiction*; and Kearns and Magidor, "Epistemicism about Vagueness and Meta-Linguistic Safety." I shall not be discussing their comparative merits and detractions here.

Here are the key features of the epistemicist treatment of borderline cases. First, excluded middle holds in borderline cases. (Epistemicism is one of many theories of vagueness that operates within a classical propositional logic.[3]) Supposing, for example, that I constitute a borderline case of being happy. Then:

> Either I am happy or it is not the case that I am happy.

Second, bivalence holds in borderline cases.[4] Thus:

> "I am happy" is true or "I am happy" is false.

Third, borderline cases beget ignorance among humans, left to their own devices. Thus if I am a borderline case of being happy then humans are not in a position to know whether I am happy.

Some important points of clarification are in order: I say "among humans" since it is no part of this brand of epistemicism that it is impossible in principle for *any* creature to know the answer in a borderline case:

> On the epistemic view, vague utterances in borderline cases are true or false and we humans have no idea how to find out which. It is quite consistent with this view that what is a borderline case for us is not a borderline case for creatures with cognitive powers far greater than any we can imagine.[5]

Epistemicism thus allows that borderline cases can be known by superbeings.

I say "left to their own devices" because epistemicism also presumably allows *us* in principle to know the truth value of borderline cases by relying on

3 By this I mean that the epistemicist accepts as true any sentence that is certified as true by the standard truth-table method and accepts as truth preserving any argument that is certified as valid by the standard method of truth tables. Examples of other approaches that endorse classical propositional logic (though this list is far from exhaustive) are the supervaluationism of Fine ("Vagueness, Truth and Logic") and Bacon's version of the view that vagueness resides in propositions and not language (*Vagueness and Thought*). These views also accept the standard inference rules for the existential and universal quantifiers. Not all of these approaches accept bivalence, however. For example, versions of supervaluationism that identify truth with supertruth will accept as supertrue any sentence of the form "*x* is bald or *x* is not bald" but will not always accept sentences of the form "'*x* is bald' is true or '*x* is bald' is false." The standard way of certifying the latter class as true relies on disquotational principles that will not be accepted by the supervaluationist who goes in for the identification of truth with supertruth.

4 That is, bivalence holds when a borderline case arises on a particular occasion of use. A single, context-dependent sentence may express a truth on one occasion and a falsehood on another, and can be borderline true on one occasion and non-borderline true on another.

5 Williamson, *Vagueness*, 212.

the testimony of superbeings (we might rely on their testimony because we are impressed enough by the performance of the superbeings in areas that we do know about). The relevant ignorance thesis is that the ordinary discriminative and intellectual resources of human beings afford no path to knowledge in borderline cases.

Further, it is tempting to say that if an utterance of "I am happy" is borderline, then, according to epistemicism, the fact it expresses is unknowable. But here we need to be very careful. By way of illustration, let us introduce a name, "Roger," with the stipulation that "Roger" picks out the number 1 if I am happy and 0 otherwise. The sentence "Roger is 1" is borderline in the way "I am happy" is borderline. But supposing I am happy, it follows that "Roger" refers to 1 and thus arguable that "Roger is 1" expresses the fact that 1 is 1. But that fact is not at all difficult to know! The issue here is that unless we operate with an extremely fine-grained notion of facts, then a fact can simultaneously be expressed by a borderline sentence and also by a precise sentence. In a case like this there will, so to speak, be a blockage to knowing the fact under the guise of the borderline sentence but there may be no similar blockage to knowing the fact under the more precise guise. Reserving "vague" and "precise" as predicates of representations, Williamson introduces the predicates "vague*" and "precise*" as predicates of objects, properties, and relations, so that a vague* object will be one that is picked out by a vague expression and precise* object picked out by a precise expression. He then points out that

> The vague description "the greatest prime number much less than 100" and the precise description "the prime number between 72 and 78" might both refer to 73 in a given context. Thus 73 would be both vague* and precise*.[6]

As he makes clear, similar points potentially apply to properties, relations, and facts themselves.[7]

Fourth, not all ignorance is ignorance due to vagueness. Moreover, not all irremediable ignorance is ignorance due to vagueness. For example, certain microphysical or mathematical questions may be deeply elusive but it need not follow that the questions are vague. Williamsonian epistemicism has a distinctive story to tell about the *source* of ignorance in the case of vagueness. The source is semantic plasticity. If a borderline sentence expresses a true proposition, then there is a false proposition that could very easily have been expressed

6 Williamson, *Vagueness*, 258.

7 Of course, this taxonomy allows that there are facts that are vague* but not precise*—in such cases, the only guises under which it is humanly possible to represent the fact are vague guises.

by that sentence on account of tiny differences in use. In this case, the sentence is *true but not determinately true*. And if a borderline sentence expresses a false proposition, there is a true proposition that it could very easily have expressed. In this case the sentence is *false but not determinately false*.[8]

I shall elaborate on some further aspects of epistemicism as needed in the discussion that follows.

### *1.2. Classical Magnets*

Tom Dougherty juxtaposes the Williamsonian explanation of moral vagueness with a second picture. According to this second picture, moral predicates pick out natural kinds, each of which serves as a "reference magnet."[9] Small differences in use would not induce a shift in reference because moral predicates referentially gravitate to these properties. Even if other properties "fit our use" a bit better, the predicates refer to the natural kinds because they are much more natural than other properties with similar extensions. He takes the category of being an orangutan as an analogy:

> Since we are assuming that there is a natural biological kind, and our usage of the term "orangutan" comes close to picking it out, this natural kind becomes the referent of the word. In this way, a natural kind can act as a "reference magnet" for a term. Let us suppose for now that there is a unique set of things that constitutes the natural kind, orangutan. (Later we will discuss the view that there are overlapping but distinct sets that are equally natural as kinds.) Now, assuming we reject a metaphysical view of vagueness, this natural kind has a precise boundary: everything is in this set or it is not. Therefore, our use of the word "orangutan" would pick out a precise set of things.[10]

8 Epistemicists also often introduce determinacy operators like, "It is determinate that," though (as explained in Fritz, Hawthorne, and Yli-Vakkuri, "Operator Arguments Revisited"), such operators threaten to behave like Kaplanian monsters. Assume for example that "Roger is identical to 1" and "1 is 1" express the same proposition and that we want to say: Determinately 1 is 1 and Not Determinately Roger is 1. Then the determinacy operator will not interact with quantifiers in the usual way. For example, we cannot reason from "Not (Determinately Roger = 1" to "$\exists x$ ($x$ is Roger and Not (Determinately $x = 1$))."

For critical discussion of a range of subtleties arising from the semantic plasticity idea, see Hawthorne, "Epistemicism and Semantic Plasticity"; and Yli-Vakkuri, "Epistemicism and Modality."

9 Dougherty, "Vague Value," 357. For further discussion of possible applications of reference magnetism in the domain of ethics, see Dunaway and MacPherson, "Reference Magnetism as a Solution to the Moral Twin Earth Problem."

10 Dougherty, "Vague Value," 357–58.

The discussion here needs a bit of cleaning up. First, assuming we buy into classical propositional logic (a commitment that, as I have said, is common to many approaches to vagueness, not just epistemicism), the difference between the reference magnetism picture and the semantic plasticity picture is not aptly captured by such claims as "Everything is an orangutan or it is not." After all, everything is bald or it is not. Given classical propositional logic, whether a term is shifty or stable has nothing to do with an excluded middle. Further, as we have seen, the primary uses of "vague" and "precise" are to representations. We should not be using ideology like "precise set" without saying what that means. Now, as we have seen, the epistemicist has "precise*" and "vague*" at their disposal—but in this sense a set can be both precise* and vague*.

Moreover, we should think of referents of moral predicates as properties and not extensions (i.e., the sets of things that predicates are true of). On the natural way of developing the magnets picture, "acted permissibly" would have expressed the same property at nearby worlds. But it need by no means have the same extension. Suppose for example that someone behaves permissibly at a dinner party at this world but not at a nearby world. That will suffice to induce a shift in extension of "acted permissibly" across worlds. But this is not the kind of shiftiness that interests the epistemicist.[11]

Finally, one should not simply assume that insofar as a moral predicate picks out a natural kind, then, even assuming there is no other natural kind with a similar extension, it *follows* that small differences in use of that predicate will be semantically sticky across the modal neighborhood. Consider, for example, a modal path from a world where "is morally good" picks out the property of being morally good to a world where "is morally good" expresses the property of being made of gorgonzola cheese, and where adjacent worlds on the path are almost the same in the distribution of microphysical properties. No matter how powerful a reference magnet moral goodness is, there will be a pair of adjacent worlds along the path, one at which "morally good" picks out moral goodness, and one at which it does not. (Arguably the extension will shift radically between that pair—use will become gradually more and more anomalous vis-à-vis the behavior of moral goodness so that eventually there is such a great mismatch that reference will shift quite dramatically to something much less anomalous. Of course I am not saying that it will jump straight from the property of moral goodness to the property of being made of gorgonzola.) The kind of reference magnetism afforded by natural kinds is one that may secure a good deal of stability but will not in general preclude the plasticity phenomenon and thus will not in full generality preclude the

11 Relatedly, I find the talk of sets as "constituting" natural kinds rather odd.

kinds of shiftiness that Williamson envisages. If the kind picked out fits use just well enough to count as the referent but still badly enough to not fit use well enough at close worlds to count as the referent, then the term will be shifty in its referent.

That said there is a picture implicit in Dougherty's sketch (albeit a very speculative one). One can think that a certain moral predicate, say "being morally permissible," picks out a natural, "jointy" property and be anti-skeptical enough to think that we are not in one of the edge cases described in the last paragraph, so that our local modal neighborhood is one in which "is morally permissible" refers to exactly the same property no matter which world from the local neighborhood is actualized. Let us call this view the "classical magnets" view of a moral predicate (where "classical" serves as a reminder that the background propositional logic is classical): we might combine classical propositional logic and bivalence with the idea that at least some moral predicates correspond to highly natural properties, which they stably refer to across nearby worlds.[12] Such predicates are not semantically plastic.

What bears emphasis is that, from the point of view of Williamsonian epistemicism, this is *not* (*pace* Dougherty) a competing explanation of moral vagueness. For given Williamsonian epistemicism, this is a view on which the relevant moral predicates are not vague at all. For recall that not all irremediable ignorance is ignorance due to vagueness. Consider phenomenal consciousness. There may be certain creatures that humans are unable, left to their own devices, to recognize as conscious even though they are. But that does not mean that the question of whether such creatures are conscious is vague. Many of us will be inclined to think that there is a fundamental phenomenon here—being phenomenally conscious—and our ignorance is simply a matter of not having the epistemic tools to probe whether the phenomenon is exhibited by the creature in question. Here a classical magnets view is tempting—it is not unnatural to think that we philosophers have done enough to see to it that "being phenomenally conscious" locks on to a kind, and that tiny differences in use *would not* have induced a semantic shift. I do not care at the moment whether this picture is correct. What I do wish to emphasize is that *if* it is correct, then insofar as ignorance due to vagueness is rooted in semantic plasticity, we should not classify our irremediable ignorance about whether a certain creature is conscious as ignorance due to vagueness.

12 A variant that I shall mention but not discuss holds that the relevant moral properties are magnetic in the sense of being easy to refer to but denies that they are highly natural and thus takes such properties to be a counterexample to Lewis's idea that naturalness correlates with easiness to refer to.

The case of moral vagueness is no different. Let us take what seems like a paradigm case of moral vagueness, one that is presented by Schoenfield:

> *Diversions*: Darryl is watching his two year old daughter play in a city park. It is permissible to divert his attention from her for 1 second. It is not permissible to divert his attention from her for 5 minutes. Is it permissible to divert his attention for 30 seconds? 31? 32? Plausibly, we can create a Sorites series, admitting of borderline cases of permissibility, out of a series of diversions whose lengths differ by a second.[13]

This certainly *seems* like a case of vagueness. But if we fully espouse the classical magnets view for "is morally permissible," it is by no means clear that we should classify the case in this way. Now of course the sentence, "It is permissible for Darryl to divert his attention for twelve seconds," may be vague for reasons having nothing to do with moral predicates: after all, there may be actions that are borderline cases of diverting attention. But assuming classical magnets is the right view for "permissible" but not for "bald" and "heap" and the other paradigms of vague predicates, it does not seem, on reflection, that the Williamsonian should see this as a case where "It is permissible that" is vague. Rather, assuming the metaphysically ambitious picture of permissibility encoded by a classical magnets view, this seems relevantly similar to the consciousness case. Just as we cannot discern the length of a rod to the nearest nanometer given our limited discriminatory perceptual capacities, so we cannot see the distribution of the special permissibility property given our limited discriminatory intellectual and discriminatory capacities. Moreover, if there is a single highly magnetic property of propositions picked out by "It is permissible that," then it does not seem likely that small shifts in use will induce semantic shifts and so the case will not have the semantic plasticity that Williamsonian epistemicism requires for ignorance due to vagueness. What may have seemed like ignorance due to vagueness thus may turn out, given the classical magnets view, to be plain old ignorance.

Two final points of clarification. First, while the Williamsonian version of epistemicism contends that ignorance due to vagueness is rooted in semantic plasticity, I leave open the possibility of a version of epistemicism that locates the distinctive source of ignorance due to vagueness elsewhere. And for all I have said, such views may offer a different take on how classical magnets relate to vagueness. Here is not the place to explore in detail what other epistemicist stories might be available about the distinctive source of ignorance due to vagueness.[14]

13 Schoenfield, "Moral Vagueness Is Ontic Vagueness," 262.

14 I note though that another well-known version of epistemicism on the market does not fit well either with the view that moral predicates are both vague and correspond to classical magnets. Here I have in mind Roy Sorensen's version of epistemicism, according to

Second, I note that, later in his paper, Dougherty presents a view that relaxes the assumption that moral predicates are magnetized to natural kinds:

> Instead there may be multiple natural extensions that are on a par with respect to naturalness.... One might think that the term "orangutan" importantly picks out something special about the metaphysical structure of the world, independently of how we represent it. But one might deny that there is a single precise set of creatures that forms a natural kind. Instead, there may be multiple equally natural sets of creatures that slightly vary in their membership.[15]

Adapted to the language of properties, and transposed to the moral case, the idea would be that there is a cloud of very natural properties in the vicinity of permissibility, none of which "is permissible" stably refers to.[16] *This* way of trying to do justice to the idea that permissibility is metaphysically special, one that denies the stability thesis of the classical magnets view, is of course much more friendly to epistemicism.

## 2. Some Anti-Epistemicist Arguments

### *2.1. Moral Vagueness and Linguistic Anthropology*

Miriam Schoenfield argues for the following thesis:

> *Central Thesis*: If a robust form of moral realism is true, and there is moral vagueness, then it is ontic vagueness.[17]

---

which the distinctive feature of vague truths is that they are truths without a truthmaker (*Vagueness and Contradiction*). Classical magnets seem paradigmatically well suited to provide truthmakers. A referee wondered whether a highly natural cutoff on a scale of moral worth might reasonably be classified as vague. Here again the crucial issue is what the relevant epistemicist story is about the source of ignorance due to vagueness as opposed to ignorance not due to vagueness. If the story about what constitutes ignorance due to vagueness is Williamsonian, then insofar as that natural cutoff induced a failure of semantic plasticity, the ignorance would not count as ignorance due to vagueness. And if the story is Sorensen's then insofar as that natural cutoff provided a truthmaker, then once again the ignorance would not count as ignorance due to vagueness.

15 Dougherty, "Vague Value," 364–65.

16 Here, "is permissible" expresses a property of properties (i.e., action types), "It is permissible that" expresses a property of propositions (a propositional operator), and "acted permissibly" expresses a property of people. I shall not fuss about which member of this family is more fundamental.

17 Schoenfield, "Moral Vagueness Is Ontic Vagueness," 259.

By "moral realism" she tell us that she means the view that "moral truths are necessary" and that they are "part of the deep underlying metaphysical structure of the world."[18] By "moral vagueness" she has in mind the thesis that moral predicates are vague. Focusing on moral permissibility, she presents a series of examples that constitute a *prima facie* case for moral vagueness. We have already looked at one of them—Diversions.

Let us suppose this situation described in Diversions obtains and the following claim is true:

> It is permissible for Darryl to divert his attention from his daughter for one second.

Is this a moral truth? One would have thought so except that Schoenfield has told us, on behalf of the moral realist, that she intends to restrict "moral truth" to "necessary truths." The proposition expressed by the above claim is not necessary, since it would be false in certain circumstances where Darryl learned that his daughter will be tortured if he diverts his attention for a second or more. Now of course there are necessary truths in the neighborhood. Such necessary truths include something of the form:

> If *c* then it is permissible for Darryl to divert his attention from his daughter for one second

where "*c*" is a placeholder for an enormously complicated description of the underlying physical facts of the scenario in question (together with phenomenal facts if one thinks of those as a metaphysical add-on).

Necessary truths in the vicinity also include:

> Actually, it is permissible for Darryl to divert attention from his daughter for one second.

(Here I am using "actually" in the way that is standard in philosophy:[19] it has a rigidifying effect, so that if "*S*" is in fact true then "Actually *S*" is necessarily true. We can think of claims that a proposition is actually the case as a claim that *at the actual world* that proposition is the case.) But I do not see much point in restricting the category of moral truths to necessary ones.

In a borderline case, the epistemicist subscribes to what Schoenfield calls the "Shifty View"—namely, that "the truth-value of an utterance: '*X* is permissible' is highly sensitive to the way the word 'permissible' is used in a linguistic

18 Schoenfield, "Moral Vagueness Is Ontic Vagueness," 259–60.

19 A standard source here is Davies and Humberstone, "Two Notions of Necessity."

community."[20] I think what Schoenfield means here is the thesis that the truth value of *certain utterances* of the form "*X* is permissible" is so sensitive. If "It is permissible for Darryl to divert his attention for nine seconds" was highly sensitive but "It is permissible for Darryl to divert his attention for one second" was stable in truth value across slight shifts in use, then the shifty view, as Schoenfield intends it, would be vindicated.

Here is Schoenfield's master argument against the Shifty View:[21]

> The problem with the shifty view is that, at least for a moral realist, it can't make good sense of moral deliberation. Suppose that Cheryl and her partner are deliberating about whether to abort a fetus at 150 days. They feel very conflicted about the issue and they spend a great deal of time deliberating, indeed, *agonizing*, over whether such an abortion would be permissible. The linguistic anthropologist then knocks on the door. "Guess what!" she says. "I've conducted a series of surveys about the way language users in your community use the word 'permissible.' Here is the data!" After dropping some thick manila folders on the coffee table, the anthropologist disappears. Fortunately, Cheryl and her partner are expert philosophers of language and they can make excellent inferences about the truth-values of sentences with vague predicates based on usage facts. Cheryl and her partner spend the night crunching through the data that the linguistic anthropologist provided. With the first rays of light, Cheryl and her partner breathe a sigh of relief. The usage facts in their community are only consistent with precisifications that permit the abortion in question. Thus, the abortion is permissible.
>
> Note that the claim that Cheryl can learn what is permissible by crunching through the data does not mean that what is permissible depends on linguistic usage, in the sense that, had we used language differently, different things would be permissible. What does, however, follow from the shifty semantic account is that Cheryl can find out that

20 Schoenfield, "Moral Vagueness Is Ontic Vagueness," 264.

21 One might have expected Schoenfield to say that if "permissible" is shifty, then permissibility is not metaphysically special in the way that the robust moral realist supposes, and hence that the shifty view is unavailable to the robust moral realist. However, she is aware of the "more relaxed" view just described in Dougherty, one that seems to combine plasticity with the thesis that permissibility is highly natural, which perhaps makes permissibility special enough for the tastes of the robust moral realist, even if it is a denizen of a highly natural cloud of candidate referents. (See "Moral Vagueness Is Ontic Vagueness," 270.)

Note that she thinks the argument that follows cannot be generalized into an argument against shifty treatments of any predicates since "nobody agonizes about borderline cases of baldness." (See "Moral Vagueness Is Ontic Vagueness," 266.)

> some abortion whose permissibility she was uncertain about, is, in fact (determinately!) permissible by collecting linguistic data. However, it does not seem like crunching through linguistic data is a way of resolving doubts about the permissibility of abortion, especially for the moral realist. Linguistic anthropologists may be helpful with all sorts of things, but solving moral conundrums is not one of them.[22]

The concern, in short, is that the Shifty View allows linguistic anthropologists to resolve moral conundrums by collecting data about linguistic usage.

The Shifty View is (at least approximately) the thesis of semantic plasticity for moral predicates. And as we have seen, the epistemicist subscribes to it. Recognizing this, Schoenfield's complaint against Williamsonian epistemicism is that it is vulnerable to the complaint articulated in the text above:

> The Williamsonian explanation yields the result that Cheryl could (in principle, though it would be extremely difficult!) resolve her deliberation about whether aborting her fetus is permissible by learning enough about her community's linguistic usage.[23]

This complaint involves some important misunderstandings that Williamson goes to some lengths to ward off. If our ignorance of the truth value of a borderline use of "I am thin" is the sort of thing that could be resolved by an anthropological investigation into usage, then that ignorance would run no deeper than our ignorance about the relevant facts of usage. But it is crucial to Williamson's epistemicist vision that the ignorance does run deeper. While the epistemicist is very much open to the metaphysical thesis that the intension of, say, "thin" supervenes on various underlying physical facts, including facts about usage, it is crucial to his vision that details of this dependence are epistemologically elusive. In this connection, Williamson points out the metaphysical dependence encoded by supervenience claims does not mean that we could somehow be in a position to know some supervenient fact once we knew some facts on which the former supervene. Against the thought that "since the supervenience generalizations are metaphysically necessary, they can be known a priori," he writes that

> as Kripke has emphasized, it is fallacious. Indeed, metaphysical necessities cannot be assumed to be knowable at all.[24]

Later he writes that

22 Schoenfield, "Moral Vagueness Is Ontic Vagueness," 265–66.

23 Schoenfield, "Moral Vagueness Is Ontic Vagueness," 267.

24 Williamson, *Vagueness*, 203.

> one should not be surprised that the known supervenience of *A*-facts on *B*-facts does not provide a route from knowledge of *B*-facts to knowledge of *A*-facts.[25]

Part of the point is that even if, say, thinness supervened on physical dimensions, knowing physical dimensions would not always put one in a position to know whether someone is thin. But these remarks are also very relevant to the idea, implicit in Schoenfield's discussion, that since the meaning of vague words depends on the sort of facts recoverable by linguistic anthropology, the question about whether, say, it is permissible for Darryl to divert attention for thirty-one seconds can be resolved by linguistic anthropology.

Williamson also points out that the slogan that meaning supervenes on use neglects "the environment as a constitutive factor in meaning."[26] But he does not wish to rest everything on that point. His idea is that even granting that some refinement of that slogan is correct (one that dispenses with a crude notion of "use"), it would not vindicate the thought that facts about meaning are somehow accessible to humans. Speaking of the imagined refined gloss on "use" he says:

> Although meaning supervenes on use there is no algorithm for calculating the former from the latter. Truth conditions cannot be reduced to statistics for assent and dissent.[27]

Consider material conditionals about Darryl, with a borderline sentence about permissibly diverting attention inserted as the consequent and with some complicated antecedent giving precise data about usage facts and precise data about Darryl's physical situation. On the epistemicist picture, even if some such an antecedent is determinately true and even if the conditional is necessary, it by no means follows that the conditional is determinately true. Consistent with the conditional expressing a necessary truth there might be a proposition easily meant by the consequent such that the conditional is false when the consequent is interpreted this way. (Thus such conditionals pose no obvious threat to the claim that if the antecedent of a conditional is non-borderline true and a conditional is non-borderline true then its consequent is non-borderline true.)

It is thus crucial to the epistemicist vision that linguistic anthropologists *cannot* resolve borderline issues in the way that Schoenfield envisages. Her objection depends on a misunderstanding of the view.[28]

25 Williamson, *Vagueness*, 204.

26 Williamson, *Vagueness*, 206.

27 Williamson, *Vagueness*, 206.

28 A referee reasonably wondered whether something in the vicinity of Schoenfield's concern could be revived for the kind of epistemicist—like Williamson—who allows that a

There is one further gap in Schoenfield's argument. Recall from our earlier discussion that it is best to think of candidate interpretations of predicates as properties but not sets. (Similarly, one should think of candidate interpretations of sentences as propositions, not truth values.) Pretend now that a linguistic anthropologist could somehow discern which property was picked out by "is permissible" on an occasion of use. Even then it would not straightforwardly follow that Cheryl's conflict would be epistemically resolved. Knowing which property is expressed by a predicate is one thing; knowing whether a certain object, act, or event instantiates it is, on the face of it, quite another. What Schoenfield's argument thus requires is not merely that the anthropologist can resolve which property "is permissible" picks out but, moreover, that they can resolve this in a way that somehow automatically answers questions about the extension of that property.

Before moving on, I would like to draw attention to one further issue. In the quoted passage, Schoenfield imagines Cheryl and her partner "spending a great deal of time deliberating" and indeed "agonizing" about whether abortion is permissible in the case she describes. Assume it is a case of vagueness. What kinds of attitudes would the epistemicist recommend? Insofar as the case is known by all parties to be borderline it would in some ways be a bit odd to spend a great deal of time deliberating. Normally when one deliberates

---

*superbeing* could in principle know the facts of reference. The referee wrote: "The thought I take it is that it's implausible that these sorts of facts (processed by a human anthropologist or a super-being anthropologist) could serve as basis for determining the truth value of 'this abortion is permissible.'" Let us think this through. Williamson's superbeing will be aware of the constitutive dependence of meaning on the environment. Moreover, epistemicists will be very sympathetic to the idea that, even at the use end of things, tiny, inscrutable micro-differences in use that are not fully captured by ordinary anthropological data may make a difference. Given this, our superbeing is going to need a lot more to work with—even on the "use" side of things—than a folder of use facts of the sort that might be cataloged by human anthropologists. So our superbeing will need to know everything relevant about the environment—perhaps up to some astonishing detail, and will likely have to know incredibly fine-grained facts about use. Perhaps the best way to imagine our superbeing, then, is as one who knows the true function from microphysical distributions to facts about reference and is also capable of surveying the microphysical lay of the land in full detail. Our superbeing says "I've surveyed the microphysical landscape, applied the algorithm and determined that the predicate 'is permissible' expresses a property that applies to the referent of 'this abortion.'" But in this setting the thought that the superbeing's reflections could not serve as a basis for determining whether the abortion is permissible has no bite. Cheryl will have no problem knowing that the abortion is permissible iff the property expressed by "is permissible" is instantiated by the event picked out by "this abortion." So if Cheryl is convinced that the superbeing knows exactly how the facts of reference supervene on microphysics and that the superbeing knows the microphysical lay of the land, then she will of course regard the superbeing as having settled the question she is worrying about.

at length about a question it is because one hopes to know the answer. If one knows a case is borderline, one knows one will not find out the answer. So why the prolonged deliberation? That is not to say that the epistemicist recommends a do-not-care attitude. Cheryl can be *concerned* that the action is impermissible. She can *fear* that the action is impermissible. She can have *nonzero credence* that the action is impermissible. But the point of extended deliberation is less clear when one knows that one is not going to discover an answer at the end of it.

*2.2. Unknowability*

Constantinescu and Dougherty both raise concerns to the effect that the epistemicist approach to moral vagueness delivers unknowable ethical truths, concerns that are cited with guarded approval by Schoenfield.[29] We have already seen that epistemicism's purported implication of unknowable moral facts is not incontestable, since a truth that is unknowable under the guise of one vague predicate may, for all epistemicism says, be knowable under another. But I shall look past that point in the following discussion.

Dougherty writes:

> How many cents are you required to spend on a taxi in order not to be late for an appointment for which you have promised to be punctual? A friend of an epistemic position may say there is a precise minimum here. But it stretches the imagination to think that we could know what this minimum is.[30]

He goes on to worry that it is arguably a conceptual truth about ethical facts that they must be action-guiding," that the postulate of unknowable ethical truths threatens to clash with that conceptual truth.[31]

Cristian Constantinescu discusses epistemicist treatments of moral vagueness and argues that it is incompatible with a nonnaturalist position that takes moral facts to be "intrinsically reason-giving."[32] His focal concern is that, while the phenomenon of unknowable facts may be unproblematic, there is something problematic about unknowable ethical facts. The picture here, as Constantinescu develops it, is that every ethical fact constitutes a normative reason for or against certain kinds of actions.[33] He then argues that there is something incoherent

29 Schoenfield, "Moral Vagueness Is Ontic Vagueness," 279.

30 Dougherty, "Vague Value," 361.

31 Dougherty, "Vague Value," 361.

32 Constantinescu, "Moral Vagueness," 155.

33 I think it is useful here to distinguish facts that are naturally expressed using explicitly ethical language and facts that are not so expressed but that have (in some cases contingently) normative significance. That *S* is on fire may well be a reason to help *S* but the proposition

about the idea that a normative reason could be unknowable, since normative reasons have to be in principle available to us as things to justify actions:

> But what doesn't seem possible is to divorce *n*-reasons [his shorthand for "normative reasons"] even from a maximally improved capacity for practical rationality. Thus, we can of course accept that there may be moral reasons for us to desist from some of our current practices, but that those reasons are inaccessible to us, due to certain biases or errors in our judgement of which we are unaware. But to recognize them as reasons means to accept that they would serve as justifications for us if our reasoning abilities were improved. What seems incoherent is the thought of an *n*-reason entirely divorced even from the sound exercise of a maximally improved capacity for practical reasoning. To claim that there are reasons which couldn't be anyone's reasons seems almost vacuous. I shall express this upshot in the form of the following epistemic constraint on normative reasons:
>
> > Epistemic constraint on reasons: If $R$ is an *n*-reason for $X$ to $\phi$, then $R$ can feature in a rational justification of the claim that $X$ ought to $\phi$, a justification which $X$ knows or could come to know if $X$'s reasoning abilities were maximally improved.[34]

An important thing to notice is that it is not the full apparatus of epistemicism but standard inference rules for the quantifiers and logical connectives, in combination with some fairly banal truths, that delivers the conclusion of unknowable moral truths. Let us use Darryl as our focal example. Take the banal truths

(a) Darryl is permitted to divert his attention for one second; and
(b) It is not the case that Darryl is permitted to divert his attention for three hundred seconds.

---

that *S* is on fire is not something expressed using normative language. My remarks can be adapted to both an expansive and restrictive conception of ethical facts, where the expansive conception includes facts of ethical significance that are not expressed using ethical language.

34 Constantinescu, "Moral Vagueness," 178–79. He then cites with approval Parfit, who wrote: "when it is true that we have decisive reasons to act in some way, this fact makes it true that if we were fully informed and both procedurally and substantively rational, we would choose to act in this way" (*On What Matters*, 1:63). If "full information" includes being fully informed about what is permissible and what is not, then this remark does not in fact lend any support at all to the claim that all ethical truths are knowable. I have not found a consensus among my informants as to what Parfit intended here. Even if "full information" included only the entire body of empirical facts, this would hardly give much support to the claim that all normative reasons are knowable since the relevant empirical facts that figure in the idealization may include unknowable ones.

We can then use standard inference rules to generate a reductio of "For all numbers of seconds $n$ (Darryl is permitted to divert his attention for $n$ seconds $\supset$ Darryl is permitted to divert his attention for $n + 1$ seconds)" in order to derive:

(c) It is not the case that for all numbers of seconds $n$ (Darryl is permitted to divert his attention for $n$ seconds $\supset$ Darryl is permitted to divert his attention for $n + 1$ seconds).

We can then exploit the duality of universal and existential quantification to reach:

(d) There is a number of seconds $n$ such that Darryl is permitted to divert his attention for $n$ seconds, but it is not the case that Darryl is permitted to divert his attention for $n + 1$ seconds.

Here is a further truth that also seems fairly banal:

(e) There is no number of seconds $n$ such that we (Darryl or any other human) are in a position to know that (Darryl is permitted to divert his attention for n seconds but that it is not the case that Darryl is permitted to divert for $n + 1$ seconds).

(If one happens to have some optimism here, then run the whole argument using milliseconds or nanoseconds—here the claim corresponding to (e) is even more secure.)

But (d) and (e) entails:

(f) There is a number of seconds such that (Darryl is permitted to divert his attention for $n$ seconds and it is not the case that Darryl is permitted to divert his attention for $n + 1$ seconds) and we are not in a position to know that (Darryl is permitted to divert for $n$ seconds and that is not the case that Darryl is permitted to divert for $n + 1$ seconds).

But this is the very kind of claim that enemies of unknowable ethical truths balk at. (Notice that the argument does not deploy words such as "determinately," "precise," and "sharp," so their semantic contribution to sentences in which they occur is neither here nor there.) Every view of vagueness that accepts the relevant banal claims (a), (b), and (e), and the validity of the relevant inferences is committed to conclusions like (f). Such views (as noted earlier—see note 3) include not only epistemicism but supervaluationism, among others.[35]

35 Williamson makes a point of emphasizing that supervaluationism is committed to claims that encode the idea that tiny differences sometimes make a difference: "Many people have found the major premise [of the sorites] implausible just because it seemed to them that

Dougherty does not seem to be suggesting that we give up claims like the fairly banal (a), (b), and (e). He thus seems to be suggesting in effect that there is a compelling ethical argument against standard inference rules for the quantifiers. Turning to Constantinescu, the following claims seem extremely plausible (at least insofar as we are comfortable with the ideology of normative reasons):

(g) For any *n*, if is a permissible for Darryl to divert attention for *n* seconds, then the fact that it is permissible for Darryl to divert attention for *n* seconds is a normative reason.
(h) For any *n*, if it is impermissible for Darryl to divert attention for *n* seconds, then the fact that it is impermissible for Darryl for *n* seconds is a normative reason.

Consider now

*Normative-Reasons Conjunction*:
(i) (*p* and *p* is a normative reason) and (*q* and *q* is a normative reason) ⊃ ((*p* and *q*) and ((*p* and *q*) is a normative reason)).[36]

This principle is obligatory for those who hold that all ethical truths are normative reasons since it is obvious that the conjunction of any pair of ethical truths is an ethical truth. But it is plausible in its own right. But putting (g) together with (h), (i), and (j), we can conclude:

(j) There is a number of seconds *n* such that (i) Darryl is permitted to divert his attention for *n* seconds but it is not the case that Darryl is permitted to divert his attention for *n* + 1 seconds, (ii) that conjunctive fact about *n* is a normative reason, but (iii) we are not in a position to know that conjunctive fact.

Thus, in effect, Constantinescu seems to be suggesting that the epistemic constraint on reasons provides good grounds for rejecting standard inference rules for the quantifiers and connectives.

Giving up the relevant inference rules strikes me as something of an overreaction to examples like Darryl. At any rate I invite readers to consider whether they wish to go in that direction and, if so, what inference rules they recommend putting in their place. Certainly these authors do not suggest any alternative logical framework in which to evaluate the claim that there are unknowable

---

there could not be a number *n* such that *n* + 1 grains make a heap and *n* do not. Supervaluationism makes the very claim that they find incredible" (*Vagueness*, 153).

36 In saying this I do not mean to suggest that *strength* of reasons can be computed in a flat-footedly additive way.

moral truths.[37] This suggests to me that they have not really confronted the choice between giving up the relevant inference rules and accepting unknowable moral truths.

But what of Constantinescu's purported connection between normative reasons and moral facts and Dougherty's purported connection between motivating reason and moral facts? Dougherty's view can be put in its best light if we accept the following principle connecting reasons to action:[38]

> *Knowledge Principle for Personal Reasons* (KRP): If $p$ is $S$'s reason for $\phi$-ing, then $S$ knows $p$.

At least on the most natural way to resolve the notoriously flexible possessive construction, this principle has a good deal going for it.[39] At any rate, I shall assume it here, which helps rather than hinders Dougherty's concerns. A similar idea is in play in Constantinescu, since he relies on the idea that "To claim that there are (normative) reasons which could not be anyone's reasons seems almost vacuous."[40] If we supplement this thought with the thought that for $p$

37 Note that a flat-footed, tripartite division of propositions into true, false, and neither does not on its face help much, as it is hard to imagine that belief in truths on the boundary will be safe enough to be known.

Of course, there are some in the Dummettian tradition that self-consciously try to preserve the knowability of moral truths by dispensing with classical logic. (See for example Wright, "Ethical Truths.") I do not have the space here to get into larger questions about the level of abductive support that is enjoyed by classical propositional logic and the standard inference rules for the quantifiers that are recommended by classical model theory. However, it should be quite obvious to readers that I hold these in high regard. Intuitionists reject the inference from the negative universal, "It is not the case that every number n is such that if n grains make a heap, n plus one makes a heap" to an existential conclusion. (Fine opts for an even weaker logic than the intuitionist one, one that precludes inferring $Q$ from ($P$ and Not ($P$ and Not $Q$)); "The Possibility of Vagueness.") What is striking, though, is that many of the writers in the ethics literature who raise knowability worries—Dougherty and Constantinescu being paradigms—are not mindful of the threat to the standard inference rules and certainly do not offer a competing logic as a working alternative.

For another defense of unknowable moral truths, see McGrath, "Moral Realism without Convergence."

38 For defenses of the idea encoded by KRP, see Dietz, "Reasons and Factive Emotions"; Unger, *Ignorance*; Hawthorne and Magidor, "Reflections on the Ideology of Reasons"; and Williamson, "Acting on Knowledge." The expression "personal reason" is borrowed from Grice, *Aspects of Reason*.

39 For a discussion of the different ways to read possessive constructions in this kind of context, see Finlay, *Confusion of Tongues*, chap. 5.

40 Constantinescu, "Moral Vagueness," 178–79.

to be someone's reason it has to be known, we get his desired conclusion, but without that supplement it is hard to see how to reach that conclusion.

Even granting KRP, it will seem excessive to many to give up standard inference rules for quantifiers and/or connectives on the basis of the lines of thoughts these authors advance. Nevertheless, it may be helpful to offer further therapy to those who are still tempted by them.

It is important to notice that there is something very misleading about the thought that if a proposition is unavailable as a personal reason then it cannot guide action. I do not want to fuss too much here about the term of art "guiding action." But—to return to a theme from the end of section 1.3—it bears emphasis that one can stand in all sorts of interesting relations to an unknowable fact, relations that can certainly have a bearing on one's planning and one's action. One can know in many cases that there is a chance that *p*. One can in some cases know that there is a significant chance that *p*. One can have significant moral concern that *p*. One can take precautions against *p*. One can fear that *p*. One can have a decently high rational credence that *p*. And so on. So the thought that unknowable propositions are dead to one as far as actions and planning are concerned does not seem to be a very good thought. Of course, assuming the personal reasons principle, there is one way that an unknowable proposition *cannot* be a guide to action—namely, by serving as a personal reason for action. But it seems extremely unpromising to elevate this relation over all others in one's account of ethical facts and not much more promising to elevate this relation over all others in one's account of what can stand as a normative reason.[41]

### 2.3. *A Proportionality Argument against Epistemicism*

Constantinescu has another argument that I think is unpersuasive but that (as we shall see in the next section) points toward some interesting issues.

The motivating concern is that, according to the epistemicist, incredibly small differences can make the difference between doing something permissible and not doing something permissible. Schoenfield's case of Darryl is adequate for our purposes here (Constantinescu discusses a very similar case): there are two periods P1 and P2 only a nanosecond apart such that it is permissible for Darryl to divert his attention for P1 but not for P2. In a case like this, the line of thought runs, an idealized agent who was apprised of what was

41 While I do not wish to put too much stock on the point, we should also recall that it is not really part of epistemicism that it is impossible *simpliciter* for humans to know borderline claims. They could, for example, learn them by testimony from superbeings. Suppose a superbeing knew a certain borderline diversion was impermissible for Darryl. The superbeing might well offer some advice: "Don't do that! It is impermissible!" Would this not put Darryl in touch with a normative reason?

going on would have praised Darryl in the one case but would have "blamed and chastised" him in the other.[42]

Constantinescu worries that "something is amiss. . . . The slightest difference (one nanosecond, one nanogram, one nano-anything) is all it takes for an option to change moral valence. This appears to violate considerations based on justice."[43]

The line of thought turns on something like the following inconsistent triad:

1. Other things being equal, very different reactions/treatments of agents are warranted as between any case C1 where Darryl diverts attention from his child in an impermissible way and a case C2 where Darryl diverts attention in a permissible way. (I say "other things being equal" simply to control for pairs of cases C1 and C2 where Darryl acts permissibly in C1 as far as diverting attention goes but commits some other sin that is not committed in C2.)
2. But if epistemicism is right there is a pair of cases incredibly close together microphysically—indeed where the difference in attention is only one nanosecond apart—where Darryl acts impermissibly in one case and permissibly in another and where other things are equal.
3. If cases are almost microphysically the same they warrant almost the same reaction/treatment.[44]

Constantinescu evidently thinks 2 is the culprit, suggesting that "we should reject the epistemicist's idea of sharp properties on moral grounds."[45]

As noted earlier, we should be cautious of applying predicates like "sharp" and "precise" to properties. Such predicates, for the epistemicist, apply to representations. We can contrive "vague*" and "precise/sharp*" along Williamsonian lines, but then we should remember that properties can be both vague* and sharp*. What Constantinescu is getting at with his own use of "sharp" is, I think, merely the epistemicist's commitment to classical propositional logic and standard rules for the quantifiers. And indeed these logical commitments

42 Constantinescu, "Moral Vagueness," 180.

43 Constantinescu, "Moral Vagueness," 181.

44 At one point Constantinescu goes so far as to say that if people *X* and *Y* are almost indistinguishable then it would be unjust for them "to receive different treatment" ("Moral Vagueness," 181). The idea that pairs of people that are almost indistinguishable should not be treated differently *at all* is indefensible. Pairs of cases that are almost the same can be chained together so that there are cases wildly different at each end. But repeated application of the principle that almost indistinguishable cases cannot be treated differently at all would have us conclude that the cases at each end of the series call for the same treatment.

45 Constantinescu, "Moral Vagueness," 180.

alone, with or without the extra commitments of epistemicism, get us the conclusion that he finds abhorrent. Grant that it is permissible for Darryl to divert his attention for one second but impermissible to divert his attention for five minutes. Armed with the relevant logical tools we can simply prove that there are two periods of time, P1 and P2, one nanosecond apart, such that it is permissible for Darryl to divert is attention for P1 but that it is not the case that it is permissible for Darryl to divert his attention for P2. We do not need to make any use of such ideology as "sharp properties." As before, what Constantinescu is in effect telling us is that we have decisive moral grounds to reject the relevant inferences rules! Again, this will seem excessive to many. And, as before, he offers no alternative logic and so, for example, we are left in the dark about how we are supposed to reason with any of the principles in the paper that he is friendly to.

As an argument against classical propositional logic and/or the standard inference rules for the quantifiers, I am completely unmoved. We should all make our peace with the sorites and realize that, for pretty much any predicate, tiny differences sometimes make a difference between the predicate being true of a thing and being false of a thing. But the inconsistent triad is not without interest. Even if we dismiss the line of thought as grounds for logical deviance we are left with the interesting decision as to whether to give up 1 or 3 in the inconsistent triad.

In support of 3, Constantinescu offers a sweeping claim about supervenience:

> *Proportionality Constraint on Supervenience*: If *P*-properties supervene on *Q*-properties, then no two things can differ greatly with respect to their *P*-properties without differing greatly also with respect to their *Q*-properties.[46]

As stated, this principle overgeneralizes. Let the *Q* properties be the family of microphysical properties. Let *P* be the singleton set containing the property of being an action that is not permissible. Consider two worlds, one where Darryl, fifty days in succession, diverts his attention from his child permissibly, but only just permissibly: in each case, if he had diverted his attention just one-hundredth of a nanosecond less, he would have acted impermissibly. The second world is extremely similar, microphysically, except that Darryl, fifty days in succession, diverts attention from his child impermissibly but only just impermissibly—on each day his attention is one one-hundredth of a nanosecond less than each corresponding day in the first scenario. Once we have made our peace with the sorites, classical propositional logic, and the relevant

46 Constantinescu, "Moral Vagueness," 182.

inference rules for the quantifiers, it is hard to preclude pairs of possible worlds like this. By hypothesis, the two worlds do not differ much with respect to the *Q* properties. But they differ a good deal with respect to the *P* properties: there are fifty instances of *being an impermissible action* in one world but the corresponding actions in the other world do not instantiate that property. Similar issues come up with myriad choices of non-ethical *P*s. Take the *P* properties to be the property of having a perfectly flat tabletop. Consider a pair of worlds, one in which there are fifty tables with perfectly flat tabletops, another with fifty tables that have vanishingly small imperfections in their tops so that none of them are perfectly flat. The underlying distribution of matter can be very similar but the difference in distribution of *P* properties is very significant. In sum, we cannot underwrite 3 by anything nearly as general as the principle that Constantinescu labels "Proportionality." Nevertheless, there are some intriguing issues in the vicinity to which I now turn.

## 3. CONTINUITY

I articulated Constantinescu's argument in terms of an inconsistent triad. Assuming some background logic of the sort previously alluded to, we are left to choose between a proportionality idea—namely, that (at a rough first pass) tiny physical differences in cases cannot render fitting significantly different attitudes and treatment—and the idea that the difference between permissible and not permissible actions (other things being equal) warrants markedly different treatments.

Something like the proportionality idea also gets advanced in Dougherty. Moved by Ted Sider's thought that one "cannot both uphold epistemicism and continue to believe that differences in vague predicates always retain the significant we previously took them to have," he writes:

> There is some plausibility to thinking that if there is only a tiny descriptive difference between the actions, then any ethical difference could not be very important.[47]

He concludes that, given epistemicism, we need to "scale back on the significance we place on applying these predicates."[48] Here again it bears emphasis that the observation that tiny differences can make a moral difference do not rely on the full epistemicist package: some humble truths together with some standard inference rules for quantifiers and connectives all by themselves

47 Sider, "Hell and Vagueness," 65; Dougherty, "Vague Value," 361.

48 Dougherty, "Vague Value," 362.

deliver the conclusion that tiny differences can separate the permissible from the impermissible. That said, an epistemicist commitment to semantic plasticity may in some cases play a distinctive motivating role in moving us to "scale back" our estimations of significance.

I shall propose a way of sharpening the proportionality idea that generates theses that can be explored with some rigor. (It will be a sharpening that does not really require the ideology of "tiny" and "significant.")[49] The idea is to find ethical *scales* that seem of foundational ethical significance to actions, and to inquire as to whether, as one moves continuously from one physical possibility to another, the values on the ethical scale vary continuously. As a ridiculously simple but conceptually instructive toy model, pretend that the only way that reality varied across time and across worlds was on one parameter: the height of Jones. Take any path through time or worlds where Jones's height varies continuously (so that, for example, there is no time in the series such that Jones is $x$ inches tall at $t$ but "jumps" in height so that at all times in some period after $t$ Jones is, for some fraction $1/n$ of an inch, at least $x$ plus $1/n$ inches tall). A proportionality thesis about some moral scale will say that the values on the ethical scale will vary continuously along that path as well. Now the world is obviously a lot more complicated than that. But we can still apply the same basic idea: once we have a notion of things varying continuously in a physical way along a path (without "jumps"), we have the means to state proportionality theses of the sort I am interested in. To fix ideas I shall be looking at some proportionality theses that focus on the microphysical terrain—here the notions of continuous variation have a natural home. A proportionality thesis will claim that the values on the ethical scale vary continuously as the microphysical terrain varies continuously. An anti-proportionality thesis will allow for discontinuous shifts. Let us look at a few proportionality theses (or perhaps, better still, "Continuity Theses") along these lines

*Some Moral Continuity Theses*:

1. If a series of possible worlds vary continuously in their microphysical profile, then insofar as they vary in moral value, they will vary continuously in their moral value.[50]

49 These ideas are touched on briefly in Dorr, Hawthorne, and Yli-Vakkuri, *The Bounds of Possibility*, 323. Discussions with Dorr have greatly influenced the writing of this section of the paper.

50 One might be tempted to instead articulate principles along the lines of, "If two cases have a small microphysical difference, then their difference on the moral scale is small," and more generally, "The size of microphysical difference corresponds to the size of the difference on the moral scale." (This kind of formulation is at least naturally suggested by Dougherty's idea that "tiny descriptive differences" cannot make for a "very important"

2. If a series of possible people vary continuously in their microphysical profile, intrinsic and relational, then insofar as they vary with regard to their moral worth, they vary continuously with regard to their moral worth.
3. If a series of possible people vary continuously in their microphysical profile, intrinsic and relational, then, insofar as they vary with regard to how meaningful their lives are, they vary continuously with regard to how meaningful their lives are.
4. If a series of possible people vary continuously in their microphysical profile, intrinsic and relational, then (insofar as they vary at all) they vary continuously with regard to how much admiration is fitting for them. (There are obviously similar principles for other reactive attitudes.)
5. If the microphysical character of the world evolves continuously over a series of times, then insofar as the intrinsic moral value of each time varies, it varies continuously.

We should immediately acknowledge one way that these theses might naturally be weakened. As noted earlier, the classical magnets view is very tempting for phenomenal consciousness. It is, for example, very tempting (for atheists) to think that in the life of a person there is a metaphysically special time that marks the last time that the person is phenomenally conscious. Armed with this picture, someone might think that there is a very easy answer to such questions as, "How could a tiny microphysical shift mark a morally significant boundary?"—shifts that will seem insignificant under the lens of microphysics will mark metaphysically significant transitions from a case in which the very special

---

ethical difference, at least in a context where one is giving "descriptive" a microphysical spin.) But such principles are challenging for a number of reasons. For one thing, certain physical variations, even sizable ones (say to the configuration of sand dunes in an uninhabited desert), may make *no* moral difference whatsoever. One can fix for this by opting for the idea that the difference on the moral scale is *no more than* the difference on the physical scale (and indeed Dougherty's own formulation is cautious in the required way). However, and perhaps more importantly, it is not clear how to give proper discipline to a notion of smallness that is cross-scale applicable. Claims like "*x* amount of time is as small as being *y* inches tall or as being *y* kilograms in mass" sound somewhat fishy, and if made at all rely on highly context-dependent, impressionistic reactions that are hard to systematize and risk relying on cross-scale notions that are too vague to be tractable. Of course there is no reason to think Dougherty is confining himself to microphysical facts when he says "descriptive facts." But the issue raised above—how to generate cross-scale notions of smallness in a way that is not excessively vague and impressionistic—is still a somewhat pressing one. I leave the challenge of rigorously developing that kind of formulation to others. The continuity theses in the text do not rely on any such notions.

phenomenon of phenomenal consciousness is present to one where it is absent. Unless one wished to combat the classical magnets picture of consciousness, it seems that one will have to be more guarded when stating continuity. The more guarded version of continuity will say, of the relevant moral scale, that if a series of cases is continuous with regard to *both* microphysical and phenomenal consciousness, then the values along the relevant moral scale will, if they vary at all, vary continuously. I shall assume that insofar as readers are drawn to the relevant picture of phenomenal consciousness, they will operate with weakenings of this sort.[51]

For each pair of a property and a moral scale, there is an anti-continuity option, one that posits a discontinuous shift along the moral scale at certain points along a sequence of cases that vary continuously in microphysical respects. Consider, for example, a series of worlds involving Darryl that vary continuously in microphysical respects and where the key difference along the series is how long Darryl diverts his attention to his child. At one end of the series Darryl diverts his attention for a second, at the other for five minutes.

51 There are potentially far more guarded versions of proportionality theses. Consider a continuous path through worlds from one where I drive off this morning with a car that I own to a world where I drive off with a car that I do not own. One might think that, while the worlds continuously vary microphysically, there is still at some point a "jump" in the "descriptive facts"—from car ownership to non–car ownership. A far more guarded vision (for some moral scale) is along the following lines: when the descriptive facts (taken *en bloc*) vary smoothly—with no "jumps" along a path, the values of the moral scale vary continuously as well. (Here I am drawing on Dougherty's preferred ideology of "descriptive facts.") The key challenge here is to articulate what it is for the descriptive facts to vary smoothly. After all, on a standard conception of properties, for any pair of cases there will be infinitely many descriptive properties had by one but not the other. When should we count a case as a cliff/jumping point in the relevant sense? I opt for the more ambitious continuity theses in the text, in part because I can more easily see them being made rigorous (since the notion of continuous variation in the microphysical state of the world is a respectable one in physics) but also in part because they may surprisingly turn out to be true for various moral scales of importance (and so, for example, and surprisingly, it may turn out that in the car-ownership case, the fundamental moral scales do not jump discontinuously at the point where I drive off with a car that I do not own.) Still, I do not want to preclude our making sense of more guarded versions of proportionality than the one in the text: if the reader finds a way they are comfortable with to articulate some other version, I invite them to see how the themes of the text play out for that alternative version.

I should mention one further caveat. Suppose the moral scale is somewhat discrete, lacking the structure of the real line. Suppose for example that there are only ten thousand possible levels of admiration. There is still an analogue to the notion of varying continuously—namely, as one moves continuously across a series of physical cases, a shift from one level of admiration being fitting to another always proceeds via all the intervening levels of admiration. (It again bears emphasis that this thesis has absolutely no need for the ideology of "small" or "tiny.")

Suppose we accept that there is a time *t* that marks the boundary between permissible and impermissible. Suppose for example, diverting attention for the period up to and including *t* is permissible, but that any longer period of diversion is impermissible. Then we might naturally think that as the physical landscape changes continuously, there is a discontinuous shift in some or all of the important moral scales—for example, perhaps there is a sharp drop-off right after the *t*-diversion world in the admiration fitting for Darryl. After all, in all the worlds in the series where the inattention was greater than the period up to and including *t*, he impermissibly diverted attention but in the other worlds he did not. According to the anti-continuity picture, a graph of the moral values across the series would display an abrupt cliff right at *t*. But the continuity lover for a particular moral scale will deny that the values from the scale are ever distributed in a cliff-like way across worlds that continuously vary in their microphysical makeup.

Other useful test cases for a continuity treatment of this or that moral scale are provided by properties that are not articulated using explicitly ethical language. Take the loving relation. One might think that if one possible individual loved no one while another possible individual loved someone, then, insofar as they were extremely close in moral worth, that would have to be because of some significantly compensating feature in the life of the loveless being—a feature that compensated for the absence of love. But reflection on sequences of worlds that vary continuously in their physical landscape, in combination with a continuity thesis about the moral-worth scale, suggests that this may be wrong. We can find a path from a lover to a non-lover that marches through continuously varying microphysical profiles. There will be pairs of cases separated by tiny microphysical differences, one of which involves the borderline presence of love, the other the borderline absence of love. And so there will be pairs arbitrarily close to each other in microphysical makeup that are divided by the absence and presence of love. If one subscribes to the continuity principle for moral worth, one will deny that moral worth abruptly drops off as one crosses the boundary to a loveless being.

Also consider principle 5, about the intrinsic value of times. Imagine a world containing a single creature who, at some point in time, dies. The defender of anti-continuity will naturally think that the intrinsic value of times sharply drops off at the point at which the creature is dead (especially if that was the only living creature left in the world).[52] The defender of continuity for the

52 I take some liberties here and elsewhere—a more careful (but also more verbose) statement would acknowledge the difference between open and closed intervals. There may not be a first time where the creature is dead; rather, the period of life might form a closed interval and the period of death an adjacent open interval with no first time.

intrinsic value of times will say that there is no discontinuous jump in the intrinsic value of times.

The choice between continuity and anti-continuity for various moral scales is an intriguing one. The last example might initially seem to suggest that anti-continuity is obviously the way to go. Is it not obvious that at the very point someone dies there will be an abrupt jump in how much concern is due to them and so on, and if there only a few people in the world, an abrupt jump (downward) in the overall value of things? But on reflection things are far from obvious here. After all, it is overwhelmingly plausible on reflection that the line between life and death is vague. Supposing we take the semantic plasticity approach of the epistemicist, we will think that one cutoff is expressed, but myriad other cutoffs could very easily be expressed. It is moreover natural to think that candidate cutoffs trail off in terms of ease of being expressed by "the boundary between that creature's life and death" in a continuous way. If none of the cutoffs are particularly natural or metaphysically privileged vis-à-vis their neighbors, then it seems plausible that moral value takes a continuous curve downward around the point of death rather than the cliff edge conducive to an anti-continuity picture.

The fan of proportionality will take a similar perspective on other cases: the boundary between loving and not loving may seem *prima facie* to enjoy such immense moral significance that there is an abrupt dip in value, fittingness of admiration, and so on, at the point in a physically continuous series of worlds that marks the boundary between a lover and a non-lover. But on reflection it seems plausible that "love" is vague, that there is no metaphysically privileged boundary in the vicinity, and on this basis one might well, for analogous reasons, begin to like the picture that the dip in value as love recedes across the series of worlds will, when graphically depicted, look like a curve rather than a sharp cliff edge.

Once one has gotten used to continuity in those cases, it is at least tempting to extend it even to the case of moral vocabulary—like "permissible"—that Dougherty focuses on. Consider the line between Darryl permissibly diverting attention and not permissibly diverting attention. Adopt a classical magnets approach and it is natural to say that there the line marks a discontinuous cliff in values along relevant moral scales. In that setting it is natural to combine an anti-continuity approach with the thesis that the exact place of a discontinuous jump is difficult to know. If someone presses, "How could a tiny physical difference mark a not-so-tiny shift in values along the relevant moral scale?" the classical magnets lover will insist that a tiny microphysical difference belies a metaphysically important difference that is hidden from view when one looks at the world merely through a microphysical lens. But once one takes an

epistemicist approach, acknowledging that the boundary between permissible and impermissible does not stand out as metaphysically significant vis-à-vis nearby boundaries, then it becomes far more tempting to think that the various moral scales—like moral worth and fittingness for admiration—do not have a cliff-like structure but rather have a curve-like structure around that boundary.

That said, I do not think the epistemicism is by any means *forced* to anti-continuity, an approach to which I now turn. Anti-continuity certainly has something to be said for it. It is far too easy to caricature the anti-continuity approach. Let us return to Constantinescu's thought that it is "unjust to praise one person and blame another when the difference between their actions was slight." It may be thought that an anti-continuity approach to the reactive attitudes would recommend heaping praise on Darryl when he just about manages to act permissibly, and blame, contempt, censure, and so on when he just about fails to act permissibly. (Indeed Constantinescu's talk of blaming and chastising the person who only just about acts impermissibly encourages this vision.) But even if anti-continuity is right, that is the wrong picture. After all, when Darryl acts only just permissibly, he acts in a way that is, for all he knows, impermissible. That is not great. And when Darryl acts only just impermissibly, he does not know he is acting impermissibly. That is not nearly as bad as knowingly acting impermissibly. Moreover, it is natural to think that norms such as "Don't divert attention impermissibly" generate what Williamson calls "derivative norms," including, in this case, the "secondary norm" that people should have the disposition not to divert their attention impermissibly.[53] Darryl's only just permissibly diverting his attention may be a tell-tale sign that he does not have a stable disposition to permissibly divert his attention and thus signal failure to comply with the secondary norm. Suppose instead that coming close to acting impermissibly was an aberration and that Darryl does have a stable disposition to act permissibly (that was interfered with in an unusual way on this occasion). Then the relevant case of Darryl's acting impermissibly, since it is ever so similar, will likely also be an aberration, and in that case, too, the secondary norm will be satisfied.

All this suggests that we should not think of the difference between the two cases as all that great. We should not think it is a boundary so dramatic as to warrant something as contrasting as heaping of praise versus heaping of blame.[54] But such a concession is perfectly compatible with the claim that the boundary between the permissible and the impermissible marks a discontin-

53 Williamson, "Acting on Knowledge."

54 Of course, this kind of response is not available to a theologian who thought that each possible life warranted one of two sharply dichotomous divine reactions, being sent to heaven and being sent to hell. Thus, even the proponent of anti-continuity might not be very comfortable with the thought that cases almost the same physically can warrant

uous boundary with regard to the fittingness of various reactive attitudes. Not knowing where the boundary lies, the lover of anti-continuity will then be committed to not knowing exactly what degree of this or that reactive attitude is most fitting in certain cases. But that is all right. Having made their peace with unknowable moral truths, it is not a particularly great additional shock to make one's peace with the fact that among such truths are facts about exactly which levels of which reactive attitudes are fitting in various cases.

If anti-continuity is defensible for this or that moral scale in connection with a property like being permissible (so that the cliff edge on the scale lines up with the boundary of the property), the key to defending such a view will, I think, be to promote thoughts such as the following: the whole point of various moral predicates is they warrant at least somewhat significant differences in reactive attitudes, mark somewhat significant differences in moral worth, and so on. One can run sorites series on such predicates as "being evil" and find physically continuous series where there is a cutoff associated with these predicates. But if the whole point of these predicates is that they warrant at least somewhat significant differences in reactive attitudes, and trigger somewhat significant differences with respect to other important moral scales, then their impact on certain of those scales had better be cliff-like.

How is this thought to be reconciled with a semantic plasticity claim to the effect that it is a semantically fragile matter which line is drawn by "evil," "permissible," and so on? The natural way to harmonize things here is to posit a kind of penumbral connection between the relevant predicates and the language of the scale.[55] Suppose, for example, in keeping with anti-proportionality, one held that the boundary on a sorties series for "being an evil person" also marked a discontinuity in the scale associated with the question, "How fitting it is to hate that person?" And suppose, in keeping with semantic plasticity, that one thought that at nearby worlds, "being an evil person" expresses different properties that draw a boundary on the same sorites series in slightly different places. A natural thought that the meaning of the question, "How fitting is it to hate that person?" also shifts in a way that is coordinate with the shift in "evil person." At the actual world, the scale associated with "How fitting it is to hate?" expresses a scale that shifts discontinuously around boundaries that mark the difference between evil and not being evil (of course the discontinuous shift may be not as dramatic as

treatments as radically different as being sent to Heaven or Hell. This issue is explored at length in Sider, "Hell and Vagueness."

55 "Penumbral connection" is the expression typically used by supervaluationists and epistemicists to mark logical, analytic, or *a priori* connections between predicates that are respected by all families of candidate interpretations. Penumbral connections obtain when individually admissible interpretations for several words are not jointly admissible.

one initially imagines because the first evil people in the series are only just evil, and the last are almost evil). Meanwhile at a nearby world where "evil" means some slightly different property, the scale associated with "How fitting is it to hate?" is different to the actual scale, by in particular having a little cliff that marks the boundary associated with that slightly different property (and thus, presumably, there is a slight shift in the meaning of "fitting" that coordinates with the slight shift in the meaning of "evil"). This view of the matter is not obviously the right one. But I submit that it is not obviously the wrong one either.

That said, I would take quite a bit of persuading in many cases to opt for anti-proportionality. Let me illustrate the relevant issues by looking at a few test cases.

Imagine someone that badly wanted a fantastic kitchen. They begin with a humdrum kitchen that gradually improves over the years, and their attitude toward their kitchen slowly evolves. At some point it is clearly fantastic. The surfaces had begun to sparkle more and more and there came a point where, given all the other changes, it just so happened that a tiny bit of extra sparkle took it over the edge to being fantastic, after which it continued to improve further. Looking back they see that there were no discontinuous shifts in their attitude to their kitchen, though by the end their attitude is extremely positive and at the beginning very negative. What should we say if, looking back, the person said to themselves, "I didn't really notice the point that marked the boundary between the kitchen not being fantastic and being fantastic. What was fitting at that point was an extra little celebration and hence a discontinuous jump in positivity. After all, what I always wanted was a fantastic kitchen!"? Someone making the speech might think to themselves, "Granted, my attitude toward my kitchen did not in fact take a noticeable jump at the point where it became fantastic. But that is because I did not know when that shift occurred. While there is no particular point in the renovation process that I can point to as a point where the shift from non-fantastic to fantastic occurred, I nevertheless know that there *was* a point where this shift did occur. And whenever it did, a marked jump in positivity of attitude was *fitting*."

A self-acknowledged desire for a fantastic kitchen is a pretty good rough and ready way to frame one's domestic ambitions. Yet I am dubious that there are good grounds for anti-proportionality here. The natural way to develop anti-proportionality for the scale of, say, fitting pride, is to suppose a penumbral connection between the question, "How much pride is fitting?" and the meaning of "fantastic," so that as the cutoff associated with the word "fantastic" moves around in nearby worlds, there is also a corresponding shift in the meaning of the question, "What level of pride in your kitchen is fitting?" (After all, it is hard to imagine that the shift from non-fantastic to fantastic would

be significant for the scale associated with the expression "the level of pride that is fitting" at a world where "fantastic" meant some other property, fantastic*.) Here, then, is the anti-proportionality vision: given the actual meaning of "the level of pride that is fitting," a certain discontinuous leap in level of pride in one's kitchen is fitting as the kitchen moves from not-fantastic to fantastic. Meanwhile, the meaning of "the level of pride that is fitting" slightly shifts its meaning at those nearby worlds, where "fantastic" slightly shifts it meaning in such a way that the meaning of "fitting pride" at *those* worlds calls for no such jump at the point at which the kitchen transitions from non-fantastic to fantastic. But, for what it's worth, the hypothesized penumbral connection does not seem to me to be especially plausible. The meaning of the question, "How much pride in your kitchen is fitting?" does not seem hostage to the semantic vicissitudes of "fantastic" in this way.

A second test case: someone slowly descends into depravity over their lifetime so that at some point they become evil and their parents become slowly more disgusted and ashamed of them. What should we say if the parents, looking back make the speech, "We didn't really notice the exact point that marked the boundary between our child not being evil and being evil. But at whatever point that shift occurred, a discontinuous jump in negative attitudes, including a marked extra degree of moral disgust, was warranted"? On reflection the case does not seem so different from the case of the kitchen. If you agree with me that it is a bit silly to posit a penumbral connection between "How much pride is fitting?" and "fantastic" that requires a discontinuous shift in what attitudes are fitting once the boundary for being fantastic is crossed, it is arguable that it is similarly implausible to posit a penumbral connection of this sort in the case of being evil.

I have voiced some hesitation about anti-continuity ideas, expressing considerable sympathy with Dougherty's idea—following Sider—that an epistemicist treatment of vague predicates "may involve scaling back the significance we place on applying these predicates."[56] However, I have provided nothing like a knockdown argument. Indeed, I do not wish to be doctrinaire here. The issues are certainly very delicate. Nor do I wish to assume a monolithic approach to continuity. For any pair of a property that can divide a physically continuous series of cases and a moral scale, one can ask whether the moral scale varies discontinuously around the boundary marked by the property or not. Perhaps a systematic defense of continuity theses for the important moral scales is possible here, where considerations of vagueness will in many cases play an important role in the defense. But, as I have indicated, anti-continuity is not

56 Dougherty, "Vague Value," 11.

dead in the water, and perhaps a selective defense of anti-continuity for some property/morally significant scale pairs is defensible. My aim here has not been to resolve this matter; it is to bring to attention a helpful way of thinking about proportionality issues that grounds them in questions of continuous variation. This approach to proportionality is one that, as far as I can tell, has not found that much life in the ethics literature thus far.

## 4. CONCLUDING REMARKS

I have found no good reason as yet to think there is anything deeply problematic about an epistemicist treatment of moral predicates. There are, to be sure, arguments in the literature to the effect that such an approach to vagueness in ethics is problematic. But these arguments are wanting. Many such arguments, if they work at all, work against any approach to vagueness that assumes classical propositional logic and some standard rules for the quantifiers, of which epistemicism is but a species of a wide genus. My hunch is that the relevant authors have not made their peace with a vital choice point: reject some mundane inference rules or else simply accept that, even when it comes to moral properties such as permissibility, tiny differences can make the difference between instantiation or non-instantiation. Nevertheless, a number of the critical discussions of epistemicism about moral vagueness point us toward hugely interesting choice points in various subdomains of ethical theorizing, ones that turn on whether to think that, as cases vary in a physically continuous ways, the boundary associated with some property marks a discontinuous shift in the values along some moral scale. Once we have made our peace with classical propositional logic and some orthodox inferences rules for the quantifiers, worries about unknowable truths come to nothing, as do worries that a tiny physical difference cannot make any moral difference. But these proportionality questions remain and the question of how to resolve them is both pressing and intriguing. I have indicated how an epistemicist approach might begin to approach them. And I would encourage those readers who opt for some alternative account of vagueness to explore proportionality issues within the framework of their favored approach.[57]

*University of Southern California and Australian Catholic University*
*jhawthor@usc.edu*

57 I am grateful to Cian Dorr, Stephen Finlay, Miriam Schoenfield, two anonymous referees, and a discussion group at Australian Catholic University for helpful comments and discussion.

## REFERENCES

Bacon, Andrew. *Vagueness and Thought*. Oxford: Oxford University Press, 2018.

Constantinescu, Cristian. "Moral Vagueness: A Dilemma for Non-Naturalism." *Oxford Studies in Metaethics*, vol. 9, edited by Russ Shafer-Landau, 152–85. Oxford: Oxford University Press, 2014.

Davies, Martin, and Lloyd Humberstone. "Two Notions of Necessity." *Philosophical Studies* 38, no. 1 (July 1980): 1–30.

Dietz, Christina. "Reasons and Factive Emotions." *Philosophical Studies* 175, no. 7 (July 2018): 1681–91.

Dorr, Cian, John Hawthorne, and Juhani Yli-Vakkuri. *The Bounds of Possibility: Puzzles of Modal Variation*. Oxford: Oxford University Press, 2021.

Dougherty, Tom. "Vague Value." *Philosophy and Phenomenological Research* 89, no. 2 (September 2014): 352–72.

Dunaway, Billy, and Tristram MacPherson. "Reference Magnetism as a Solution to the Moral Twin Earth Problem." *Ergo* 3, no. 25 (2016): 639–79.

Finlay, Stephen. *Confusion of Tongues: A Theory of Normative Language*. Oxford: Oxford University Press, 2014.

Fine, Kit. "The Possibility of Vagueness." *Synthese* 194, no. 10 (October 2017): 3699–725.

———. "Vagueness, Truth and Logic." *Synthese* 30, nos. 3–4 (April–May 1975): 265–300.

Fritz, Peter, John Hawthorne, and Juhani Yli-Vakkuri. "Operator Arguments Revisited." *Philosophical Studies* 176, no. 11 (November 2019): 2933–59.

Grice, Paul. *Aspects of Reason*. Oxford: Clarendon Press, 2001.

Hawthorne, John. "Epistemicism and Semantic Plasticity." In *Metaphysical Essays*, 185–210. Oxford: Oxford University Press, 2006.

Hawthorne John, and Ofra Magidor. "Reflections on the Ideology of Reasons." In *The Oxford Handbook of Reasons and Normativity*, edited by Daniel Star, 113–42. Oxford: Oxford University Press, 2018.

Kearns, Stephen, and Ofra Magidor. "Epistemicism about Vagueness and Meta-Linguistic Safety." *Philosophical Perspectives* 22 (2008): 277–304

McGrath, Sarah. "Moral Realism without Convergence." *Philosophical Topics* 38, no. 2 (Fall 2010): 59–90.

Parfit, Derek. *On What Matters*, vol. 1. Oxford: Oxford University Press, 2011.

Schoenfield, Miriam. "Moral Vagueness Is Ontic Vagueness." *Ethics* 126, no. 2 (January 2016): 257–82.

Sider, Ted. "Hell and Vagueness." *Faith and Philosophy* 19, no. 1 (January 2002): 58–69.

Sorensen, Roy. *Vagueness and Contradiction*. Oxford: Oxford University Press, 2001.

Unger, Peter. *Ignorance: A Case for Scepticism*. Oxford: Oxford University Press, 1979.

Williamson, Timothy. "Acting on Knowledge." In *Knowledge First: Approaches in Philosophy and Mind*, edited by J. Adam Carter, Emma C. Gordon, and Benjamin Jarvis, 163–81. Oxford: Oxford University Press, 2017.

Williamson, Timothy *Vagueness*. Milton Park, UK: Routledge, 1994.

Wright, Crispin. "Truth in Ethics." *Ratio* 8, no. 3 (December 1995): 209–26.

Yli-Vakkuri, Juhani. "Epistemicism and Modality." *Canadian Journal of Philosophy* 46, nos. 4–5 (August 2016): 803–35.

*Journal of Ethics and Social Philosophy*
VOL. 23, NO. 2 · NOVEMBER 2022
*https://doi.org/10.26556/jesp.v23i2.1957*


# FORGIVING THE MOTE IN YOUR SISTER'S EYE

## ON STANDINGLESS FORGIVENESS

*Kasper Lippert-Rasmussen*

Much recent philosophical exploration of the prerequisites of holding agents responsible has focused on the issue of standing to blame. In this article, I extend this exploration to a related, but in this respect unchartered, phenomenon: *forgiveness*. This topic lies downstream from wrongdoing and blame. Consider the following, typical sequence of events: wrongdoing occurs, the victim does (or does not) have standing to blame, they either blame or do not, we object if they blame while lacking standing (say, on the grounds that the blame is hypocritical), and eventually the relationship between wrongdoer and victim is (or is not) repaired through the victim's forgiveness of the wrongdoer. Many philosophers have examined either the act of blaming itself, or something relevant to the acquisition of standing to blame, to explain why we object that someone has no standing to blame. Here I argue that there is something that comes *after* blame for which our account of standing has implications. Specifically, I submit that one can lack standing to forgive in a way that is similar to the way one can lack standing to blame hypocritically even while abstaining from—perhaps even renouncing one's right to—blame altogether.[1]

Consider: relationship therapists report that when partners are confronted with evidence of their infidelity they sometimes go on the offensive and start to blame those they have deceived for having been unjustifiably neglectful in ways that partly explain their affairs.[2] Sometimes there is something to the counter-accusation. Imagine you are the deceived party in one of these cases. And imagine that, after pointing a finger at your past blameworthy neglect, and without having addressed the issue of her own infidelity, your partner magnanimously states that she forgives you, suggesting that this is a suitable point at

1 I thank an anonymous reviewer for helpfully suggesting this way of framing of the arguments I present in this article.

2 Meyers, "Why Cheaters Blame Their Innocent Partners."

which to end the conversation and move on. Very probably, you would want to continue the conversation, pressing *your* points, even if you accept—and even if you *say* that you accept—that your past neglect was blameworthy. Would you accept the forgiveness being offered? Most of us would dismiss it as an offer your partner has no standing to make given that the wrong she has committed against you is much greater than your wrong. In this article, I want to support the idea that there is such a thing as (not) having the standing to forgive, and I shall try to make some sense of what is going on when people dismiss forgiveness *despite* conceding that they have wronged the other party in the way for which they are being offered forgiveness.

Broadly speaking, forgiveness can be dismissed in two ways.[3] You *directly* dismiss it if you: deny that you did what you are being forgiven for; concede that you did it, but deny that it was wrongful; or, finally, concede that what you did was wrongful, but claim that you have a valid excuse for it. In my opening example none of these bases of direct dismissal capture your reason for dismissing the forgiveness of your past neglect. Your dismissal is *indirect*, because you are neither challenging the truth of the claim about blameworthiness that the forgiveness presupposes, nor challenging whether, in principle, your act is forgivable. Your dismissal is indirect, because what you are submitting is that, in virtue of facts about the forgiver, or the forgiver's relation to you, the forgiver has no standing (a notion I explain in section 2) to forgive you for your blameworthy action.

In this article, my focus is on indirect dismissals of forgiveness, and I explore these dismissals in the light of indirect dismissals of blame. Forgiving and blaming are closely connected—most obviously, because forgiving simply is ceasing to blame in the right way. Hence, if one lacks standing to blame, one also lacks standing to forgive. Or so I shall argue. Call this inference the *Simple Argument*. While the Simple Argument is one important thought underlying this article, it far from summarizes it. For instance, while the Simple Argument might make it reasonable to expect that the norms regulating blame regulate forgiveness as well, it does not establish this. Perhaps standingless forgiveness is morally wrongful for reasons other than standingless blame, or, unlike standingless blame, not wrongful at all.

In the recent literature on standing to blame, many philosophers argue that a hypocrite lacks standing to blame for an act even if that act is blameworthy, and that standingless hypocritical blame is *pro tanto* morally wrongful.[4] I shall

3 See Cohen, *Finding Oneself in the Other*, 119.

4 Cohen, *Finding Oneself in the Other*, 115–42; Dworkin, "Morally Speaking," 182–88; Fritz and Miller, "The Unique Badness of Hypocritical Blame," "When Hypocrisy Undermines Standing to Blame," and "Hypocrisy and Standing to Blame"; Herstein, "Understanding Standing"; Isserow and Klein, "Hypocrisy and Moral Authority"; McKiernan, "Standing Conditions and Blame"; Piovarchy "Situationism, Subjunctive Hypocrisy, and Standing to

defend analogous claims about forgiving: a forgiver can lack standing to forgive someone else for an act even if that act is forgivable (henceforth: the *Standinglessness Claim*); and standingless, hypocritical forgiveness—like that manifested in my opening example—is *pro tanto* morally wrongful (henceforth: the *Wrongness Claim*). I also try to defend the more cautious *Conditional Claim* that, for each of the two claims about lacking standing to blame, if that claim is true, then so is the corresponding claim about forgiveness, i.e., the Standinglessness Claim and the Wrongness Claim. As indicated, I am not aware of any previous discussions tying standing to forgive to standing to blame, though in the philosophical literature on forgiveness the question of whether standing to forgive a wrong requires one to be the victim of that wrong is familiar.[5] This question is peripheral to my concerns.

Section 1 identifies the sense of the term "forgive" at stake in this article, and Section 2 defines the relevant notion of indirect dismissal of forgiveness. Section 3 defines hypocritical forgiveness and argues that the hypocritical forgiver lacks standing to forgive, thus supporting the Standinglessness Claim. Section 4 explains why the hypocrite's standing to forgive is annulled. It appeals to the idea that hypocritical forgivers display insufficient, or deficient, commitment to the norms whose violation they are forgiving. Section 5 defends the Wrongness Claim, submitting that, like hypocritical blame, hypocritical forgiveness is wrongful because it involves relating to the recipient (person being forgiven) as an inferior. Section 6 concludes.

## 1. WHAT IS IT TO FORGIVE?

Forgiveness is a complex and varied phenomenon. However, my discussion examines the following communicative notion of forgiveness:

*F* (the forgiver) forgives *W* (the wrongdoer) for $\phi$-ing if, and only, if:

1. *F* communicates to *W* that *F* believes that *W*'s $\phi$-ing was blameworthy;

Blame" and "Hypocrisy, Standing to Blame and Second-Personal Authority"; Radzik, "On Minding Your Own Business"; Roadevin, "Hypocritical Blame, Fairness, and Standing"; Rossi, "The Commitment Account of Hypocrisy," "Feeling Badly Is Not Good Enough"; Smith, "On Being Responsible and Holding Responsible"; Statman, "Why Disregarding Hypocritical Blame Is Appropriate"; Todd, "A Unified Account of the Moral Standing to Blame"; and Wallace, "Hypocrisy, Moral Address, and the Equal Standing of Persons."

5 Hughes and Warmke, "Forgiveness," sec. 4; Piovarchy, "Hypocrisy, Standing to Blame and Second-Personal Authority," 605; Russell, "The Who, the What, and the How of Forgiveness," 2–3; Zaragoza, "Forgiveness and Standing," 612–19. See also my discussion of condition 4 in section 1.

2. *F* communicates to *W* that, henceforth, *F* either releases *W* from some or all of the duties to *F* that *W* has acquired, by $\phi$-ing, to respond to the blame for $\phi$-ing from *F* (i.e., *F* exercises, and thereby renounces, her normative power to change wrongdoer norms), *or* renounces whatever liberty rights *F* has acquired against *W* to blame W for *W*'s $\phi$-ing (i.e., *F* exercises, and thereby renounces, her normative power to change victim norms);[6]
3. The setting of *F*'s communicative act is of the right sort; and
4. *F* is either the victim of the wrongdoing or suitably related to the victim of the wrongdoing, and *W* is either the person who wronged *F* by $\phi$-ing or suitably related to the wrongdoer.[7]

On this definition, to forgive is to perform a speech act.[8] However, the extension of "forgiving" is broader than that. Specifically, there is a sense of forgiving where "forgiveness centrally concerns how you feel about the wrongdoer as a person."[9] While one might never have communicated forgiveness to the person who has wronged one, one might have forgiven her in one's heart, i.e., one might completely "dissociate her wrongdoing from the way [one feels] about her."[10] Conversely, one can perform the speech act of forgiving someone and nonetheless continue to resent one's wrongdoer for what she did.

This dual reference of "forgiving" explains why we can sometimes say, of those who have forgiven in the communicative sense, that they have forgiven insincerely. We mean that their thoughts about the wrongdoer are still very much shaped by her wrongdoing. Forgiving is an impure performative.[11] When you say "I forgive you," I can intelligibly have a skeptical thought: "You *say* you've forgiven me, but have you *really*?" Here I am exclusively interested in the pure performative sense of forgiving. By stipulation, the question "But did

6 See Nelkin, "Freedom and Forgiveness," 175–83; Pettigrove, *Forgiveness and Love*, 9–12; Riedener, "The Standing to Blame, or Why Moral Disapproval Is What It Is," 185–87; Warmke, "The Normative Significance of Forgiveness," 688, 697–99; cf. Allais, "Wiping the Slate Clean," 47–50.

7 Forgiving, in my sense, does not require any uptake by the recipient, but see Fricker, "What Is the Point of Blame?" 172; and Brunning and Milam, "Oppression, Forgiveness, and Ceasing to Blame," 15–57.

8 In the relevant terminology, my definition focuses on declarative speech acts of forgiving.

9 Allais, "Wiping the Slate Clean," 49; Adams, "Forgiveness," 294; Murphy and Hampton, *Forgiveness and Mercy*, 21.

10 Allais, "Wiping the Slate Clean," 57; Brunning and Milam, "Oppression, Forgiveness, and Ceasing to Blame," 155.

11 Austin, *How to Do Things with Words*, 83–84; see also Warmke, "The Normative Significance of Forgiveness," 694–98.

she *really* forgive me?" makes no sense: you have uttered "I forgive you" (or something equivalent), and in the context we are in there is no room for doubt about whether, in the relevant speech act–focused sense, you have forgiven me. The development of an account of standingless speech acts of forgiving is important in itself. Perhaps certain aspects of forgiveness are specific to communicative forgiveness. And it is possible that the account will also cast light on standingless emotion-centered forgiveness.[12]

Before proceeding, let me speak specifically to each of conditions 1–4. Condition 1 implies that when you inform someone who appears to have wronged you that what they did was not wrong, or was excusable, you are not forgiving them, but doing something else. You are denying that blame was merited in the first place—in which case, there is no room for forgiving either.

Those who forgive will often have previously (emotionally or communicatively) blamed. However, they may never have got quite as far as blaming. They may have felt, merely, that they were ready to blame, or would be blaming at some point. On my analysis neither of these sequences identifies a necessary precursor of forgiveness. Condition 1 requires the forgiver to express a belief to the effect that the wrongdoer has acted in a blame*worthy* way and, thus, that she is *entitled* to blame the wrongdoer, not that she actually blames the wrongdoer. This makes sense, because, on the present account, what one does when one forgives is renounce the *right to* blame (see 2). Suppose that I have never blamed my partner for a certain wrong she committed against me, and that I realize she feels bad about what she did. Surely, I can forgive her despite my never having blamed her until now. In doing this I forgo any right to blame her at a later point in time. On the other hand, if I think I had no right to blame her, I am prevented from thinking that I can renounce such a right.

Condition 2 implies that, in forgiving, one must convey to the person one forgives that one believes she did something blameworthy, and that one believes one has the standing to blame her.[13] One must convey that, in the absence of forgiveness, one would be entitled to continue, or to start, to blame and entitled to receive an uptake to one's blame: "In expressing resentment or indignation to another person, you standardly demand that she acknowledge her fault to you, or more generally, that she enter an exchange with you that

12 Some argue that blame is "incipiently" communicative: Darwall, *The Second-Person Standpoint*, 120; Fricker, "What Is the Point of Blame?" 177–80; McKenna, *Conversation and Responsibility*, 176; Smith, "Moral Blame and Moral Protest," 39; cf. Driver, "Private Blame"; Macnamara, "Taking Demands out of Blame," 151–56. The same could be true of forgiveness; on communicative forgiveness, see Warmke, "The Normative Significance of Forgiveness," 691.

13 Cf. Calhoun, "Changing One's Heart," 95; and Novitz, "Forgiveness and Self-Respect," 309–11.

constitutes her being held accountable by you or her giving account to you."[14] If I utter "I forgive" while communicating that there is nothing to forgive, or that there is but I have no standing to forgive it, I am not really forgiving. Condition 2 also ensures that forgiving is not merely ceasing to blame.[15] Forgiving is something one does, not something that merely happens to the forgiver, e.g., because she forgets all about the wrong in question or simply stops caring about it. This is trivially true of communicative forgiving, because to forgive in this sense involves performing a speech *act*.[16]

Finally, according to 2, forgiveness admits of degrees. This corresponds well with the way in which people actually forgive. In many cases, forgiveness is total, and the forgiver renounces any claim against, and any liberty rights in relation to, the wrongdoer's blameworthy action. However, forgiveness can be less than total. Thus it may be that I renounce the right to bring up your wrong as a conversational matter and start blaming you at will, but do not renounce the right to blame you again should you start blaming me for a similar wrong that I commit against a third party.

According to 3, the setting of the communicative act has to be of the right sort. Quite what that means is a complex issue that we can ignore for present purposes. However, to see the need for this qualification, suppose that I utter "I forgive you" to my wrongdoer while she points a gun to my head threateningly, leaving me in no doubt as to what will happen if I do not "forgive" her. Certainly, I have performed the locutionary act of uttering a string of words people often utter when they forgive, but given the coercion my utterance does not have the illocutionary force of forgiveness.

Condition 4 places a limit on who can perform an act of forgiving. Third parties can blame someone for their wrongdoing. Wrongdoers can blame themselves for their own wrongdoing. However, only the *victims* of the wrongdoing—or, as my definition allows, those suitably related to the victims of the

14 Riedener, "The Standing to Blame, or Why Moral Disapproval Is What It Is," 186–87.

15 See Allais, "Wiping the Slate Clean," 43–44; Brunning and Milam, "Oppression, Forgiveness, and Ceasing to Blame," 146; Hieronymi, "Articulating an Uncompromising Forgiveness," 530; Milam, "Reasons to Forgive," 243; Murphy, "Forgiveness and Resentment," 506; Pettigrove, *Forgiveness and Love*, 4, 97. Similarly, to refuse to forgive is essentially to continue to insist on the right to blame and on the duty of the blameworthy party to respond to the blame (Radzik, "On Minding Your Own Business," 583).

16 But something similar is also true of forgiving understood as an emotion. As Hieronymi points out: to swallow a pill that erases blame (as an emotion) is not to forgive in an emotion-focused sense ("Articulating an Uncompromising Forgiveness," 530). Swallowing a pill that makes one perform an act meeting conditions 1–4—assuming that 3 does not rule out this possibility on the ground that swallowing a pill means that the setting is not right—counts as forgiving.

wrongdoing—can forgive a wrongdoer.[17] They can forgive the wrongdoer, moreover, and not just anyone who is somehow (thinly) related to her.[18] As Linda Radzik puts it,

> the ability to grant or withhold forgiveness requires a special kind of standing. Some argue that only the victims of the wrong, and perhaps their close loved ones, have such standing. An employee who has been cheated by the boss can forgive, but the other co-workers are in no position to do so. Others grant that some non-victims can also have the standing to forgive or refuse to forgive, but only in virtue of a special need for support on behalf of the victim or a special obligation or relationship that the third party holds to either the victim or the wrongdoer.[19]

## 2. DISMISSING FORGIVENESS AS STANDINGLESS

Applying the notion of communicative forgiveness introduced in the previous section, I propose the following account of what it is to indirectly dismiss forgiveness as something the forgiver lacks standing to give:

> *Disjunctive View of Indirectly Dismissing Forgiveness*: *W* indirectly dismisses *F*'s forgiveness for *W*'s $\phi$-ing on grounds of lack of standing if and only if:
>
> 5. *W* denies that she has any duties to *F*, as a result of $\phi$-ing, to respond to *F*'s blaming of her for $\phi$-ing, that *F* can free her from, or
> 6. *W* denies that *F* has acquired any of the liberty rights against *W* to blame *W* for $\phi$-ing that *F* can renounce.[20]

17 See Chaplin, "Taking It Personally."

18 Murphy, "Forgiveness and Resentment," 506. If one can wrong oneself, then one can forgive oneself in the same ways that one can forgive others. This is not to deny that one can forgive oneself for wronging others, but when one does so, one does it in a sense different from that in which one forgives others for wronging oneself. Self-forgiveness, like self-blame (see Shoemaker, "The Trials and Tribulations of Tom Brady"; and Tierney, "Hypercrisy and Standing to Self-Blame"), raises interesting and complex issues of its own and I shall largely set it aside here.

19 Radzik, "On Minding Your Own Business," 582; Griswold, *Forgiveness*, 117; but see Pettigrove, "The Standing to Forgive," esp. 593–95; Walker, "Third Parties and the Scaffolding of Forgiveness," 495.

20 The rights and duties in question are conversational. Such rights and duties are different from, because less stringent than, say, the right to life and liberty and duties not to kill or enslave. Thus, while they can permissibly be enforced by silencing, or ignoring, others' utterances, they cannot be enforced with lethal force. However, this—unlike the normative structure that rights discourse imposes—is not important for present purposes.

On the disjunctive view, then, to indirectly dismiss forgiveness is to repudiate a claim that the communicative act of forgiving presupposes in virtue of 2. This is the claim that the recipient of the forgiveness has a duty, to the forgiver, to provide uptake to the forgiver's acts of blaming should she engage in such acts, or that the forgiver holds a liberty right against the recipient to blame her.[21] Accordingly, in indirectly dismissing forgiveness the intended recipient of the forgiveness claims, in effect, that the act of forgiving has misfired—the speaker's utterance is meant to have the illocutionary force of an act of forgiveness, but it fails to do so because condition 2 is not satisfied. The condition is unsatisfied because the speaker has neither a liberty right to blame nor a claim right to an uptake to her blame.[22] Accordingly, the forgiver lacks the moral authority to forgive required (as my definition of communicative forgiveness makes plain) by forgiveness. This is not to deny that unsuccessful acts of forgiveness involve uttering the same words—performing the same locutionary acts—as those uttered in otherwise similar felicitous acts of forgiveness. Nor is it to deny that to forgive one must represent oneself as having the normative authority that, according to 2, communicative forgiveness requires.[23] Indirectly, dismissible forgiveness is in many ways like an act of consenting on behalf of someone else. In the absence of special precursors, such as delegation, one does not have the normative authority to consent on another's behalf. Hence, even if one performs the same locutionary act as that involved in the corresponding felicitous illocutionary act of consenting, one still fails to consent in the relevant sense.[24] Nor, finally, does my account imply that an agent who engages in an act of infelicitous forgiveness has not wholeheartedly formed an intention to put her negative reactive attitudes to the wrongdoer behind her.

The disjunctive view has three important implications. First, it implies that when one dismisses forgiveness indirectly, one brackets the question of whether the act for which one is being offered forgiveness was blameworthy and simply denies that the forgiver has the standing to blame in the way that her forgiveness presupposes. Second, in principle indirectly dismissing forgiveness can be a rather unemotional activity. In particular, in indirect dismissals, the potential recipient of the forgiveness need not be implying that the forgiver morally ought not, all things considered, to forgive. Indeed, consistently with the disjunctive view, the standingless forgiver might be morally required to offer forgiveness (standingless forgiveness, and thus infelicitous or merely apparent—a qualification I

21 Duff, "Blame, Moral Standing and the Legitimacy of the Criminal Trial," 129, and "Responsibility and Reciprocity," 780–85.

22 Compare Lippert-Rasmussen, "Praising without Standing," 5–7.

23 See Riedener, "The Standing to Blame, or Why Moral Disapproval Is What It Is," 193, 195–96, 199–200.

24 Piovarchy, "Hypocrisy, Standing to Blame and Second-Personal Authority," 611.

take as read in my next two points) because the offer of forgiveness will turn the forgiver into an apparent moral exemplar capable of serving as an inspiration to many others. Likewise, consistently with the disjunctive account, there could be situations in which someone ought to accept forgiveness even though there is no wrong needing to be forgiven. Similarly, there may be situations in which a wrongdoer should accept forgiveness even though the forgiver is not the victim of wrongdoing and thus not the person with standing to forgive. This might be the case, for example, because the wrongdoer's self-blame is driving her toward suicide; only forgiveness from the person she falsely believes to be the victim of her wrong will prevent her from going down that route. Third, the present account is silent on whether forgiveness that fails to satisfy condition 2 is morally objectionable. Specifically, it is consistent with the possibility that an infelicitous attempt to forgive (i.e., an act that purports and was meant to be an act of forgiveness but is not) is *pro tanto* morally wrongful because, say, the speaker has culpably represented herself as possessing a certain normative authority that she in fact lacks.

## 3. HYPOCRITICAL FORGIVING

Against this conceptual background, I will now ask: Can forgiveness can be hypocritical? If it can, can the hypocritical forgiveness be appropriately dismissed, indirectly, as standingless? There is a natural way of understanding these questions. When someone mentions "hypocritical forgiveness," the sort of case likely to spring to mind is one where someone, Tartuffe style, pretends to forgive, conscious that, at heart, she will continue to nurse a grudge while aiming to appear magnanimous.[25] This is *not* the sort of hypocritical forgiveness I have in mind. Rather, the sort of hypocritical forgiveness I shall examine is the following:

> *F* hypocritically forgives *W* for $\phi$-ing, if and only if:
>
> 7. *F* attempts to forgive (in the communicative sense defined in section 1) *W* for $\phi$-ing;
> 8. *F* believes, or should believe, that there are others such that she herself has done (or would have done) things to them that are both relevantly similar to $\phi$-ing and contextually relevant;[26]

25 Crisp and Cowton, "Hypocrisy and Moral Seriousness," 343–44; see also section 2.

26 Condition 8 implies that, in cases of hypocritical forgiveness, *F* need not believe that she has $\phi$-ed in a way that wronged *W*. It suffices that *F* believes that *F* has done similar wrongs to someone, and that she does not think she has any reason to hope for forgiveness from others for these wrongs, and actually does not even see them as wrongs. What, according

9. Non-coincidentally, *F* does not suitably make herself, or accept herself being made, the target of forgiveness from others for *her* own conduct that is relevantly similar to $\phi$-ing; or
10. *F* (a) does not believe there are morally relevant differences between *W*'s conduct and her own putatively similar $\phi$-ing of the kind that justify her forgiving *W* while not making herself, or accepting herself being made, the target of others' forgiveness, nor does *F* (b) have a belief to this effect for reasons she can, or should be able to, see are not sufficient reasons.

This definition successfully captures a range of cases in which we would naturally consider the forgiveness hypocritical but for reasons other than the deception involved in the Tartuffe case. Indeed, given the definition, Tartuffe-style forgiveness may qualify as non-hypocritical forgiveness if the Tartuffe forgiver publicly and proportionately blames herself for her greater wrong while publicly forgiving the lesser wrongdoer, though at heart she has no regrets about her own greater wrong whatsoever and continues to resent the lesser wrongdoer.

Condition 7 reflects the fact that, trivially, to forgive hypocritically one has to attempt to forgive in the relevant communicative sense. Conditions 8 and 9 provide the meat of the explanation of why *F*'s forgiveness is hypocritical. Their satisfaction means that *F* fails to recognize that *W* has a right to blame *F*, and hence a right to renounce blaming *F*, with a foundation no less solid than *F*'s own putative right to blame *W*. Hence, *F* does not have the moral authority over *W* that forgiveness requires. The "would have done" in condition 8 allows for counterfactual hypocrisy. Thus, I might blame someone for something I have not done myself while also knowing that I would have done the same thing myself had I been in that person's situation.[27] Roughly speaking, condition 8 is informed by this thought: the fact that *F* has done (or would have done under relevant hypothetical circumstances) something relevantly similar to *W* undermines *F*'s right to blame *W* and demand uptake of that blame by *W*.

The purpose of conditions 9 and 10 is to exclude certain cases of hypocritical forgiveness—cases, that is, involving mere incoherence, and cases involving an assumed moral difference between one's own act of forgiveness and that

8, *F* has to believe is that she has performed a certain action, and that, whether she believes this or not, the action is both relevantly similar to $\phi$-ing and contextually relevant. *F* need not believe that she has performed an action under that description.

27 Piovarchy, "Situationism, Subjunctive Hypocrisy, and Standing to Blame."

of others who satisfy conditions 7 and 8.[28] Condition 9 is designed to rule out cases where *F* is simply incoherent and we are dealing with what we might describe as a merely incoherent forgiver. This forgiver might have as readily ended up (and with a suitable frequency does end up) blaming herself for $\phi$-ing while not blaming *W* for doing similar things to her, so it is sheer coincidence that, in this case, she ends up forgiving her victim for, say, a minor wrong committed against her while failing to see that she is a potential recipient of even more magnanimous forgiveness from the victim for her own greater wrongdoing. While such a forgiver could display various vices—incoherence, for a start—hypocrisy is not among these.[29] Accordingly, in forgiving her wrongdoer such a person might not engage in an act of (wrongful) hypocritical forgiveness.[30] A forgiver who satisfies 9 is one who does not see that she herself is an appropriate target of (more severe) blame by those she has wronged. For that reason, the normative relation between her and the person she forgives is relevantly similar to the normative relation that exists between the person she has wronged and herself. One indication that condition 9 is satisfied is that the hypocritical forgiver sees herself as magnanimous when she forgives the person she has wronged but does not see that this person—her own victim—would manifest even greater magnanimity if they were to forgive her for her greater wrong. The hypocritical forgiver might, in these circumstances, take herself to be entitled to the other's forgiveness, or simply think the other's forgiveness is not needed to repair her damaged relation to her victim.

The purpose of 10 is to exclude foreseeable defeaters of hypocrisy. It says that *F* is warranted in believing that there is a morally relevant difference

28 Conditions 9 and 10 should align with one's views about what undermines standing to blame. Different theorists might want to tweak them so that they fit their own views on this matter. I suspect Todd ("A Unified Account of the Moral Standing to Blame") would want to revise the conditions so they handle cases in which a forgiver's conduct does not manifest lack of commitment to the norm for a violation of which the forgiveness is being offered (see section 4). Rather differently, Fritz and Miller ("Hypocrisy and the Standing to Blame") might wish to adjust the conditions to handle cases in which the forgiveness at issue manifests a differential blaming position. I think 9 and 10 are capable of being developed in these ways, and that for present purposes we can set aside questions about what exactly would be required to deliver the sought-after alignments.

29 Fritz and Miller, "Hypocrisy and the Standing to Blame," 122.

30 I am assuming here that hypocrisy cannot be a wholly objective matter. Specifically, I believe that someone who forgives hypocritically must *either* have certain attitudes (e.g., the attitude of not seeing one's own wrongdoing as something that renders one a suitable target of blame and forgiveness from one's victim) *or* be in a position such that she ought to have seen that having attitudes of this kind is appropriate and that the reason why she nevertheless does not have this attitude is some kind of exception seeking in her own favor (cf. Piovarchy, "Hypocrisy, Standing to Blame and Second-Personal Authority," 618–20).

(located in the differential effects of forgiveness, for example) between her own forgiving of *W* and *W*'s forgiving of her that morally justifies her act of forgiving and justifies her, again morally, in not accepting forgiveness from *W*. Suppose, for instance, that *F* forgives *W* because *W* is psychologically fragile and consumed by guilt, whereas *F* is robust enough to live with a powerful sense of guilt. If *that* is *F*'s sole reason for forgiving *W*, while not considering herself an appropriate recipient of forgiveness, clearly *F* is not manifesting the vice of hypocrisy.[31]

In my view, hypocritical forgiving, as I have defined it, can be rightly dismissed as standingless. In support of this view I offer, first, a case of political hypocrisy, in addition to the example involving forgiveness for infidelity offered in the introduction:

> *Dresden*: Suppose that, in contrast to what actually happened, in the years after World War II the German state never apologized for Nazi atrocities but simply ignored the horrors inflicted on hundreds of millions by Hitler's regime. Suppose, with this as the background, that at a prominently staged fiftieth anniversary ceremony in Dresden town hall, counting among its invitees the Israeli ambassador, the German state through its representatives officially forgives the Allies for the militarily largely pointless terror bombing of Dresden in the final months of World War II—bombing that resulted in the deaths of tens of thousands of innocent German civilians.

Plausibly, the invitees, as the intended recipients of this forgiveness, are in a position to dismiss it as hypocritical even if they concede that the terror bombing of Dresden was blameworthy. Conditions 7–10 seem to be satisfied, 7 trivially so. Condition 8 is satisfied because the German state and its representatives know, or should know, that if the terror bombing in question was wrong, the Holocaust was a much greater wrong, and a relevant one, too, given the overall context of the Dresden attack and the invitees. On account of the systematic failure to address the wrongs of the Holocaust, 9 is satisfied in Dresden. And 10

31 One might motivate 9 and 10 by appealing to Piovarchy's analysis of lack of standing. Neither the merely inconsistent forgiver nor the forgiver who thinks there is a morally relevant difference between the wrong committed against her by the recipient of her forgiveness and the (greater) wrong the forgiver has committed against the recipient of her forgiveness makes—or thinks she is entitled to make—"a second-personal demand on others, while failing to accept the authority of others to make the same kind of second-personal demand on them" (Piovarchy, "Hypocrisy, Standing to Blame and Second-Personal Authority," 614). Both accept that others have the relevant authority, but they simply fail to notice it—in the former case as the result of a benign oversight and in the latter as the result of a mistaken belief that there are reasons to exercise that authority in the forgiver's case but not in the case of those who have wronged the forgiver.

we can assume to be satisfied, because the reason for the discrepancy between what the German state is forgiving and what it seeks forgiveness for, in connection with Dresden, are wholly explained by its own reluctance to face up to its own wrongdoing.

Having brought out the intuitive plausibility of the view that hypocritical forgiveness is standingless, I want now to offer a separate argument for the view:

11. If *F* has standing to forgive *W* for *ϕ*-ing, then *F* has standing to renounce a liberty right against *W* to blame *W* for φ-ing or standing to renounce a claim right against *W* that *W* provides uptake to *F*'s blaming *W* for *ϕ*-ing.
12. If *F* has either standing to renounce a liberty right against *W* to blame *W* for *ϕ*-ing or standing to renounce a claim right against *W* that *W* provides uptake to *F*'s blaming *W* for *ϕ*-ing, then *F* either has a liberty right against *W* to blame *W* for *ϕ*-ing or a claim right against *W* that *W* provides uptake to *F*'s blaming *W* for *ϕ*-ing.
13. If *F* has either a liberty right against *W* to blame *W* for *ϕ*-ing or a claim right against *W* that *W* provides an uptake to *F*'s blaming *W* for *ϕ*-ing, then *F* has standing to blame *W* for *ϕ*-ing.
14. So, if *F* has standing to forgive *W* for *ϕ*-ing, then *F* has standing to blame *W* for *ϕ*-ing.

This argument is clearly valid, since 11–13 are three linked conditionals and its conclusion is a conditional with the antecedent of 11 as its antecedent and the consequent of 13 as its consequent. Hence, the crucial question is whether the premises are true. Arguably, 11 follows relatively straightforwardly from 2 in my definition setting out what communicative forgiveness is, i.e., the claim that: *F* communicates to *W* that, henceforth, *F either* releases *W* from some or all of the duties to *F* that *W* has acquired, by *ϕ*-ing, to respond to the blame for *ϕ*-ing from *F* . . . ; *or* renounces whatever liberty rights *F* has acquired against *W* to blame *W* for *W*'s *ϕ*-ing. And 12 strikes me as a conceptual truth. One cannot have the standing to renounce a right unless one has that right. Finally, 13 is a plausible account of what it is for *F* to have standing to blame *W* (in a communicative sense) for *ϕ*-ing: surely, here, *either F* has a liberty right against *W* to blame *W* for *ϕ*-ing *or W* has a duty to *F* to provide an uptake to *F*'s blaming *W* for *ϕ*-ing.[32] One reason why this account of standing to blame is attractive is

32 Compare Lippert-Rasmussen, "Praising without Standing," 5–7. Not everyone accepts that there is something like standing to blame (Bell, "The Standing to Blame"; Dover, "The Walk and the Talk"; King, "Skepticism about the Standing to Blame"). For reasons of space here I am simply relying on the assumption that skepticism about standing to blame can

that when people dismiss someone as not having the standing to blame they need not be claiming that the person should not (morally) engage in blaming. After all, standingless blame (like standingless forgiveness) may be morally justified in virtue of its good consequences.

These, then, are my arguments for the claim that forgiveness can be standingless. While the first, intuitive argument, appealing to Dresden (or for that matter, the opening example of the cheating forgiver), carries greater weight for me, I think the second definition-based argument is also forceful.

## 4. WHAT UNDERMINES STANDING TO FORGIVE?

If hypocritical forgiveness is standingless, what is it about the hypocrite that undermines her standing to forgive? I think the answer to this question is the following:

> *Commitment Account*: What deprives the hypocrite of her standing to forgive others is the fact that she is not genuinely committed to the norm that her forgiveness presupposes.[33]

This account—which is meant to mirror the intuition shared by several theorists who regard commitment to a norm as necessary for standing to blame while not corresponding to any specific fleshing out of that intuition—explicates the two examples of hypocritical forgiveness I have presented in a satisfying way. Through her unwillingness to address her own infidelity, and even more vividly through her affair itself, the cheating partner manifests a lack of commitment to the norm on which her forgiveness is based, i.e.,

---

be defeated. Skeptics about standing to blame are invited to assess the argument above as one that shows what would follow, as regards standing to forgive, if there were such a thing as standing to blame.

33 Theorists who have defended a commitment account of hypocritical blame include Crisp and Cowton, "Hypocrisy and Moral Seriousness"; Friedman, "How to Blame People Responsibly," esp. 274–75, 276–77, 282; Riedener, "The Standing to Blame, or Why Moral Disapproval Is What It Is"; Rossi, "The Commitment Account of Hypocrisy" and "Feeling Badly Is Not Good Enough"; and Todd, "A Unified Account of the Moral Standing to Blame." Riedener argues that it is a constitutive rule of blaming that you "don't have the authority to blame someone in light of a norm if you don't take it seriously yourself," submitting that taking the norm seriously is exactly what the hypocritical blamer does not do ("The Standing to Blame, or Why Moral Disapproval Is What It Is," 196). Perhaps a similar analysis applies no less well to hypocritical forgiveness: that is, it is a constitutive rule of forgiving that you do not have authority to forgive someone for a violation of a particular norm unless you take it seriously, and you do not do the latter when you fail to acknowledge that your similar, or more serious, violation of the very same norm makes you someone who is also a potential, and perhaps more appropriate, target of (blame and) forgiveness.

essentially, the norm that spouses should not deceive each other and ought to support one another emotionally. Similarly, the imagined German state in the Dresden case shows a lack of principled commitment to the norm of not killing civilians. It fails to apply the norm in a case where this would reflect badly on Germany.

In other cases, however, the commitment account seems to deliver the wrong answers. In passing—light heartedly, but not hypocritically—I might forgive someone. It is fairly obvious that I care little about the wrong committed against me, and that I think of the forgiveness in a rather business-like way. Possibly, I forgive in a way manifesting no greater commitment to the norm at issue than a hypocritical forgiver does, with the difference that the latter is seriously upset about another's violation of the norm. Yet, it would seem odd to say that my standing to forgive is undermined. A case such as this seems to be a counterexample to the commitment account.

This challenge can be met by specifying the lack of commitment that undermines standing to forgive more precisely.[34] Thus, it might be suggested that one is committed to a norm in the relevant, objective sense if and only if one has always complied with the norm (or complied with it to a sufficiently high degree). On this understanding of commitment, the forgiver in the previous paragraph might be fully committed to the norm they forgive another person for violating. I suspect that this notion of commitment is far too crude. In many cases compliance with a norm is a good indicator of commitment, but it is neither necessary nor sufficient for the commitment. That it is not necessary emerges, for instance, in Friedman's acknowledgement that the weak-willed hypocrite is "fully committed to" the norm she violates.[35] That it is not sufficient is shown by the subjunctive, hypocritical blamer (or forgiver). This individual has been fortunate enough never to violate a particular norm, perhaps because she has never been in a situation where she would gain from its violation. However, had such an occasion arisen, she would have flouted the norm—indeed, she presently desires to do just that should an occasion arise—to whatever extent her self-interest dictated. This individual surely lacks commitment to the norm in question.[36] Plausibly, blame and forgiveness from such an agent can sometimes be dismissed as hypocritical.

34 Perhaps only lack of commitment biased in one's own favor, or in favor of those whom one somehow sympathizes with, undermines standing.

35 Friedman, "How to Blame People Responsibly," 281.

36 Piovarchy, "Hypocrisy, Standing to Blame and Second-Personal Authority," 619. Cf. "I will not attempt fully to analyze the sort of commitment at issue; however, it consists, minimally, in endorsement of the value as a genuine value, together with at least some degree of motivation to act in accordance with the value" (Todd, "A Unified Account of the Moral Standing to Blame," 355).

My own response to the present challenge is rather different. I wish to stress two things. First, if the present counterexample works against the commitment account, it also works against an analogous commitment account of standing to blame. Hence, it supports the Conditional Claim, i.e., the claim that for each of the Standingless Claim and the Wrongness Claim (about blame), if that claim is true, then so is the corresponding claim about forgiveness. If the commitment accounts are to be rejected both in relation to blame and in relation to forgiveness, that is some reason to think that hypocrisy might not undermine standing. At any rate, plainly, we will have stronger reason to think that hypocrisy *does* undermine standing if we can explain what it is about hypocrisy that undermines standing—whether to blame or to forgive. Second, if counterexamples of the kind I sketched above successfully defeat the commitment account, we will need an alternative explanation of what it is about the hypocrite that undermines her standing to forgive. The literature on standing to blame suggests that a widely supported candidate would be:

> *Moral Equality Account*: What deprives the hypocrite of her standing to forgive others is the fact that, in virtue of her hypocritical forgiveness, she denies or violate the moral equality of persons.[37]

The animating idea here is that hypocritical forgivers deny, or violate, the moral equality of persons because they see themselves as being in a position to blame others for minor wrongs even though they themselves have committed greater wrongs against others and fail to acknowledge those greater wrongs.

Unfortunately, this account is defeated by the case of the *hyper*critical forgiver. The hypercritical forgiver finds it very difficult to forgive herself, but very easy to forgive others. If this person treats anyone as an inferior, thereby implicitly denying, or violating, moral equality, presumably it is herself.[38] Yet, when she forgives others, they cannot dismiss her forgiveness as standingless in the light of her failure to treat herself as an equal in relation to her acts of forgiveness.

The obvious response to this objection is to embrace something like the following modification of the moral equality account:

> *Anti-Superiority Account*: What deprives the forgiver of her standing to forgive others is the fact that, in virtue of her hypocritical forgiveness, she affirms her moral superiority over other persons.[39]

37 Fritz and Miller, "Hypocrisy and the Standing to Blame," 125; Wallace, "Hypocrisy, Moral Address, and the Equal Standing of Persons," 328, 335; but see Riedener, "The Standing to Blame, or Why Moral Disapproval Is What It Is," 191.

38 Cf. Murphy, "Forgiveness and Resentment," 505.

39 Cf. Lippert-Rasmussen, "Praising without Standing," 669.

On this account, plainly, the hypercritical forgiver retains her standing to forgive. She does not affirm her moral superiority over others—far from it. Ultimately, however, the anti-superiority account is flawed, and this drives us back to the commitment account (assuming we started there). Consider two aristocrats, both of whom think that, in a wide range of cases, aristocrats should forgive wrongs done to them by other aristocrats but almost never forgive wrongs committed against them by commoners. Both, then, affirm superiority over the commoners. Suppose now that both aristocrats forgive a commoner who has committed the same minor wrong against each of them. And assume that the first aristocrat has not committed any wrongs against the commoner she is forgiving, while the second has committed much greater wrongs against the commoner than those she is forgiving. On the anti-superiority account, both commoners can indirectly dismiss the forgiveness they are being offered, since both aristocrats affirm their superior moral status relative to the commoners.[40] However, in addition to this the second commoner can legitimately claim that, because the aristocrat has wronged her to a much greater degree, she is in no position to allocate the blame in the first place, and thus in no position to forgive. Hence, what undermines the second aristocrat's position to forgive is not her denial of moral equality, but the fact that she has committed greater wrongs against the recipient of her forgiveness.

I accept that some will take issue with this objection to the anti-superiority account, and, for that matter, with my previous objection to the moral equality account. Even they, however, should accept that what undermines the standing to blame—be that a denial of moral equality or an affirmation of one's own superiority—can be present in the case of forgiveness as well. Once this is accepted, it is hard to see how friends of the moral equality, or the anti-superiority, account of standing to blame could deny that there is such a thing as lacking the standing to forgive. If this is granted, we have strong support for the Standinglessness Claim (see introduction). This claim is true whichever of the three accounts of standing to blame I have discussed in this section is correct.

40 Lippert-Rasmussen, "Why the Moral Equality Account of the Hypocrite's Lack of Standing to Blame Fails," 669–72. It might be objected that while both aristocrats affirm their own superiority explicitly, only one of them does so implicitly through her pattern of forgiveness. In reply, I must say that I fail to see how what one affirms, or denies, implicitly can undermine one's standing to perform certain acts if, when one says that very same thing explicitly (perhaps at the very moment one forgives), that does not undermine one's standing.

## 5. The Wrongfulness of Hypocritical Forgiveness

Let me now turn to the question of what makes hypocritical forgiveness wrongful. I want to defend two claims: that if hypocritical blame is *pro tanto* wrongful, then so is hypocritical forgiveness; and that hypocritical forgiveness is *pro tanto* wrongful. I defend these two claims by scrutinizing four accounts of why hypocritical blame is *pro tanto* wrongful.

In the previous section I considered the moral equality account of standing to forgive and to blame. On my conception of standing, the mere fact that your forgiveness is standingless does not in itself show that it is *pro tanto* wrongful. However, Fritz and Miller and Wallace all seem to take their accounts of why hypocritical blame is standingless to also be accounts of why hypocritical blame is *pro tanto* wrongful:[41]

> *Moral Equality Account of the Wrongfulness of Hypocritical Blame (or Forgiveness)*: Hypocritical blaming (or forgiving) is *pro tanto* wrongful because it involves the blamer's (or forgiver's) denying the moral equality of the addressee(or recipient) or treating this person as if she is not a moral equal.[42]

If this is the correct account of hypocritical blame, the analogous, parenthesized account of the *pro tanto* wrongfulness of hypocritical forgiveness is also correct. After all, on my account a hypocritical forgiver is involved in hypocritical blame (or, at least, must believe themselves to be entitled to blame where, as a matter of fact, such blame would be hypocritical). I think the moral equality account of the wrongfulness of hypocritical forgiveness captures a crucial element of what is intuitively objectionable about hypocritical forgiveness. For, intuitively, what is objectionable about the deceitful partner's forgiveness is the way in which she relates to her partner as someone whose entitlements, in relation to holding each other accountable, are lesser than her own, and that way of relating to others is built into hypocritical forgiveness by definition.[43]

41 Fritz and Miller, "Hypocrisy and the Standing to Blame," 122; and Wallace, "Hypocrisy, Moral Address, and the Equal Standing of Persons," 332.

42 Because Fritz and Miller propose an account of the wrongness of hypocritical blame only, the "(or Forgiveness)" represents an extension of their account. A similar point applies to the other instances of "(or Forgiveness)" and other parenthesized instances of "forgive" or derivatives of "forgive" in the accounts introduced in this section.

The formulation of the account here accommodates the intuition that the hypercritical blamer (or forgiver) does not act in a *pro tanto* wrongful way because she does not relate to others as a superior.

43 This is part of what makes the forgiveness in the imagined Dresden case intuitively objectionable, though other factors might be at play here as well.

Not everyone accepts the moral equality account of the wrongfulness of hypocritical blame, so let us consider three other accounts and ask how they apply to hypocritical forgiveness. In a recent article, Isserow and Klein suggest:

> *Desert Account of the Wrongfulness of Hypocritical Blame (or Forgiveness)*: Hypocritical blaming (or forgiving) is *pro tanto* wrongful because it involves doing something to acquire (or actually acquiring) more esteem in the eyes of others than one deserves in a context where attributions of faults and virtues are typically tied to comparative esteem.[44]

If this account is correct, the equivalent explanation of hypocritical forgiveness is also correct. After all, alongside forgiving another's minor fault the hypocritical forgiver omits to address her own faults in a way that seems to involve trying to acquire, or actually acquiring, more esteem than she merits: that acquisition is the upshot of her avoidance of deserved blame. Also, by actively conveying a false impression of magnanimity the hypocritical claimer lays claim to underserved esteem.

It might be objected that in *some* cases avoiding having one's esteem lowered in deserved ways, or having one's esteem boosted in undeserved ways, will move one closer to possession of the amount of esteem that one deserves. It will do so, for example, if, for other reasons, one's level of actual esteem diverges from one's level of deserved esteem. In my view, this objection might well defeat the desert account. However, in the present context I need only note that assessments of the objection will be symmetrical across the desert account of moral wrongfulness of hypocritical blame and the desert account of moral wrongfulness of hypocritical forgiveness—they will apply as powerfully, or feebly, to both.

It can also be objected that, implausibly, the desert account seems to imply that forgiveness is *pro tanto* wrongful. After all, part of what one does when one forgives is renouncing one's right to blame the wrongdoer in a way that this person actually deserves. Hence, if the forgiver acts in accordance with this renouncement, the wrongdoer receives less blame than she deserves, and therefore, probably, more esteem than she deserves. However, my account of what forgiveness involves does not speak to the issue of esteem. It is compatible

44 Isserow and Klein, "Hypocrisy and Moral Authority," 209. The context qualification is not one that Isserow and Klein themselves suggest. However, it seems that without (and perhaps even with) this restriction, their account is overinclusive. They note that since "an agent can undermine their moral authority in many ways, [their own] account construes hypocrisy as multiply realizable" (Isserow and Klein, "Hypocrisy and Moral Authority," 193). I would add that, similarly, the undermining of one's moral authority is similarly multiply realizable, and that hypocrisy is just one way in which it can be realized—as, in effect, acknowledged by Isserow and Klein, "Hypocrisy and Moral Authority," 205–6.

with it that someone who is forgiven for her wrongs should have the esteem she has in the eyes of others lowered in proportion to the wrong despite the forgiveness. Hence, even if the empirical conjecture involved in the present challenge is correct, it would not challenge the desert account.

Third, in a recent article Cristina Roadevin defends:

> *Reciprocity Account of Hypocritical Blame (or Forgiveness)*: Hypocritical blaming (or forgiving) is *pro tanto* wrongful because it involves a failure to reciprocate to the recipient of blame (or forgiveness) on the part of the blamer (or forgiver), i.e., the blamer (or forgiver) demands something from the recipient while rejecting a relevantly similar demand from her.[45]

Hypocritically forgiving someone who has wronged you while displaying disproportionately little attention to your own similar, or greater, wrongs against the recipient of your forgiveness amounts to a failure of reciprocity relevantly like that involved in hypocritical blame. One expects others to take one's own complaints against their wrongful actions seriously by accepting one's forgiveness (thereby acknowledging one's entitlement to blame), yet does not honor the expectation that one will take the similar or greater complaints of others seriously, e.g., by apologizing and asking for forgiveness. Hence, from the perspective of reciprocity, hypocritical forgiveness and blame are wrongful on exactly the same grounds.

Consider, finally, a view defended by Thomas Scanlon:

> *Falsehood Account of Hypocritical Blaming*: Hypocritical blaming is *pro tanto* wrongful because it involves the suggestion of a false claim, i.e., the claim that the blamer's and blamee's moral relationship is impaired as a result of the blamee's, not the blamer's, faults.[46]

This account can readily be generalized to cover hypocritical forgiveness:

> *Falsehood Account of Hypocritical Forgiving*: Hypocritical forgiving is *pro tanto* wrongful because it involves the suggestion of a false claim, i.e., the claim that the forgiver's and the recipient of forgiveness's moral relationship is impaired as a result of the recipient's, not the forgiver's, faults, and is now partly, or fully, repaired as a result of the forgiver's (hypocritical) forgiveness.

Again, I am not championing falsehood accounts of the wrongfulness of hypocritical blame or forgiveness. I am simply contending that the suggestion of

45 Roadevin, "Hypocritical Blame, Fairness, and Standing," 137; cf. Duff, "Responsibility and Reciprocity," esp. 780–85.

46 Scanlon, *Moral Dimensions*, 122–23, 128–29.

a false claim about what modifies the relation between the involved parties is as involved, or implicit, in cases of hypocritical forgiveness as it is in cases of hypocritical blame. The deceitful partner's forgiveness suggests that she is the party with legitimate cause to withhold goodwill and trust from her deceived partner, and therefore the one with discretion to either restore or not restore their relationship. So, if the false suggestion is wrongful in the case of hypocritical blame, the same seems true when hypocritical forgiveness is at issue.

I have now supported the Wrongness Claim—the claim that hypocritical forgiving is *pro tanto* wrongful. Such forgiving is wrongful, I have argued, because it denies the moral equality of the recipient or treats her as if she is a moral equal. I have also supported the conditional claim that if hypocritical blame is *pro tanto* wrongful, then so is hypocritical forgiveness. I have pointed out that several familiar accounts of the wrongfulness of hypocritical blame imply that, likewise, hypocritical forgiveness is also wrongful. Admittedly, this does not show that no account of the wrongfulness of hypocritical blame could imply that while hypocritical blame is *pro tanto* wrongful, hypocritical forgiving is not, but it does confer a degree of robustness on my conditional claim about the Wrongness Claim.

## 6. CONCLUSION

If the arguments in this article are sound, one can lack the standing to forgive in ways that would be hypocritical in the way I have described; certainly one can do so if, as many philosophers think, one can lack the standing to blame in this way. Hypocritical forgiveness is *pro tanto* wrongful because, like hypocritical blame, it involves denying moral equality or treating the addressee as if she is not a moral equal.[47] At any rate, if hypocritical blame is *pro tanto* wrongful for that reason, then so is hypocritical forgiveness.[48]

*University of Aarhus*
*lippert@ps.au.dk*

47 Recall that I have discussed two anti-superiority accounts: one of what undermines the standing of the hypocrite to forgive (section 4), and one of the *pro tanto* moral wrongness of hypocritical forgiveness (section 5). I reject the former account. However, I am sympathetic to the latter.

48 A previous version of this paper was presented at the Society for Applied Philosophy's annual conference on July 3, 2021. I thank Chris Bennett, John W. Devine, Nir Eyal, Alejandra Mancilla, Massimo Renzo, and two anonymous reviewers for helpful comments. This work was funded by the Danish National Research Foundation (DNRF144).

## REFERENCES

Adams, Marilyn McCord. "Forgiveness: A Christian Model." *Faith and Philosophy* 8, no. 3 (July 1991): 277–304.

Allais, Lucy. "Wiping the Slate Clean: The Heart of Forgiveness." *Philosophy and Public Affairs* 36, no. 1 (Winter 2008): 33–68.

Austin, J. L. *How to Do Things with Words*. Oxford: Clarendon Press, 1962.

Bell, Macalaster. "The Standing to Blame: A Critique." In Coates and Tognazzini, *Blame: Its Nature and Norms*, 263–81.

Brunning, Luke, and Per-Erik Milam. "Oppression, Forgiveness, and Ceasing to Blame." *Journal of Ethics and Social Philosophy* 14, no. 2 (December 2018): 143–78.

Calhoun, Cheshire. "Changing One's Heart." *Ethics* 103, no. 1 (October 1992): 76–96.

Chaplin, Rosalind. "Taking It Personally: Third-Party Forgiveness, Close Relationships, and the Standing to Forgive." In *Oxford Studies in Normative Ethics*, vol. 9, edited by Mark Timmons, 73–94. Oxford: Oxford University Press, 2019.

Coates, D. Justin and Neil A. Tognazzini, eds. *Blame: Its Nature and Norms*. New York: Oxford University Press, 2013

Cohen, G. A. *Finding Oneself in the Other*. Princeton: Princeton University Press, 2013.

Crisp, Roger, and Christopher J. Cowton. "Hypocrisy and Moral Seriousness." *American Philosophical Quarterly* 31, no. 4 (October 1994): 343–49.

Darwall, Stephen. *The Second-Person Standpoint*. Cambridge, MA: Harvard University Press, 2006.

Dover, Deniela. "The Walk and the Talk." *Philosophical Review* 128, no. 4 (October 2019): 387–422.

Driver, Julia. "Private Blame." *Criminal Law and Philosophy* 10 (June 2016): 215–20.

Duff, R. Anthony. "Blame, Moral Standing and the Legitimacy of the Criminal Trial." *Ratio* 23, no. 2 (April 2010): 123–140.

———. "Responsibility and Reciprocity." *Ethical Theory and Moral Practice* 21, no. 4 (May 2018): 775–87.

Dworkin, Gerald. "Morally Speaking." In *Reasoning Practically*, edited by Edna Ullmann-Margalit, 181–88. New York: Oxford University Press, 2000.

Fricker, Miranda. "What Is the Point of Blame?" *Noûs* 50, no. 1 (May 2016): 165–83.

Friedman, Marilyn. "How to Blame People Responsibly." *Journal of Value Inquiry* 47, no. 3 (June 2013): 271–84.

Fritz, G. Kyle, and Daniel Miller. "Hypocrisy and the Standing to Blame." *Pacific Philosophical Quarterly* 99, no. 1 (March 2018): 118–39.

———. "The Unique Badness of Hypocritical Blame." *Ergo* 6, no. 19 (2019): 545–69.

———. "When Hypocrisy Undermines Standing to Blame: A Response to Rossi." *Ethical Theory and Moral Practice* 22, no. 2 (May 2019a): 379–84.

Griswold, Charles L. *Forgiveness: A Philosophical Exploration*. New York: Cambridge University Press, 2007.

Herstein, Ori J. "Understanding Standing: Permission to Deflect Reasons." *Philosophical Studies* 174, no. 12 (December 2017): 3109–32.

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

Hughes, Paul M., and Brandon Warmke. "Forgiveness." *Stanford Encyclopedia of Philosophy* (Summer 2017). https://plato.stanford.edu/entries/forgiveness/#StanForg.

Isserow, Jessica, and Colin Klein. "Hypocrisy and Moral Authority." *Journal of Ethics and Social Philosophy* 12, no. 2 (November 2017): 191–222.

King, Matt. "Skepticism about the Standing to Blame." In *Oxford Studies in Agency and Responsibility*, vol. 6, edited by David Shoemaker, 265–288. Oxford: Oxford University Press, 2019.

Lippert-Rasmussen, Kasper. "Praising without Standing." *Journal of Ethics* 26, no. 2 (June 2022): 229–46.

———. "Why the Moral Equality Account of the Hypocrite's Lack of Standing to Blame Fails." *Analysis* 80, no. 4 (October 2020): 666–74.

Macnamara, Coleen. "Taking Demands out of Blame." In Coates and Tognazzini, *Blame: Its Nature and Norms*, 141–61.

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

McKiernan, Amy L. "Standing Conditions and Blame." *Southwest Philosophy Review* 32, no. 1 (January 2016): 145–52.

Meyers, Sheri. "Why Cheaters Blame Their Innocent Partners." news.com.au, November 29, 2013. https://www.news.com.au/why-cheaters-blame-their-innocent-partners/news-story/6b56c157ba66540524925d640565fobb.

Milam, Per-Erik. "Reasons to Forgive." *Analysis* 79, no. 2 (April 2019): 242–51.

Murphy, Jeffrie G., "Forgiveness and Resentment." *Midwestern Studies in Philosophy* 7, no. 1 (September 1982): 503–16.

Murphy, Jeffrie G., and Jean Hampton. *Forgiveness and Mercy*. Cambridge: Cambridge University Press, 1988.

Nelkin, Dana K. "Freedom and Forgiveness." In *Free Will and Moral Responsibility*, edited by Ishtiyaque Haji and Justin Caouette, 165–88. Newcastle: Cambridge Scholars Press, 2013.

Novitz, David. "Forgiveness and Self-Respect." *Philosophy and Phenomenological Research* 58, no. 2 (June 1998): 209–315.

Pettigrove, Glen. *Forgiveness and Love*. Oxford: Oxford University Press, 2012.

———. "The Standing to Forgive." *The Monist*, 92, no. 4 (October 2009): 583–603.

Piovarchy, Adam. "Hypocrisy, Standing to Blame and Second-Personal Authority." *Pacific Philosophical Quarterly* 101, no. 4 (July 2020b): 603–27.

———. "Situationism, Subjunctive Hypocrisy, and Standing to Blame." *Inquiry* (forthcoming). Published online ahead of print, January 8, 20202. https://www.tandfonline.com/doi/full/10.1080/0020174X.2020.1712233.

Radzik, Linda. "On Minding Your Own Business: Differentiating Accountability Relations within the Moral Community." *Social Theory and Practice* 37, no. 4 (October 2011): 574–98.

Riedener, Stefan. "The Standing to Blame, or Why Moral Disapproval Is What It Is." *Dialectica* 73, nos. 1–2 (March–June 2019): 183–210.

Roadevin, Cristina. "Hypocritical Blame, Fairness, and Standing." *Metaphilosophy* 49, no. 2 (January 2018): 137–52.

Rossi, Benjamin. "The Commitment Account of Hypocrisy." *Ethical Theory and Moral Practice* 21, no. 3 (August 2018): 553–67.

———. "Feeling Badly Is Not Good Enough: A Reply to Fritz and Miller." *Ethical Theory and Moral Practice* 23, no. 1 (January 2020): 101–5.

Russell, Luke. "The Who, the What, and the How of Forgiveness." *Philosophy Compass* 15, no. 3 (March 2020): 1–9.

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

Shoemaker, David. "The Trials and Tribulations of Tom Brady: Self-Blame, Self-Talk, Self-Flagellation." In *Self-Blame and Moral Responsibility*, edited by Andreas Brekke Carlsson, 28–47. Cambridge: Cambridge University Press, 2022.

Smith, Angela M. "Moral Blame and Moral Protest." In Coates Tognazzini, *Blame: Its Nature and Norms*, 27–48.

———. "On Being Responsible and Holding Responsible." *Journal of Ethics* 11. no. 4 (December 2007): 465–84.

Statman, Daniel. "Why Disregarding Hypocritical Blame Is Appropriate." *Ratio* (forthcoming). Published online ahead of print, July 13, 2022. https://doi.org/10.1111/rati.12345.

Tierney, Hannah. "Hypercrisy and Standing to Self-Blame." *Analysis* 81, no. 2 (April 2021): 262–69.
Todd, Patrick. "A Unified Account of the Moral Standing to Blame." *Nôus* 53, no. 2 (June 2019): 347–74.
Walker, Margaret Urban. "Third Parties and the Social Scaffolding of Forgiveness." *Journal of Religious Ethics*,41, no. 3 (July 2013): 495–512.
Wallace, R. Jay. "Hypocrisy, Moral Address, and the Equal Standing of Persons." *Philosophy and Public Affairs* 38, no. 4 (November 2010): 307–41.
Warmke, Brandon. "The Normative Significance of Forgiveness." *Australian Journal of Philosophy* 94, no. 4 (January 2016); 687–703.
Zaragoza, Kevin. "Forgiveness and Standing." *Philosophy and Phenomenological Research* 84, no. 3 (May 2012): 604–21.

*Journal of Ethics and Social Philosophy*
VOL. 23, NO. 2 · NOVEMBER 2022
*https://doi.org/10.26556/jesp.v23i2.1740*


# MORAL DISAGREEMENT AND PRACTICAL DIRECTION

*Ragnar Francén*

Whenever *A* judges that $\phi$-ing is morally wrong and *B* judges that $\phi$-ing is not morally wrong, they disagree. At least, this is an intuition that most of us share. It also seems quite clear that (again, intuitively) this is not only a contingent fact about *some* such pairs of moral judgments. Rather, it holds always, or by necessity, that, if we recognize one person as judging that an act is morally wrong, and recognize another person as judging that this act is not morally wrong, then we think that they disagree. This paper presents and argues for a novel account of such moral disagreements. In short, the idea is as follows. Moral judgments are attitudes that one can act in accordance and discordance with, and there is a disagreement between two moral judgments if there is at least some act that is in accordance with one judgment but in discordance with the other. I argue that this account is available for theories of moral judgments for which the standard accounts of moral disagreements are not available (e.g., contextualist/relativist theories).

The two standard types of accounts of moral disagreements both presuppose that the class of moral wrongness judgments is uniform, though in different ways. According to the *belief account*, the disagreement is doxastic: *A* and *B* have beliefs with conflicting contents. This presupposes "belief uniformity": that the content of moral concepts and beliefs is invariant between judges and contexts in such a way that, whenever *A* believes that $\phi$-ing is morally wrong and *B* believes that $\phi$-ing is not morally wrong, their beliefs have mutually inconsistent contents. Otherwise, there are at least possible disagreements between pairs of judgments, like *A*'s and *B*'s above, that the belief account cannot explain. Consequently, cognitivist views that accept belief uniformity—all forms of absolutist (aka invariantist or non-relativist) cognitivism—go hand in hand with such accounts. According to the *attitude account*, moral disagreements are non-doxastic: *A* and *B* have clashing practical attitudes, e.g., desires that cannot be satisfied simultaneously. This presupposes "attitude uniformity": that moral judgments are always accompanied by, or consist of, desire-like attitudes. Otherwise, there are at least possible disagreements between pairs of judgments

like *A*'s and *B*'s above, that the attitude account cannot explain. This account can be used by many non-cognitivist views according to which moral judgments (necessarily) consist of desires, and perhaps other views that imply that moral judgments are necessarily accompanied by desires (e.g., cognitivist views combined with some strong form of moral motivation internalism).

Both uniformity claims are philosophically controversial, however, and a number of theories about moral judgments in the recent literature—most obviously contextualist theories, but also others—imply the denial of *both* uniformity claims. (More about this in section 1.) Such theories therefore face a challenge when it comes to accounting for moral disagreements, since they cannot (at least in any simple way) use one of the two standard accounts above. More specifically, the challenge they face is that of finding (in the absence of belief uniformity and attitude uniformity) a trait that all moral judgments share, such that disagreement can be explained in terms of that trait. The new account of moral disagreement presented in this paper offers an answer to this challenge.

The idea is that a non-doxastic account is available also without attitude uniformity. Even if deontic moral judgments are not desires, and are not always accompanied by desires, it is characteristic of them that they have *practical direction* in the same sense as desires. Intuitively we think of people as *acting in accordance or discordance* with their (and our) moral wrongness judgments. This is to recognize that moral wrongness judgments have *practical direction* in the sense that they are judgments that one can act in accordance (or discordance) with. And it seems that we do, at least pre-theoretically, recognize that moral judgments can have practical direction without being (or being accompanied by) desires. For to the extent that we recognize that people can accept moral judgments without accompanying motivation and desires, we may speak of them as failing to desire (or be motivated) to *act in accordance with* their moral wrongness judgments.

The first step of my argument is to show that we can use the feature of having practical direction to account for deontic moral disagreements. According to the practical direction (PD) account, developed in section 2, there is a disagreement between two deontic moral judgments if there are acts that are in accordance with one judgment but in discordance with the other. The second step is to establish that we can make sense of the idea that deontic moral judgments necessarily have practical direction—i.e., that this is one way in which the class of deontic moral judgments is uniform—even if they are not necessarily accompanied by desires. Even if we recognize this possibility pre-theoretically, it is not obvious that it can be defended on philosophical grounds, since a plausible starting point seems to be that desires, but not beliefs, necessarily have practical direction. I will argue that we can make sense of this idea in

section 3. In section 4, I argue that the PD account can handle various types of moral disagreements that are more complex than those discussed in section 2. I also tentatively suggest a way in which the account can be extended so that it also handles evaluative disagreements (not only deontic moral disagreements) and nonmoral normative disagreements.

But first, section 1 gives a bit more background: I describe theories that reject both uniformity claims, and discuss the relevance in metaethics of an account of disagreement available for such theories.

## 1. PRELIMINARIES

Several different theories in the metaethical literature imply the rejection of both uniformity claims. Many of these can be subsumed under the label

> *Content Relativism*: Deontic moral judgments are beliefs, the content of which can vary between believers.

This includes various forms of moral subjectivism and contextualism (a.k.a. "indexical relativism" and "speaker relativism"), according to which the content of a person's moral judgments depends on her moral standard, in such a way that, e.g., "morally wrong" refers to different properties when used by different persons (such as failure to maximize well-being for some and lack of respect of autonomy for others).[1] It also includes "moral culture relativism" according to which the content of moral judgments can vary between cultures or societies, depending on which moral values govern the particular societies. Further, it includes "metaethical pluralism," according to which different metaethical analyses can be correct for different people's moral judgments, so that, e.g., when some people accept moral claims they have beliefs about nonnatural, *sui generis* normative properties of actions, while others have beliefs about some natural properties of actions (such as well-being maximization).[2] All of these content-relativist views imply that the content of moral beliefs varies in such a way that when *A* thinks that $\phi$-ing is wrong, and *B* thinks that $\phi$-ing is not wrong, then, in contrast to surface appearance, the propositions that they believe need not be inconsistent. Furthermore, even though they are often combined with the idea that moral judgments are intimately connected to motivation and desires, they almost always allow for possible cases of moral judgments that

1 Dreier, "Internalism and Speaker Relativism"; Harman, "Moral Relativism Defended" and "Moral Relativism"; Phillips, "How to Be a Moral Relativist"; Prinz, *The Emotional Construction of Morals*; Wong, *Moral Relativity*.

2 Francén, *Metaethical Relativism* and "Moral and Metaethical Pluralism "; Gill, "Metaethical Variability, Incoherence, and Error" and "Indeterminacy and Variability in Metaethics."

are not accompanied by desires. Thus, they reject both uniformity claims, and can account for moral disagreement in neither of the two standard ways.[3]

Some other views also reject both uniformity claims. One is "variationism," according to which some moral judgments are beliefs and some are desires.[4] Furthermore, there are forms of non-cognitivism according to which moral judgments are identified with a cluster of dispositional tendencies that normally come together—of which the disposition to be motivated or have a desire is one—but that allow that there can be moral judgments without all dispositional tendencies being in place.[5] To simplify the discussion in what follows, I will often use content relativism as my example when I present the PD account, though it works for the other kinds of views as well.

Theories that reject both uniformity claims are minority views among metaethicists. But the availability of a satisfactory account of moral disagreement for such theories should not interest only their proponents. First, the alleged failure to account for disagreement is probably the reason against accepting a view of this sort most commonly cited by non-cognitivists (who accept attitude

3 There are also more recent forms of "assessor relativism" according to which the propositional contents of moral beliefs do not vary between different persons, but the content is such that it can be true relative to some people's moral standards and false relative to others' (Brogaard, "Moral Contextualism and Moral Relativism" and "Moral Relativism and Moral Expressivism"; Kölbel, "Indexical Relativism versus Genuine Relativism" and "Moral Relativism"; Shafer, "Constructivism and Three Forms of Perspective-Dependence in Metaethics" and "Assessor Relativism and the Problem of Moral Disagreement"; Egan, "Relativist Dispositional Theories of Value"). It has been suggested that such theories, in contrast with content relativism, can explain disagreement intuitions. First, there is a common content to disagree about, and second, from each person's perspective—when she is assessing the two statements from her own moral standard, at most one of the two statements can be true (Brogaard, "Moral Contextualism and Moral Relativism" and "Moral Relativism and Moral Expressivism"; Kölbel, "Moral Relativism"; MacFarlane, "Relativism and Disagreement"). But this has also been contested. Suppose that *A* and *B* disagree over whether it is always morally wrong to lie. According to assessor relativism, the proposition that it is always wrong to lie can be true or false only relative to specific standards. If it is true relative to *A*'s standard and false relative to *B*'s standard, then *A* will judge it to be true and *B* judge it to be false—which is why we have what might seem like a disagreement. But *A* and *B* may still agree about the truth value of the relevant proposition (lying is always wrong) relative to each specific standard: e.g., that it is true relative to *A*'s and false relative to *B*'s standard. If so, they do not disagree about its truth value at all—after all, it has truth values only relative to standards, and they agree about these truth values. Cf. Dreier, "Relativism (and Expressivism) and the Problem of Disagreement"; Francén, "No Deep Disagreement for New Relativists." If this is correct, assessor relativism cannot account for moral disagreement as disagreement in belief.

4 This is suggested by Gill, "Metaethical Variability, Incoherence, and Error" and "Indeterminacy and Variability in Metaethics"; Loeb, "Moral Incoherentism."

5 Björnsson and McPherson, "Moral Attitudes for Expressivists and Relativists."

uniformity) as well as absolutist cognitivists (who accept belief uniformity). Hence, an important part of the defense of both non-cognitivism and absolutist cognitivism depends on this failure.

Second, and relatedly, arguments against each uniformity claim are highly influential in metaethics. Many non-cognitivists and contextualists/relativists have argued against belief uniformity based on the diversity of moral opinions (between individuals, groups, and/or communities). This kind of argument goes roughly as follows: the (actual or potential) diversity in moral opinions—or differently put, the fact that people apply moral terms and concepts to different kinds of acts—indicates (e.g., shows or is best explained by) that when different people use moral terms and concepts, they do not always, or at least need not, refer to the same properties. Thus, when two people accept the same moral claim—e.g., both judge that it is morally wrong to eat meat—the cognitive content of their moral judgments may differ.[6] Such arguments are, of course, controversial. Absolutist cognitivists contend that diversity does not, in the end, support the rejection of belief uniformity. But diversity arguments and their conclusion nonetheless represent one main strand in metaethics.

Attitude uniformity is also highly contested. This is clear from the debate over moral motivation internalism and externalism. Externalists argue that there can be moral judgments that are entirely unaccompanied by motivation and desires.[7] Most motivational internalists also argue that moral judgments and desires/motivation can come apart under certain conditions, for example, when the judge is practically irrational, or if her judgment is part of a moral practice where most moral judgments motivate.[8] If either of these views is correct, some moral judgments are unaccompanied by desires—which means that an account of moral disagreement in terms of clashing desires will fail

6 Arguments of this kind can be found in, e.g., Blackburn, *Spreading the Word*; Harman, "Moral Relativism"; Horgan and Timmons, "New Wave Moral Realism Meets Moral Twin Earth"; Tersman, *Moral Disagreement*; Wong, *Moral Relativity*; Hare, *The Language of Morals*. There are also less direct arguments against belief uniformity. It has, for example, been argued that theories that imply the rejection of a stable belief content better explain the connection between moral judgments and motivation (Dreier, "Internalism and Speaker Relativism"; Prinz, *The Emotional Construction of Morals*).

7 Brink, *Moral Realism and the Foundations of Ethics*; Gert and Mele, "Lenman on Externalism and Amoralism"; Mele, "Internalist Moral Cognitivism and Listlessness"; Roskies, "Are Ethical Judgments Intrinsically Motivational?"; Stocker, "Desiring the Bad"; Svavarsdóttir, "Moral Cognitivism and Motivation."

8 For views of the former kind, see Smith, *The Moral Problem*; van Roojen, "Moral Rationalism and Rational Amoralism"; Wedgwood, *The Nature of Normativity*; Korsgaard, "Skepticism about Practical Reason." For views of the latter kind, see Bedke, "Moral Judgment Purposivism"; Dreier, "Internalism and Speaker Relativism"; Tresan, "The Challenge of Communal Internalism."

to account for instances of disagreement where such judgments are involved. In other words, externalists and many internalists reject the form of attitude uniformity required for the attitude account to cover all moral disagreements.[9]

Proponents of each uniformity claim have used considerations such as those above to argue against the other uniformity claim. But to the extent that we find some plausibility in both kinds of arguments—against attitude uniformity and against belief uniformity—this lends (*prima facie*) support (and has indeed been used to support) theories that imply the rejection of both uniformity claims. Given this, removing what by many is seen as the main reason against accepting such theories—i.e., their alleged inability to account for moral disagreements—*might* alter our conclusion about which kind of theory gains most overall support from the arguments.

The main aim of this paper, then, is to develop a novel plausible account of moral disagreements, the PD account, which requires neither belief uniformity nor attitude uniformity, thus giving theories that reject both uniformity claims a way of explaining our disagreement intuitions. It should be noted that other suggestions have been made as to how, e.g., content relativism can explain disagreement intuitions. Such proposals include explanations in terms of metalinguistic negotiations and presuppositions of shared standards.[10] I will *not* try to evaluate such previous proposals in this paper, but will merely observe that it is controversial whether they succeed.[11] Also, several philosophers have suggested that moral contextualists/relativists could explain disagreements in terms of clashing practical attitudes.[12] I will not examine the details of these proposals, but as noted above, it is hard to see that they can escape the problem

9 This is not to say that all forms of internalism according to which moral judgment and motivation can come apart imply that attitude accounts of moral disagreements fail. There are non-cognitivists who argue that moral judgments are desire-like states and that such states are dispositions to motivate under normal circumstances See, e.g., Björnsson, "How Emotivism Survives Immoralists, Irrationality, and Depression"; Eriksson, *Moved by Morality*; Gibbard, *Thinking How to Live*.

10 For proposals of the first kind, see Plunkett and Sundell, "Disagreement and the Semantics of Normative and Evaluative Terms," "Dworkin's Iinterpretivism and the Pragmatics of Legal Disputes," and "Antipositivist Arguments from Legal Thought and Talk"; Bolinger, "Metalinguistic Negotiations in Moral Disagreement." For proposals of the second kind, see Dreier, "The Supervenience Argument against Moral Realism"; Harman, "What Is Moral Relativism?"; López de Sa, "Presuppositions of Commonality" and "Expressing Disagreement."

11 Marques, "What Metalinguistic Negotiations Can't Do"; MacFarlane, "Relativism and Disagreement"; Finlay, "Disagreement Lost and Found"; Marques and García-Carpintero, "Disagreement about Taste"; Beddor, "Subjective Disagreement."

12 Björnsson and Finlay, "Metaethical Contextualism Defended"; Dreier, "Transforming Expressivism" and "Relativism (and Expressivism) and the Problem of Disagreement"; Marques, "Doxastic Disagreement" and "Disagreeing in Context"; Harman, "Moral

that some cases of intuitive disagreement are left unexplained, *as long as* moral judgment and non-cognitive attitudes sometimes come apart.[13] This is *at least* a *prima facie* problem for these views, a problem that motivates the search for alternative solutions. Perhaps it can be argued that *most* disagreement intuitions can be explained by an attitude account and that those that cannot can be explained away.[14] I see no principled reason to reject such debunking explanations, but as with the other suggestions above, I will not try to evaluate this strategy here. Instead, this paper focuses on presenting and developing a positive case in favor of the new PD account of moral disagreement: a non-doxastic account that does not require practical attitudes to be present for a moral disagreement to occur—and that therefore does not require explaining away certain disagreement intuitions.

## 2. THE PRACTICAL DIRECTION ACCOUNT OF MORAL DISAGREEMENT

### *2.1. Practical Direction*

Deontic moral judgments have practical direction in the following sense: intuitively, there are ways of acting that are to act in accordance with, or follow, them. If I judge it morally right or obligatory to give to charity, then if you do so, you act in accordance with my judgment. If I believe that it is morally wrong to steal but still do it, then I act against (or in discordance with) my own judgment. I take this to be fairly uncontroversial; this is how we intuitively think and talk about deontic moral judgments.

---

Relativism"; Sundell, "Disagreements about Taste"; Wong, *Moral Relativity*; Finlay, *Confusion of Tongues*; Jackson, "The Argument from the Persistence of Moral Disagreement."

13 The most elaborate and complex account of this sort is probably that of Finlay, *Confusion of Tongues*, ch. 8. However, in the end, since Finlay characterizes "fundamental disagreements as involving a basic conflict in preferred ends" (234), his view seems to require the presence of the relevant practical attitudes (preferences)—i.e., attitude uniformity—for this kind of disagreement to occur. (For a more elaborate argument to the effect that contextualists cannot explain disagreements in terms of clashing practical attitudes, see Eriksson, "Explaining Disagreement.")

14 Could content relativists instead hold that, in cases where moral judgments are not accompanied by the relevant practical attitude (assuming that we acknowledge the possibility of such cases), we do not have disagreement intuitions to start with? This is much less plausible. As long as we really identify one person, *A*, as holding that eating meat is wrong, and another, *B*, as holding that it is not wrong, even if we stipulate that they lack certain additional practical attitudes, the (pre-theoretically) intuitively plausible verdict is that they disagree about whether eating meat is wrong. Of course, one's intuitions *might* start to waver as an effect of accepting content relativism, but then the pre-theoretic intuition still needs explaining. (Thanks to an anonymous referee for bringing up both this and the explaining-away strategy.)

Moral judgments share this trait with a bunch of other things. *Desires* have practical direction in the relevant sense: intuitively, there are ways of acting that are to act in accordance with, or follow, them. For example, if I eat ice cream, I act in accordance with my desire to eat ice cream. Many other things also have practical direction in this sense: they are things that we intuitively think that one can act in accordance or discordance with. One can act in accordance or discordance with verbal *orders*. One may succeed or fail to follow, or build in accordance with, *blueprints* for building constructions. Similarly, we can succeed or fail to assemble furniture in accordance with the *assembly instructions*. When we play chess (or any other game), making certain moves is to follow (i.e., play in accordance with) the *rules*. If we have a *shopping list*, shopping certain things is to shop in accordance with the list.

For all of these things, we experience them as directing us toward (or against) certain specific courses of actions (e.g., shopping for certain groceries, assembly of certain parts in a certain order, performance of actions judged obligatory), and the performance of these actions is then experienced as being in accordance (or discordance) with the thing.[15] We will return to the question of what it is that makes them things that we think of as having practical direction in this sense, what it is that unites them. For now, let us just grant that, *intuitively*, certain mental attitudes and other things are such that there are ways of acting that are to act in accordance with them.

Most things do not have practical direction in the intended sense—my kitchen table does not point me toward a specific action that would be to act in accordance with it, neither does the sun, the Eiffel Tower, the number three, etc. More to the point, in contrast to desires, ordinary descriptive beliefs do not have practical direction in this sense. Consider the belief that grass is green, or the belief that all horses can fly. These beliefs can be true, if they represent reality correctly. But just like my kitchen table (and the other things above), they do not point to specific ways of acting that would, intuitively, be to act in accordance, follow, or comply with them. The difference between beliefs and desires that I am after here is related to the common idea that desires, unlike beliefs, are attitudes that can be fulfilled or satisfied, rather than true. This I take it is another way of capturing the idea that desires, but not beliefs, have a practical implication, in the sense that they in some way seem to give us direction, or point to certain specific actions (or types of actions) such that doing those actions would be to comply with the attitude.[16]

15 "Specific courses of actions" should be understood as including specific token actions, specific types of actions, or specific sets of (types of) actions.

16 The distinction is also plausibly related to the idea that desires and beliefs have different directions of fit. I will not pursue this similarity here, however. The notion of directions

The idea can be further clarified by considering two potential concerns to the effect that ordinary descriptive beliefs also can have practical direction. If I believe that Alva will go for a swim tomorrow, and Alva actually does this, has she not acted in accordance with my belief? In one sense she has—she has acted as I believed she would. But this does not mean that the belief has practical direction in the sense intended above. My belief merely aimed to represent reality, not point out an act as one to be done. So it does not make sense to say that Alva could have *followed* or *complied with* the belief, in the sense of doing what it told her to do. In this way it differs from a judgment that Alva ought to swim tomorrow.[17]

Consider next my belief that a lion will kill me unless I run away. When I run, do I not act in accordance with that belief? This is also different from the things that have practical direction discussed above. Running away is not to act in accordance with belief *considered by itself*. That belief *by itself* does not point to any specific action, which is to follow the belief. This is obvious from the fact that I might desire to be killed by a lion. Contrast this with, e.g., a blueprint. If the blueprint specifies that bricks are to be used for the façade, then even if you hate brick façades, *in order to follow the blueprint*, you need to use bricks. That is, there is some act (way of building) that is in accordance with the blueprint *considered by itself*, irrespective of your goals, etc. Deontic moral judgments also have practical direction in this sense: if I judge $\phi$-ing morally wrong (right) then $\phi$-ing is to act in discordance (accordance) with the judgment *considered by itself*.

### *2.2. The Practical Direction Account of Moral Disagreement*

Deontic moral judgments can concern what we are morally allowed, obliged, and disallowed to do. Which acts are in accordance and discordance with deontic moral judgments of these different kinds? Plausibly, $\phi$-ing is to *act in discordance* with judgments to the effect that $\phi$-ing is morally disallowed—e.g., that it is morally wrong to $\phi$, that you ought not $\phi$, or that it is morally obligatory not to $\phi$. Of course, $\phi$-ing is to *act in accordance* with judgments that "tell you to $\phi$," i.e., moral judgments to the effect that $\phi$-ing is morally required, or that $\phi$-ing is something that you ought to do. But we should also say that $\phi$-ing is to

of fit is highly theoretically contested. Depending on how you understand the idea of direction of fit, the claim that moral judgments have world-to-mind direction of fit will be highly controversial, whereas I hope that most can agree that, in some intuitive sense, $\phi$-ing is to act in accordance with your judgement that $\phi$-ing is morally obligatory.

17 Note that also, e.g., deontic judgments about past actions have practical direction in the relevant sense. The judgment that it was wrong of Alva to swim yesterday states that Alva's swimming was not a thing to be done, so not swimming would have been for Alva to act in accordance with the practical implication of the judgment. (This also means that, if Alva had (magically) known about the judgment prior to her act, she could have followed the judgment in the sense of choosing to avoid the act that it states is not to be done.)

act in accordance with weaker moral judgments: judgments to the effect that $\phi$-ing is morally permissible (or not impermissible, i.e., not wrong). If you think that it is morally OK to eat tomatoes, then eating tomatoes is in line with that judgment. To summarize:

> *Discordance*: Acting in discordance with a moral judgment, MJ, is to act in a way that MJ says (or implies) is *disallowed*.
>
> *Accordance*: Acting in accordance with MJ is to act in a way that MJ says (or implies) is morally *allowed* (or *not morally disallowed*).

If a judgment has no implications about the moral status of $\phi$-ing—e.g., the judgment that snow is white—then $\phi$-ing is not to act in accordance (or discordance) with the judgment.

Here is a first stab at a criterion of disagreement in terms of practical direction: there is a moral disagreement between two persons if they accept moral judgments such that one cannot act in accordance with both.[18] Consider this pair of judgments:

> *K*: In situation *S*, it is morally wrong to kill *Q*.
>
> *U*: In situation *S*, it is morally wrong not to kill *Q*.

If I am in situation *S*, then I cannot act in accordance with both of these judgments. According to our initial criterion, then, there is a conflict between

18 Note the similarity between this preliminary statement of the PD account of disagreement and Stevenson's famous idea of disagreement in attitude, that is, the idea that moral disagreement is disagreement in the sense that it "involves an opposition of attitudes both of which cannot be satisfied" (Stevenson, "The Emotive Meaning of Ethical Terms" 2). The discussion below shows that Stevenson's account, just like the preliminary version of the PD account, needs to be amended to handle many cases of moral disagreement. Furthermore, while Stevenson's ambition was to provide an account of moral disagreement that was plausible given non-cognitivism, my aim is to argue that the PD account is available even if moral judgments are not constituted by (or necessarily accompanied by) desires, since moral judgments, even if they are not desires, are mental states that one can act in accordance with. Since this is my objective, it makes more sense to state the account in terms of "states that one can act in accordance with" rather than "states that can be satisfied." When we talk about desires, both expressions sound felicitous: "eating that ice cream *satisfied* (or fulfilled) my desire to eat something sweet" and "I thereby acted in accordance with my desire." For moral judgments, it sounds strange to say that, by donating to charity, I *satisfy* my judgment that one ought to do so. But it sounds perfectly felicitous to say that I acted in accordance with my judgment. (The same holds for chess rules, shopping lists, blueprints, etc.) We might think that this is just a choice of words, but to make the account plausible, it matters. My goal is to make sense of our disagreement intuitions, and to do that I aim to show that thinking of moral judgments as things that one can act in accordance with makes intuitive sense, and that disagreement can be accounted for in terms of this notion.

the judgments, and *K* and *U* disagree. There are two problems with this initial criterion. The first is that the criterion overgeneralizes. Consider *X*'s judgment that grass is green. There is no disagreement between *K*'s judgment and *X*'s judgment. But it does hold that one cannot act in accordance with both—for the simple reason that one cannot act in accordance at all with the latter. What this makes evident is that the initial criterion fails to capture the idea that there is a disagreement if two judgments have practical directions that clash with each other—since it can be fulfilled without one of the judgments even having practical direction. So we might consider the idea that there is a disagreement between two judgments if one cannot act in accordance with one judgment without acting in discordance with the other. On this criterion, there is a disagreement between *K* and *U*, but not between *K* and *X*.

But this amendment runs into the second problem. Consider the following two judgments.

> *U*: In *S*, eating meat is morally wrong.
>
> *K*: In *S*, eating meat is not morally wrong (both doing so and not is permissible).

Clearly there is a disagreement here. But in this case, it is possible to act in accordance with one judgment without acting in discordance with the other. If I eat vegetarian, then I have acted in accordance with *U*'s judgment. But I have not acted in discordance with *K*'s judgment. This illustrates that for two normative judgments to be in conflict it need not hold for every action that they give different verdicts about whether it should be performed—it suffices that they give different verdicts for some action. Consequently, what we should say is that there is a practical direction disagreement between two moral judgments, as long as there is at least some act that is in accordance with one judgment but in discordance with the other judgment. In the case at hand, eating meat is such an action.

> *Practial Direction Account*: There is a deontic moral disagreement between *A* and *B* if and only if they accept moral judgments respectively ($J_A$ and $J_B$) such that there is at least some act that is in accordance with $J_A$ and in discordance with $J_B$.

This account, it seems, captures the practical dimension of deontic moral disagreements: they are disagreements about how to act, and plausibly two people are in such a disagreement if they accept judgments such that some ways of acting are in accordance with one judgment but in discordance with the other.

The existence of such acts shows that the two judgments have, as it were, clashing practical implications (with regard to some action). There is at least some action that, if it is performed, will be in line with one of the judgments but go against the other.

Let us see how the PD account explains some simple cases of deontic moral disagreement.

$J_1$: $\phi$-ing is wrong (or: you ought not $\phi$; not $\phi$-ing is obligatory).

$J_2$: $\phi$-ing is not wrong (or: is permissible, is right, or is obligatory).

The PD account implies that there is a disagreement here: $\phi$-ing is in accordance with $J_2$ but in discordance with $J_1$.

$J_3$: $\phi$-ing is morally obligatory.

$J_4$: $\phi$-ing is not morally obligatory.

Here, the PD account implies that there is disagreement since not-$\phi$-ing is in accordance with $J_4$ (because $J_4$ implies that not $\phi$-ing is permissible), and in discordance with $J_3$.

$J_5$: $\phi$-ing is morally wrong.

$J_6$: Not-$\phi$-ing is morally wrong.

On the one hand, if genuine moral dilemmas are possible, then there is no disagreement, since both $J_5$ and $J_6$ can hold (indeed, one person can accept both). And this is accurately captured by the PD account. $\phi$-ing is in discordance with $J_5$, but not in accordance with $J_6$. For under the presupposition that genuine moral dilemmas are possible, $J_6$ is silent on $\phi$-ing; it does not imply that $\phi$-ing is morally permitted. On the other hand, if genuine moral dilemmas are not possible, then there is a disagreement, which is captured by the PD account: the only way to act in accordance with $J_5$ is to not $\phi$, which is to act in discordance with $J_6$. I am not taking a stand on whether genuine moral dilemmas are possible or not, but either way, the PD account captures the presence/absence of disagreement.

The PD account can capture our disagreement intuitions about these cases since they can be construed as disagreements about what to do. In section 4, we will look at deontic moral disagreements that are less straightforward disagreements about how to act, and I will argue that the PD account can nonetheless handle those cases.

I have now argued that the idea that deontic moral judgments have practical direction, in the intuitive sense described in section 2.1, can be used

to explain moral disagreement intuitions. In section 3, I give this intuitive explanation more substance, through presenting a theory about what it is that makes moral judgments (and certain other things) have practical direction. There I will also argue that moral judgments can have practical direction even if they are not desires, so that the PD account is available to content relativists. Before that, however, let us consider a potential objection that helps clarify the account.

### *2.3. Endorsement*

According to the PD account, there is a deontic moral disagreement whenever the practical directions of moral judgments clash. It might seem that this view is committed to the more general view that all clashes in practical direction between mental attitudes constitute disagreements. But then there are counterexamples. Imagine a drug addict who wants to take heroin and his father who wants him not to. Do they, by virtue of their wanting different things—and the fact that there is some act, i.e., taking heroin, that is in accordance with one of these attitudes and in discordance with the other—disagree about something, e.g., about what the drug addict is to do? Not necessarily, it would seem. The drug addict might well desire that he not want to take heroin and think that he should not. So, they might want inconsistent things without disagreeing. Arguably, this holds in many cases of clashing desires.

But there is a good reason to think that all pairs of moral judgments with clashing practical directions constitute disagreements (as the PD account holds), even though this is not so for all clashing desire pairs. For the clash of two attitudes to constitute disagreement, they plausibly have to involve endorsement (or taking a stand) of some sort. To *judge* that it is wrong to eat meat, in contrast to just entertaining the idea that it is wrong, is to endorse (or accept) the moral claim in question. In contrast, the drug addict's desire to take heroin does not involve endorsement. We can compare with disagreements about factual matters, conceived of as clashing propositions. There is no disagreement between someone who merely entertains the thought that grass is green and someone who entertains the thought that grass is blue—because entertainings, in contrast to beliefs, do not involve endorsement.

I do not have a view about what endorsement is, or in which way moral judgments involve endorsement, given the kind of content relativism and other views that are in focus in this paper. But I take it that any theory about moral judgments that is not a nonstarter will have to be able to say that *accepting* a moral claim involves endorsement in some sense. If it can, it can avoid the implication that all cases of attitudes with clashing practical direction are disagreements.

## 3. PRACTICAL DIRECTION WITHOUT DESIRES

### *3.1. The Challenge: Explaining Practical Direction*

It is part of the aim of this paper to show that the PD account of moral disagreement can be used by content relativism and other theories about moral judgments that imply the denial of both belief uniformity and attitude uniformity. For this to work, these theories about moral judgments must be able to accommodate the view that deontic moral judgments necessarily have practical direction, otherwise the PD account of moral disagreement cannot explain all cases of deontic moral disagreement. But—and this is the challenge in this section—there is *prima facie* reason to doubt that the theories in question can do this.

On the face of things, there seem to be only two straightforward ways to get the result that moral judgments necessarily have practical direction. First, since desires (but not beliefs) have practical direction, one way would be to hold that moral judgments are, or are necessarily accompanied by, desires. But that is not available to views that reject attitude uniformity, i.e., views that say that (all or some) deontic moral judgments are beliefs not accompanied by desires. Second, nonnaturalists can hold that since moral judgments are beliefs about special normative facts—i.e., facts that are categorically prescriptive, as Mackie put it—these facts have practical direction built into them.[19] They could hold that moral judgments have practical direction indirectly, through being beliefs in facts with practical direction. Most variants of content relativism cannot hold this, however. While strictly nonnaturalist forms of content relativism are possible, the variants of content relativism that are actually defended in philosophical literature do not take this form. Rather, they tend to imply that (all or most) moral beliefs are beliefs with a naturalistic content. And I want to suggest an account of moral disagreements that works for these theories. Thus, neither of these two straightforward explanations of the practical direction of moral judgments is available for content relativism. This means that these theories face a challenge of explaining the practical direction of deontic moral judgments.[20]

In the next two subsections, I will argue that, once we understand in virtue of what certain attitudes (and other phenomena) have practical direction, this challenge can be answered.

19 Mackie, *Ethics*.

20 We should also note that if some theory of moral judgments fails to explain the practical direction of moral judgments, this does not only take away a possible explanation of moral disagreement, but it is also a serious shortcoming in itself. As we noted above, it seems to be a platitude that moral judgments necessarily have practical direction. This, then, is a challenge for all theories that imply that moral judgments are beliefs about naturalistic properties and are not necessarily accompanied by desires.

*3.2. An Account of Practical Direction*

What does it take to have practical direction—that is, what makes us experience some things as things that one can follow, or act in accordance with? We can start by asking in virtue of what desires have practical direction. One thought is that desires have practical direction—i.e., are such that we can act in accordance with or follow them—in virtue of being motivational states. In contrast to beliefs, they are attitudes that intrinsically and/or necessarily motivate (or dispose) us to act in certain specific ways. One might then think that this is what makes us experience acting in those ways as acting in accordance with the desire. This would also explain why ordinary beliefs do not have practical direction, since ordinary beliefs do not necessarily motivate (at least on a standard Humean picture of motivation). If this view about practical direction is correct, then views that imply that moral judgments are not necessarily motivating also imply that they do not necessarily have practical direction.

But if we broaden our perspective, we see that the simple view above is incorrect. As noted above, some things other than desires have practical direction. For example, we can act in accordance or discordance with juridical laws, build in accordance with blueprints, play in accordance with chess rules, shop in accordance with shopping lists, and assemble furniture in accordance with assembly instructions. But these phenomena do not necessarily or intrinsically motivate anyone to act in accordance. There can be assembly instructions and laws that no one is motivated to follow—but they would still be assembly instructions and laws, and thus the sorts of things that *can be followed*. Consequently, the simple motivation view is too crude: once we have identified types of things that have practical direction, we clearly can think of at least individual tokens of those types (e.g., individual laws or assembly instructions) that people are not motivated to act upon, but that we still intuitively think of as things that one can act in accordance with.

It seems to me that our initial view was on the right track, however, in that it focuses on action-influencing tendencies. Even though many types of things that have practical direction do not necessarily or intrinsically motivate people to act, what unites them is that they are types of things whose function for us is to influence actions in specific ways. They are either mental attitudes—like desires—or human communicative acts or constructions—like blueprints, chess rules, juridical laws, requests, and orders—that point out certain ways of acting, and, in different ways, function to guide or influence people to perform (or not perform) those acts. The idea is that the fact that the main function or role they have for us is to get people to act in specific ways, explains (and makes

sense of) why we think of those ways of acting as ways of acting in accordance with, complying with, or following the desires, blueprints, chess rules, etc.

More precisely, in order to explain the fact that we think of the blueprint and law, etc., that no one is motivated to use as having practical direction, we should say that when a *type* of thing—e.g., blueprints or assembly instructions—has as its (main) function to influence actions in specific ways, then we will see tokens of that kind—e.g., individual blueprints and assembly instructions—as being things that one can follow or act in accordance with, even if that token does not actually influence anyone to act. As long as those "anomalous" tokens have traits that make them count as belonging to such an action-influencing type, they will also count as having practical direction. (They are, as it were, free riders of the type to which they belong.) I suggest that this is what unites the things that we think of as having practical direction.

If we compare the different types of things that have practical direction, we see that they gain their action-influencing function in different ways. Desires (arguably) have it due to their intrinsic nature (at least if we accept a common form of functionalism about desires). Laws, blueprints, assembly instructions and (verbal) orders have that role due to social conventions. Shopping lists have it due to people's decisions: when I create a list with the purpose of using it to guide my shopping, I thereby give it the function and also make it a shopping list.

The latter examples make it clear that things that do not have practical direction thanks to their intrinsic nature can still become things with practical direction if they gain an action-influencing function in some human practice. A piece of paper with words referring to groceries does not, as such, have practical direction. It could be an enumeration of the first words that come to the author's mind or a list of things in her fridge. If so, there is no action that is to act in accordance with it. It is only if it is or starts to be used with the function of guiding behavior that it becomes a shopping list and becomes a thing that we think of as something that one can act in accordance with. That is, *qua* list, it does not have practical direction, but *qua* shopping list, it does.

Further, for at least many of the phenomena that we think of as having practical direction, this feature is also integrated in the concept in question as a necessary condition: e.g., in order for *x* to count as a shopping list, there have to be ways of acting that are to act in accordance with *x*; and in order for *x* to be a rule in chess, there has to be some way to play that is in accordance with it. Consequently, these phenomena have practical direction by conceptual necessity.

We also have an explanation of why ordinary descriptive beliefs do not have practical direction. When I believe that it is sunny today, the function of this belief, taken by itself, is not to influence or guide me to act in any *specific* way. We might think that one of its functions is to influence behavior, so that the

belief, when combined with, e.g., a desire to be out in the sun, or to stay away from the sun, or to dance a little dance when it is sunny, disposes the agent to act. But the very fact that its function (if we think it has such a function) is to dispose for different actions when combined with different desires, shows that there is not *one specific way* of acting that can be thought of as acting in accordance with the belief.

### *3.3. The Practical Direction of Moral Judgments*

Whether or not we accept that deontic moral judgments are intrinsically or necessarily motivating, we should accept that they have a practical role in our social lives: as central parts of our moral practice they function to influence actions in specific ways. Specific subtypes of deontic moral judgments are tied to specific ways of acting. Judging an action morally wrong generally, though perhaps contingently, functions to dispose the judge to *avoid* the action. This is often also why we tell people, and sometimes argue with them, that certain types of acts are wrong—not just to ensure that they have correct views on this matter, but to influence them to avoid these types of actions. In these ways, moral-wrongness judgments have a specific practical function in our social practice in the following sense: they (typically) play a certain causal role in how people interact with each other and in how they regulate their actions; this role is such that it ties the judgments in question to one specific way of acting (avoidance). Similarly with other subtypes of deontic moral judgments. This action-influencing function is undoubtedly a salient feature of deontic moral judgments. Indeed, any theory of moral judgments that is not consistent with, or that cannot account for, moral judgments having at least a contingent action-influencing function of this sort will be implausible to start with.

With the account of practical direction defended in the previous section, this explains why we think of deontic moral judgments as having practical direction. Since judgments that $\phi$-ing is morally wrong have the function to influence actions in a specific way—namely so that $\phi$-ing is avoided—it makes sense that we think of avoidance of $\phi$-ing as being to act in accordance with, or to comply with, such judgments. It makes sense, in the same way as it makes sense that we think of blueprints and chess rules as being things that one can act in accordance with.

This explanation of why deontic moral judgments have practical direction does not require that they are intrinsically or necessarily connected to motivation to act. Above we noted that most things that we think of as things that one can act in accordance with are not intrinsically motivational—e.g., blueprints, shopping lists, chess rules, and juridical laws. They influence actions because they are part of a practice where people are contingently interested in

following them. Many cognitivists hold similar views about moral judgments. Also, as noted above, there may well be token blueprints and juridical laws, for example, that do not motivate anyone, and that never serve to influence or guide action. As long as they have traits that make them belong to the relevant type of things (e.g., blueprints and juridical laws) which typically have as their function to guide behavior, they will be experienced as things that one can act in accordance with. Similarly, this account allows that moral judgments entirely unaccompanied by motivation still count as having practical direction.

We now have an explanation of moral judgments being mental attitudes that one can act in accordance with that does not require that they are desires, or are necessarily accompanied by desires. For example, the explanation is available to cognitivist theories according to which moral judgments have an action-influencing function due to a strong but contingent relation to motivation. To be clear, I do not here propose some specific view about the mechanisms behind the action-influencing function of moral judgments. The point is instead that the general idea about why moral judgments have practical direction is compatible with different views on this. Cognitivists will hold that deontic moral beliefs are a subclass of our beliefs connected in some special (but contingent) way to motivation. They may say that our motivation to avoid immoral acts is partly due to the special content (or character) of moral-wrongness judgments, innate psychological mechanisms, and/or moral upbringing, perhaps in combination with a complex pattern of social reactive attitudes toward people who act in ways perceived as immoral and a drive to avoid being subject to such attitudes.[21]

There is no reason to think that the kind of theories in focus in this paper—those that reject both uniformity claims—would be unable to use this account of practical direction. Content relativists hold that the wrongness beliefs of different individuals (or in different societies) have different contents, so on a general level what they will need to hold is that beliefs with different content serve the relevant action-influencing role for different individuals (or in different societies). More concretely, one common group of content-relativist theories is especially congenial with the idea that moral judgments have an action-influencing function. These are views according to which the content of moral beliefs (at least in normal cases) depends on the judge's moral standard, where her moral standard is taken to consist in some subclass of

21 This should also make it clear that the claim that moral judgments have an action-influencing function should not be construed in such a way that functionalism about mental states implies that they are, e.g., desires. Rather, the action-influencing role that makes us see, e.g., wrong judgments as connected to one way of acting (avoiding the action judged wrong), can be due to external and contingent motivation. (Thanks to a referee for pushing me to be explicit about this.)

her motivational states (states of the kind that non-cognitivists *identify* moral judgments with), or an idealized version of such a subclass.[22] As we saw in section 1, these relativist theories allow that moral judgments, on the one hand, and desires or motivation, on the other hand, can come apart. Therefore they cannot account for (all) moral disagreements in terms of clashing desires. However, since these theories still imply that there is a tight connection between moral judgment and motivation to act, they have no problem accounting for the action-influencing aspect of moral judgments, and thus for the practical direction of moral judgments—which means that the account of disagreement in terms of practical direction is available to them.

Similarly, other theories that reject both uniformity claims are consistent with moral judgments having an action-influencing function. Variantists, who hold that some moral judgments are beliefs and some are desires, can hold that the action-influencing role is served sometimes by beliefs and sometimes by desires (or alternatively: that most but not all moral judgments motivate because most moral judgments are desires, only some are beliefs).[23] Non-cognitivist theories that identify moral judgments with a cluster of dispositional tendencies that normally come together—of which the disposition to be motivated or have a desire is one—but that allow that there can be moral judgments without all dispositional tendencies being in place, can explain the practical direction of moral judgments in the suggested manner, since they hold that, in normal cases, moral judgments influence action in one specific direction.[24]

To sum up (and once again focusing on cognitivist content relativism as our example): one can hold that moral beliefs have practical direction even though beliefs in general do not. Like other beliefs they are not intrinsically or necessarily motivating. But, unlike other beliefs, they function to influence action in a specific direction. Or differently put, the role they have in our psychologies and moral practices is (for reasons that cognitivists may disagree about) such that they are tied to specific ways of acting. In this respect, they are much like shopping lists or assembly instructions: pieces of paper with words or drawings do not generally have practical direction, but some such pieces of papers have

22 Such theories include those defended in Brogaard, "Moral Contextualism and Moral Relativism"; Dreier, "Internalism and Speaker Relativism"; Harman, "Moral Relativism"; Kölbel, "Moral Relativism"; Prinz, *The Emotional Construction of Morals*; Wong, *Moral Relativity*.

23 For variantist views, see Gill, "Metaethical Variability, Incoherence, and Error" and "Indeterminacy and Variability in Metaethics"; Loeb, "Moral Incoherentism."

24 For an expressivist view of this kind, see Björnsson and McPherson, "Moral Attitudes for Expressivists and Relativists."

been given an action-influencing role (through individual decisions or social conventions) uniquely tying them to certain specific ways of acting.

Furthermore, just like with the concept of a shopping list or blueprint, the practical direction has become a conceptually necessary feature for counting as a deontic moral judgment: by conceptual necessity, a judgment that $\phi$-ing is *O* is not a judgment to the effect that $\phi$-ing is morally obligatory unless $\phi$-ing is to act in accordance with the judgment.

### *3.4. Summing Up*

If the idea about practical direction just defended is correct, then theories about moral judgments that reject both uniformity claims about the class of moral judgments can nonetheless maintain that deontic moral judgments necessarily have practical direction. And if the PD account of moral disagreements presented in section 2 is correct, then we can account for deontic moral disagreements in terms of practical direction. Taken together, this would mean that the PD account allows also theories that reject both uniformity claims to explain moral disagreements. Before we draw that conclusion, however, we need to see how the PD account can handle certain types of moral disagreements not yet discussed.

## 4. THE PRACTICAL DIRECTION ACCOUNT AND COMPLEX MORAL DISAGREEMENTS

Recall what the PD account says:

> *Practical Direction Account*: There is a deontic moral disagreement between *A* and *B* if and only if they accept moral judgments respectively ($J_A$ and $J_B$) such that there is at least some act that is in accordance with $J_A$ and in discordance with $J_B$.

In section 2, I described how this account handles some simple types of moral disagreement—disagreements that can easily be construed as straightforward disagreements about what to do. I will now turn to more complex types of deontic moral disagreements, not as obviously explainable by the PD account. In sections 4.1–4.6, I will argue that the PD account can handle these disagreements as well—the trick is to find the (sometimes fairly complex) act that is in accordance with one and in discordance with the other judgment. In section 4.7, I outline how the PD account can (perhaps) be extended from deontic to evaluative moral disagreements.

*4.1. Conditional Moral Judgments*

$J_9$: $\phi$-ing is morally permitted if and only if $\phi$-ing maximizes happiness.

$J_{10}$: $\phi$-ing is morally permitted whether or not $\phi$-ing maximizes happiness.

*Problem*: $J_9$ and $J_{10}$ intuitively disagree, but since $\phi$-ing is neither in accordance nor discordance with $J_9$, it would seem that the PD account does not capture that.

*Solution*: Intuitively, this is a disagreement about whether it is ok to $\phi$ in a situation where $\phi$-ing does not maximize happiness. So, there is one action type—i.e., *$\phi$-ing in situation where $\phi$-ing does not maximize happiness*—which is in accordance with $J_{10}$, but in discordance with $J_9$. Thus, allowing such "situation-based individuations" of acts, lets the PD account handle disagreement between conditional moral judgments.

*4.2. Conjunctive moral judgments*

$J_{11}$: Both *j*-ing and *y*-ing is morally permitted

$J_{12}$: Doing both *j* and *y* is not morally permitted

*Problem*: Since $J_{12}$ does not state that doing *only one* of *j* or *y* is impermissible, neither of these individual acts is in discordance with $J_{12}$. So how can the PD account capture the disagreement?

*Solution*: There is one kind of act that is in discordance with $J_{12}$ and in accordance with $J_{11}$—the combinatory act of both *j*-ing and *y*-ing. Allowing "combinatory acts" thus lets the PD account handle such disagreements.

*4.3. Disjunctive and Quantified Moral Judgments*

$J_{13}$: $\phi$-ing is morally permitted or *y*-ing is morally permitted.

$J_{14}$: Neither $\phi$-ing nor *y*-ing is morally permitted.

*Problem*: Intuitively, this is a disagreement (about whether at least one of $\phi$-ing and *y*-ing is permitted). But since $J_{13}$ is disjunctive, and therefore does not imply that any of the two acts mentioned is permitted, neither of these acts is in accordance with $J_{13}$. So how can the PD account explain this kind of disagreement?

*Solution*: The disagreement at hand can be construed as a disagreement about whether it is permissible to perform one of the two acts ($\phi$ or *y*) *given that the other act is not permissible*. For example, $J_{13}$ implies that, in a situation where *y*-ing is not morally permitted (i.e., a situation where the second disjunct does not hold), *j*-ing is permitted. So the situation-individuated act *j-ing in a*

*situation where y-ing is not morally permitted* is in accordance with $J_{13}$. Further, this act is in discordance with $J_{14}$, since $J_{14}$ implies that *j*-ing is not permitted.

More generally it holds that disjunctive moral judgments have, for each moral disjunct, implications for what is allowed or not allowed *if the situation is such that none of the other disjuncts hold*. This, then, is the clue to finding the relevant acts in relation to moral disagreements involving disjunctive judgments, that is, acts that are in accordance with one judgment and discordance with the other.

Furthermore, we can make sense of moral disagreements involving quantifiers—e.g., disagreement about whether *at least one* act, *four* acts, *most* acts, *a few* acts, etc., are wrong—in the same way. This is because such quantified statements, for our purposes here, can be seen as equivalent to disjunctions.[25] For example, the judgment that there is at least one act that is permissible in the domain consisting of $\phi$ and *y* is equivalent to $J_{13}$ above.

Consider also the disagreement between one who thinks that at least one action (of all possible actions) is morally wrong and another who rejects this. Then for each action, *x*, there is a disagreement about whether it is wrong to *perform x given that no other acts are wrong*. The PD account captures this since this (italicized) act is in accordance with one and in discordance with the other judgment. Another example:

$J_{15}$: Most of $A_1$–$A_3$ are morally wrong.

$J_{16}$: It is not the case that most of $A_1$–$A_3$ are morally wrong.

$J_{15}$ is equivalent to a disjunctive judgment: [15i] both and only $A_1$ and $A_2$ are wrong, or [15ii] both and only $A_1$ and $A_3$ are wrong, or [15iii] both and only $A_2$ and $A_3$ are wrong, or [15iv] all of $A_1$–$A_3$ are wrong. $J_{16}$ is equivalent to: [16i] both and only $A_1$ and $A_2$ are not wrong, or [16ii] both and only $A_1$ and $A_3$ are not wrong, or [16iii] both and only $A_2$ and $A_3$ are not wrong, or [16iv] all of $A_1$–$A_3$ are not wrong.

We can now find acts (one for each disjunct) that are in accordance with $J_{16}$. For example: $J_{16}$ implies that doing both $A_1$ and $A_2$ is permissible (not wrong) (i.e., that 16-i holds), in a situation where disjunctions (16ii–iv) do not hold. So, *doing both $A_1$ and $A_2$ in a situation where 16ii–iv do not hold* is in accordance with $J_{16}$. And doing both $A_1$ and $A_2$ is in discordance with $J_{15}$, since $J_{15}$ implies that at least one of these acts is wrong. This means that the PD account implies that there is a disagreement since there is an act that is in accordance with one judgment and in discordance with the other.

Other disagreements involving quantifiers can be handled similar ways.

25 Allowing for infinite disjunctions, when the domain existentially quantified over is infinite, should not cause problems in this context.

*4.4. Inter-normative Discourse*

MW: $\phi$-ing is morally wrong.

PW: $\phi$-ing is not prudentially wrong.

AW: $\phi$-ing is not aesthetically wrong.

*Problem*: Intuitively, there is no disagreement between MW, on the one hand, and either AW or PW on the other. But it might seem that the PD account gives us disagreement (and thus overgeneralizes). For $\phi$-ing is to act in accordance with PW and AW, but in discordance with MW.

*Solution*: If there are different specific kinds of normative wrongs (or reasons), then none of these are all-things-considered wrongs (reasons). If so, a judgment to the effect that an act is morally or prudentially wrong does not tell us not to do it. Consequently, doing the act is not in discordance with the judgment, and consequently there is no disagreement between MW and PW/AW.

But do we then get the unwanted implication that no such judgment can be involved in PD disagreements at all? No. Similar to conditional and disjunctive moral judgments, we can find other (situation-individuated) acts that are in accordance/discordance with such judgments. MW in effect tells us that $\phi$-ing is not to be done, if the situation is one where moral reasons are the only reasons relevant to the all-things-considered status of the act—either because they are the only or the strongest reasons. This means that *$\phi$-ing in a situation where only moral reasons are relevant* (in the above sense) is to act in discordance with MW. Likewise, *$\phi$-ing in a situation where only prudential reasons are relevant* is to act in accordance with PW. So a "simple moral disagreement case" like that between $J_1$ ($\phi$-ing is morally wrong) and $J_2$ ($\phi$-ing is not morally wrong) is not a disagreement about what to do *simpliciter*. Rather it is a disagreement about whether to *$\phi$ in a situation where only moral reasons are relevant*. (The reason that there is no disagreement between MW and PW/AW is that there is no situation-based individuation of an act that is in discordance with MW but in accordance with PW/AW.)

We should note, however, that some moral (and prudential, etc.) judgments are all-things-considered judgments. One may judge, e.g., that *$\phi$-ing is all-things-considered wrong for moral reasons.* Indeed, this might be what people often have in mind when they think that $\phi$-ing is morally wrong. If this is how we construe $J_1$ and $J_2$, then this is a simple disagreement about whether to $\phi$. Also, for such all-things-considered judgments, there can be disagreements between different normative domains: if someone judges that *$\phi$-ing is all-things-considered not wrong, for prudential reasons*, then there is a disagreement about whether to $\phi$ between this judgment and the previous moral judgment.

*4.5.* Pro Tanto *Reasons*

$J_{17}$: There is a *pro tanto* moral reason to $\phi$.

$J_{18}$: There is no *pro tanto* moral reason to $\phi$.

*Problem*: Intuitively there is a disagreement here, but the PD account seemingly does not account for it. For there is no action that is in accordance or discordance with $J_{17}$ (or $J_{18}$). This is what to expect, since this is not a simple disagreement about what to do—two people who accept $J_{17}$ and $J_{18}$ may well agree about what is to be done (all things considered).

*Solution*: Even though it is not a disagreement about what to do *all things considered*, it can be construed as a conflict about what to do in a situation where there are no reasons against $\phi$-ing to consider. The act of *not $\phi$-ing in such a situation* goes against $J_{17}$, but is in accordance with $J_{18}$.

*4.6. Summing Up*

The driving thought behind the PD account is the plausible idea that all deontic moral judgments are, in more or less direct ways, directed toward actions—they are judgments about how to act. The most straightforward deontic moral judgments are thus judgments that one can act in accordance or discordance with, and the PD account uses that feature of them to account for disagreement: there is a disagreement between such simple deontic judgments if there are acts that are in accordance with one judgment but in discordance with the other. The disagreement consists, as it were, in a clash in practical direction between the two judgments. We have now considered deontic moral judgments that have a more complex structure—conditional, conjunctive, disjunctive, and quantified judgments, judgments about *pro tanto* reasons, and about different kinds of normative reasons. I have argued that also for these kinds of judgments, it is possible to find (types of) actions that are in accordance/discordance with the judgments, and that the PD account therefore accounts for disagreements between such judgments.

*4.7. Evaluative Disagreements*

I have argued that the PD account can handle *deontic* moral disagreements. This is the main purpose of this paper. In this subsection I briefly outline how the PD account could also be developed to handle *evaluative* moral disagreements. Consider the following judgments:

$J_{19}$: $x$ is good.

$J_{20}$: $x$ is not good.

*Problem*: Evaluative judgments are not (in a simple way) judgments that can be acted in accordance with. So the PD account cannot handle them.

*Solution*: It is quite plausible that that one's negative and positive attitudes, or preferences, can be in accordance/discordance with one's (and others') evaluative judgments. This idea is congenial with views according to which having value is to be *desirable* (or to be a thing that it is *fitting to desire*). Reasonably, if I judge that $x$ is desirable, then desiring $x$ is to desire in accordance with that judgment. (Other attitudes than desires might be relevant depending on what kind of evaluative judgments we are concerned with: about intrinsic moral values, character evaluations, aesthetic evaluations, prudential evaluations, etc.). According to the present rough suggestion, then, there is a disagreement between $J_{19}$ and $J_{20}$ because:

> Liking/desiring/appreciating/admiring $x$ (or preferring $x$ to something neutral) is in accordance with $J_{19}$ but in discordance with $J_{20}$.

Even though I find it plausible that some account of evaluative disagreements along these lines will work, the exact relation between values and positive/negative attitudes is a complicated and disputed issue, and the details needs to be worked out elsewhere.[26]

## 5. CONCLUDING REMARKS

The main aim of this paper has been to develop a plausible account of moral disagreements, the PD account, which requires neither belief uniformity nor attitude uniformity. The PD account is a non-doxastic account of moral disagreements, similar to accounts in terms of clashing desires, but does not require that all moral judgments are, or are accompanied by, desires. This is possible because (i) moral judgments necessarily have practical direction, in

26 One further issue, both regarding deontic and evaluative disagreements, concerns disagreements that involve epistemic modals:

$J_{21}$: $\phi$-ing might be morally wrong.

$J_{22}$: $\phi$-ing is (definitely) not morally wrong.

There seems to be a disagreement here, but not about whether $\phi$-ing is wrong (although there is a potential disagreement about that), rather about whether it is certain that $\phi$-ing is not wrong. This means that there is no disagreement about what to do here and, consequently, the disagreement is not captured by the original PD account.

One possible solution is to say that, here, $\phi$-ing is in accordance with a moral claim endorsed through the acceptance of $J_{22}$ but in discordance with a moral claim *actively held open through the acceptance of* $J_{21}$. I tentatively suggest that this is what the disagreement consists in. To handle these sorts of cases, the statement of the PD account would need to be amended accordingly.

the sense that they are judgments that one can act in accordance with, even if they are beliefs that are only contingently accompanied by desires, and (ii) moral disagreement can be construed as clashes in practical direction. If this is correct, theories about moral judgments that reject both uniformity claims—most prominently various versions of content relativism—cannot be dismissed due to an inability to account for moral disagreement.

It might be objected that the PD account fails to explain one aspect of moral disagreement. It accounts for moral disagreements as practical disagreements (about what to do). But we also experience them as disagreements *about whether something is the case (or is true)*: e.g., about whether it is wrong to eat meat. In this way, they have the appearance of conflicts in belief (beliefs in contradictory propositions).

Here I can only outline one possible way for defenders of the PD account to handle this appearance. Plausibly, the account should be understood as claiming that, in normative domains, what on the surface has the same appearance as disagreements in belief (at least partly: they also appear to be practical disagreements), ultimately, under the surface, is another kind of disagreement. This is similar to what many non-cognitivists have argued: that we can make sense of moral judgments and moral discourse having an (in many respects) absolutist cognitivist surface appearance, while they have a different underlying nature. If such a strategy works, then the surface appearance of moral disagreement need not be problematic.

Indeed, on pain of being complete nonstarters, content relativists (and others rejecting belief uniformity) need some strategy like this, *irrespective of how they explain disagreement*. This means that the appearance of moral disagreements is not an *extra* explanatory burden for these views. In short: consider the class of all judgments *that killing is morally wrong*. We represent this class *as if* they were judgments with the same propositional content: that is, we represent them (in language and thought) with the same that-clause. Content relativists (and others rejecting belief uniformity) hold that this surface—this way of representing them—is misleading: the judgments in the class need not share content. (Rather they share something else that makes them judgments *that killing is wrong*.) If content relativists can make sense of this much, which they must do to get off the ground, then they have also, it seems, made sense of the fact that we think and talk of moral disagreements *as if* they are conflicts between beliefs with contradictory propositional content. For the beliefs between which there are disagreements (PD disagreements, if I am right), are judgments that we represent as if they are beliefs with conflicting contents: e.g., a judgment *that killing is wrong* and a judgment *that killing is not wrong*. Given this, it makes sense that we think of it as a disagreement *about whether killing is*

*wrong* even though the conflict that is actually there is a clash in practical direction.[27] To summarize: the PD account lets content relativists (and others who reject both uniformity claims) acknowledge that people morally disagree in cases where we intuitively think they disagree; then it remains to be explained why moral judgments and moral disagreements have an absolutist surface appearance, but that is a challenge that content relativists need to tackle anyway.

Finally, let me propose that, even if the dialectical significance of the PD account partly comes from the fact that it requires neither belief uniformity nor attitude uniformity, it might be considered part of the complete picture of moral disagreement even if one of the uniformity claims holds. First, if attitude uniformity holds, e.g., because some variant of non-cognitivism is correct, then moral judgments have practical direction (are such that they can be acted in accordance with) by virtue of being practical desire-like states. Consequently, the PD account is applicable. We may then perhaps see the attitude account of moral disagreement as an instance of the more general PD account (where the latter explains in which way the practical attitudes in question clash).[28] (The question of how this relates to other ways of understanding disagreements in attitude will have to be discussed elsewhere.) Second, even if absolutist cognitivism (and thereby belief uniformity) is correct, and moral disagreement can be accounted for in terms of beliefs in contradictory propositions, we should acknowledge that when two people disagree about whether something is morally disallowed, allowed, or obligatory, they *also* disagree in the sense that at least some way of acting is in accordance with one person's judgment but in discordance with the other person's judgment.[29]

*University of Gothenburg*
*ragnar.francen@filosofi.gu.se*

27 Furthermore, if the PD account works as I have argued above, then it is extensionally equivalent to an absolutist cognitivist belief account. That is, the two accounts imply that we disagree in exactly same cases (namely, the cases that intuitively count as moral disagreements). This means that descriptions of moral disagreements as if they were conflicts between beliefs in contradictory propositions can be used, without extensional mismatches, to represent conflicts consisting of clashes in practical direction.

28 Thanks to an anonymous referee for suggesting this.

29 I am grateful for comments from two anonymous referees for this journal, and for discussion with Gunnar Björnsson, John Eriksson, and Alva Ståge. I am also grateful for valuable comments from the audiences when earlier versions of this paper were presented at the Gothenburg research seminar in practical philosophy, at the Language and Metaphysics of Normativity conference, Uppsala (2016), the Normative Disagreement Workshop, Oslo (2016), the Value Disagreement Conference, Lisbon (2017), the Future of Normativity Conference, University of Kent (2018), and the Second Groningen Metaethics Workshop,

## REFERENCES

Beddor, Bob. "Subjective Disagreement." *Nous* 53, no. 4 (December 2019): 819–51.

Bedke, M. S. "Moral Judgment Purposivism: Saving Internalism from Amoralism." *Philosophical Studies* 144, no. 2 (May 2009): 189–209.

Björnsson, Gunnar. "How Emotivism Survives Immoralists, Irrationality, and Depression." *Southern Journal of Philosophy* 40, no. 3 (March 2002): 327–44.

Björnsson, Gunnar, and Stephen Finlay. "Metaethical Contextualism Defended." *Ethics* 121, no. 1 (October 2010): 7–36.

Björnsson, Gunnar, and Tristram McPherson. "Moral Attitudes for Expressivists and Relativists: Solving the Specification Problem." *Mind* 123, no. 489 (January 2014): 1–38.

Blackburn, Simon. *Spreading the Word: Groundings in the Philosophy of Language.* Oxford: Clarendon Press, 1984.

Bolinger, Renée Jorgensen. "Metalinguistic Negotiations in Moral Disagreement." *Inquiry* 65, no. 3 (2020): 352–80.

Brink, David O. *Moral Realism and the Foundations of Ethics.* Cambridge: Cambridge Univerity Press, 1989.

Brogaard, Berit. "Moral Contextualism and Moral Relativism." *Philosophical Quarterly* 58, no. 232 (July 2008): 385–409.

———. "Moral Relativism and Moral Expressivism." *Southern Journal of Philosophy* 50, no. 4 (December 2012): 538–56.

Dreier, James. "Internalism and Speaker Relativism." *Ethics* 101, no. 1 (October 1990): 6–26.

———. "Relativism (and Expressivism) and the Problem of Disagreement." *Philosophical Perspectives* 23 (2009): 79–110.

———. "The Supervenience Argument against Moral Realism." *Southern Journal of Philosophy* 30, no. 3 (Fall 1992): 13–38.

———. "Transforming Expressivism." *Noûs* 33, no. 4 (December 1999): 558–72.

Egan, Andy. "Relativist Dispositional Theories of Value." *Southern Journal of Philosophy* 50, no. 4 (December 2012): 557–82.

Eriksson, John. "Explaining Disagreement: Contextualism, Expressivism and Disagreement in Attitude." *Belgrade Philosophical Annual* 32 (2019): 93–113.

———. *Moved by Morality: An Essay on the Practicality of Moral Thought and Talk.* Uppsala, Sweden: Filosofiska Institutionen Uppsala Universitet, 2006.

Finlay, Stephen. *Confusion of Tongues: A Theory of Normative Language.* Oxford:

University of Groningen (2019). The research was supported by the Riksbankens Jubileumsfond Grant DNR: P16-0061:1.

Oxford University Press, 2014.
———. "Disagreement Lost and Found." In *Oxford Studies in Metaethics*, vol. 12, edited by Russ Shafer-Landau, 187–205. Oxford: Oxford University Press, 2017.
Francén, Ragnar. *Metaethical Relativism: Against the Single Analysis Assumption.* Göteborg: Department of Philosophy, University of Gothenburg, 2007.
———. "Moral and Metaethical Pluralism: Unity in Variation." *Southern Journal of Philosophy* 50, no. 4 (December 2012): 583–601.
———. "No Deep Disagreement for New Relativists." *Philosophical Studies* 151, no. 1 (October 2010): 19–37.
Gert, Joshua, and Alfred R. Mele. "Lenman on Externalism and Amoralism: An Interplanetary Exploration." *Philosophia* 32 (May 2005): 275–83.
Gibbard, Allan. *Thinking How to Live.* Cambridge, MA: Harvard University Press, 2003.
Gill, Michael B. "Indeterminacy and Variability in Metaethics." *Philosophical Studies* 145, no. 2 (August 2009): 215–34.
———. "Metaethical Variability, Incoherence, and Error." In *Moral Psychology*. Vol. 2, *The Cognitive Science of Morality: Intuition and Diversity*, edited by Walter Sinnott-Armstrong. Cambridge, MA: MIT Press, 2007.
Hare, R. M. *The Language of Morals.* Oxford: Clarendon Press, 1952.
Harman, Gilbert. "Moral Relativism." In *Moral Relativism and Moral Objectivity*, edited by Gilbert Harman and Judith Jarvis Thomson, 1–64. Cambridge: Blackwell, 1996.
———. "Moral Relativism Defended." *Philosophical Review* 84, no. 1 (January 1975): 3–22.
———. "What Is Moral Relativism?" In *Values and Morals*, edited by A. I. Goldman and J. Kim, 143–61. Dordrecht: D. Reidel Publishing Company, 1978.
Horgan, Terence, and Mark Timmons. "New Wave Moral Realism Meets Moral Twin Earth." *Journal of Philosophical Research* 16 (1991): 447–65.
Jackson, Frank. "The Argument from the Persistence of Moral Disagreement." In *Oxford Studies in Metaethics*, vol. 3, edited by Russ Shafer-Landau, 75–86. Oxford: Oxford University Press, 2008.
Korsgaard, Christine M. "Skepticism about Practical Reason." *Journal of Philosophy* 83, no. 1 (January 1986): 5–25.
Kölbel, Max. "Indexical Relativism versus Genuine Relativism." *International Journal of Philosophical Studies* 12, no. 3 (2004): 297–313.
———. "Moral Relativism." In *Lectures on Relativism*, edited by Dag Westerståhl and Torbjörn Tännsjö, 51–72. Göteborg: Philosophical Communications, 2005.
Loeb, Don. "Moral Incoherentism: How to Pull a Metaphysical Rabbit Out of a

Semantic Hat." In *Moral Psychology*. Vol. 2, *The Cognitive Science of Morality: Intuition and Diversity*, edited by Walter Sinnott-Armstrong. Cambridge, MA: MIT Press, 2008.

López de Sa, D. "Expressing Disagreement: A Presuppositional Indexical Contextualist Relativist Account." *Erkenntnis* 80 (March 2015): 153–65.

———. "Presuppositions of Commonality: An Indexical Relativist Account of Disagreement." In *Relative Truth*, edited by Max Kölbel and M. García-Carpintero, 297–310. Oxford: Oxford University Press, 2008.

MacFarlane, John. "Relativism and Disagreement." *Philosophical Studies* 132 (January 2007): 17–31.

Mackie, J. L. *Ethics: Inventing Right and Wrong*. London: Penguin Books, 1977.

Marques, Teresa. "Disagreeing in Context." *Frontiers in Psychology* 6, no. 257 (March 2015).

———. "Doxastic Disagreement." *Erkenntnis* 79 (March 2014): 121–42.

———. "What Metalinguistic Negotiations Can't Do." *Phenomenology and Mind* 12 (August 2017): 40–48.

Marques, Teresa, and Manuel García-Carpintero. "Disagreement about Taste: Commonality Presuppositions and Coordination." *Australasian Journal of Philosophy* 92, no. 4 (2014): 701–23.

Mele, Alfred R. "Internalist Moral Cognitivism and Listlessness." *Ethics* 106, no. 4 (July 1996): 727–53.

Phillips, David. "How to Be a Moral Relativist." *Southern Journal of Philosophy* 35, no. 3 (Fall 1997): 393–417.

Plunkett, David, and Timothy Sundell. "Antipositivist Arguments from Legal Thought and Talk: The Metalinguistic Response." In *Pragmatism, Law, and Language*, edited by Graham Hubbs and Douglas Lind, 56–75. New York: Routledge, 2014.

———. "Disagreement and the Semantics of Normative and Evaluative Terms." *Philosophers' Imprint* 13, no. 23 (December 2013): 1–37.

———. "Dworkin's Interpretivism and the Pragmatics of Legal Disputes." *Legal Theory* 19, no. 3 (September 2013): 242–81.

Prinz, Jesse. *The Emotional Construction of Morals*. Oxford: Oxford University Press, 2007.

Roskies, Adina. "Are Ethical Judgments Intrinsically Motivational? Lessons from 'Acquired Sociopathy.'" *Philosophical Psychology* 16, no. 1 (2003): 51–66.

Shafer, Karl. "Assessor Relativism and the Problem of Moral Disagreement." *Southern Journal of Philosophy* 50, no. 4 (December 2012): 602–20.

———. "Constructivism and Three Forms of Perspective-Dependence in Metaethics." *Philosophy and Phenomenological Research* 89, no. 1 (July 2014): 68–101.

Smith, Michael. *The Moral Problem*. Oxford: Blackwell, 1994.
Stevenson, Charles L. "The Emotive Meaning of Ethical Terms." In *Facts and Values: Studies in Ethical Analysis*. New Haven: Yale University Press, 1963.
Stocker, Michael. "Desiring the Bad: An Essay in Moral Psychology." *Journal of Philosophy* 76, no. 12 (December 1979): 738–53.
Sundell, Timothy. "Disagreements about Taste." *Philosophical Studies* 155, no. 2 (September 2011): 267–88.
Svavarsdóttir, Sigrun. "Moral Cognitivism and Motivation." *Philosophical Review* 108, no. 2 (April 1999): 161–219.
Tersman, Folke. *Moral Disagreement*. Cambridge: Cambridge University Press, 2006.
Tresan, Jon. "The Challenge of Communal Internalism." *Journal of Value Inquiry* 43, no. 2 (June 2009): 179–99.
Van Roojen, Mark. "Moral Rationalism and Rational Amoralism." *Ethics* 120 (April 2010): 495–525.
Wedgwood, Ralph. *The Nature of Normativity*. Oxford: Clarendon Press, 2007.
Wong, David B. *Moral Relativity*. Berkeley: University of California Press, 1984.

*Journal of Ethics and Social Philosophy*
VOL. 23, NO. 2 · NOVEMBER 2022
*https://doi.org/10.26556/jesp.v23i2.2113*


# THE BEST AVAILABLE PARENT AND DUTIES OF JUSTICE

*Jordan David Thomas Walters*

In a recent paper published in *Ethics*, Anca Gheaus argued for the best available parent view, which holds that the right to parent should track "the child's, as well as third parties', but not the potential parents', interests."[1] In this short note, I argue that the best available parent view, in its present formulation, struggles to accommodate for our weighty duty not to perpetuate historical injustices. I offer an alternative view that reconciles this tension. Let us begin with Gheaus's view.

The status quo is, as Gheaus puts it, that we should "allocate child custody to procreators [because they] hold the moral right to parent their offspring, unless they renounce it or lose it for child abuse and neglect."[2] The status quo constitutes not only the moral and philosophical leanings of many, but also the legal order in which we were raised. For although some of us might have had better parents than others in some respects, the state took no interest in allocating children to the best available parent.

Gheaus thinks that the status quo fails to properly justify parental authority over children. We might reconstruct Gheaus's argument for the best available parent view as follows:

1. Rights to control the life of another must be justified in two ways: consent or legitimate interests.
2. Children have not yet developed into fully autonomous agents and therefore cannot give consent.
3. By 2, parental rights must be justified by appeal to the child's legitimate interests.

1 Gheaus, "The Best Available Parent." Following Gheaus, when I refer to "rights" in this essay, I am referring to moral rights unless otherwise specified.

2 Gheaus, "The Best Available Parent," 434.

4. Since "childrearing can have negative externalities … there is a presumption in favour of child-rearing that advances children's interests as much as possible, while respecting third-parties' rights."[3]
5. By 4, "there is a presumption in favour of the view that the right to parent is held by the person who would make the best available parent for a child and who is willing to rear her."[4]

Gheaus points out that the "best" in the best available parent view should be understood comparatively. That is, if we are given a choice between parents $P_1 \ldots, P_n$, and we know that $P_{101}$ is the best, we ought to specify the right to parent to $P_{101}$, since that would make things go best. On the comparative construal of "best," even if $P_{101}$ is only 1 percent better than $P_{100}$, we would be making a moral mistake if we specified the right to $P_{100}$. For our purposes, it will be useful to formulate a version of Gheaus's principle of parental control rights that makes explicit this comparative construal of bestness:

> *Gheaus's Comparative Principle* (GCP): "The right to parent is held by the person who, among those willing to parent, is going to advance the child's legitimate interests best."[5]

But there is something strange about trying to apply GCP, at least in certain contexts. As Gheaus herself acknowledges: "In non-ideal circumstances many people are unjustly poor and suffer from social exclusion," which gives rise to the worry that "the best available parent view compounds injustice by denying them a right to parent."[6] I think this is a worry that we should take seriously. When we are thinking about parental rights, we should not ignore such circumstances. We should recognize that all children enter the world with a particular history, bound by a particular set of institutions, social practices, and familial relations. And we ought to take this into account when theorizing about how to specify parental rights. Reflecting on how the best available parent view applies in such circumstances, Gheaus writes:

> Another person's claim to parent a newborn cannot easily over-ride the claim of the loving and adequate gestational mother to exercise global authority over the child. *But the view allows for this possibility.* Suppose another adult wishes to parent the child, an adult who is not yet in a loving relationship with the newborn but whose abilities to exercise

3 Gheaus, "The Best Available Parent," 435.
4 Gheaus, "The Best Available Parent," 435.
5 Gheaus, "The Best Available Parent," 434.
6 Gheaus, "The Best Available Parent," 459.

> beneficial authority over the child significantly surpass those of the gestational mother. In this case the best available parent view may mandate the exclusion of the gestational parent from exercising parental authority over the child, but not from continuing an intimate relationship with her.[7]

I want to focus on Gheaus's claim that the best available parent view allows for the possibility that we might specify parental rights in such a way that a gestational mother could lose their parental rights on the grounds that there exists another person who would be a better parent than the gestational mother. This seems mistaken to me.

To see why, consider the case of the residential school system in Canada. Between 1831 and 1996, more than 130 residential schools operated within Canada. The stated purpose of the residential school system was to assimilate Indigenous youth into Canadian society. To assimilate, teachers prohibited students from speaking their own language, wearing traditional clothes, and practicing Indigenous spiritual traditions. An estimated 3,200 Indigenous children died from overcrowding in the schools and many students suffered physical and sexual abuse.[8]

Despite their recent decline in the 1990s, the effects of the residential school system are still unravelling. While I was writing this article, an unmarked mass grave of 215 Indigenous children was discovered in British Columbia.[9] Seven hundred fifty-one unmarked mass graves have been discovered near a former residential school in Saskatchewan.[10] And 182 unmarked graves have been discovered in British Columbia.[11] It is perhaps an understatement to note that recent empirical research supports the claim that survivors of the residential school system face vast health inequalities in Canadian society.[12] The historical trauma and lasting effects of colonization that survivors of the residential school system deal with may have led to a situation where survivors may lack "personal parenting resources" through no fault of their own.[13]

7 Gheaus, "The Best Available Parent," 458, emphasis added.

8 Miller, "Residential Schools in Canada."

9 Watson and Dickson, "Remains of 215 Children Found Buried at Former B.C. Residential School, First Nation Says."

10 Eneas, "Sask. First Nation Announces Discovery of 751 Unmarked Graves Near Former Residential School."

11 Migdal, "182 Unmarked Graves Discovered Near Residential School in B.C.'s Interior, First Nation Says."

12 Wilk, Maltby, and Cooke, "Residential Schools and the Effects on Indigenous Health and Well-Being in Canada."

13 Gheaus defines "personal parenting resources" as dispositions that are efficient for the task of child-rearing, e.g., emotional stability and a tendency to nurture. See Gheaus, "The Best

Given this backdrop of genocide and colonization, it seems strange that "the best available parent view may mandate the exclusion of the gestational parent from exercising parental authority over the child, but not from continuing an intimate relationship with her."[14] For any act of specifying the right to parent an Indigenous child to the *best* available parent would seem to come at the expense of compounding this historical injustice, and would likewise serve to reimpose a form of assimilation that initially drove the residential school project. Now, Gheaus does acknowledge that her view might advocate for a sort of "leveling up" of would-be parents who are victims of injustice; so, in some circumstances, the right does not hold since it is silenced by our weightier obligations to reparations. And yet recall that Gheaus does allow for the possibility that the best available parent view might exclude the gestational mother from exercising parental authority over their child (while allowing them to maintain an intimate relationship with the child).[15]

But why should this be a live possibility simply because there exists another would-be parent who would be a better parent than the gestational mother? In the case of would-be parents who are survivors of the residential school system, admitting of this possibility seems to reinstitute a form of colonial interference with Indigenous peoples.[16] That is, it assumes that we can weigh the value of

*X*: having a child raised by the best available parent

against

*Y*: our duty not to perpetuate historical injustices.

Yet talk of *X* outweighing *Y* in our normative theorizing, in this case, seems to yield the wrong verdict. Part of taking our duty not to perpetuate historical

---

Available Parent," 450. I should note that I most certainly do not want to assume that Indigenous individuals would not be the "best parents" for their children in the actual world. My thought here is that, due to discrimination, interference by the Canadian state, etc., it is a live possibility that some Indigenous individuals have adequate personal parenting resources as opposed to optimific personal parenting resources. But this claim is a conditional claim about a possible world, which says: even if such-and-such conditions were to hold in Indigenous communities, it would be wrong to specify the content of moral paternal rights in the way that GCP does. An anonymous referee aptly notes that similar considerations might arise due to individuals affected by structural injustices, which might hinder their personal parenting resources.

14 Gheaus, "The Best Available Parent," 458.

15 Gheaus, "The Best Available Parent," 458.

16 Although note that the interference would not only arise by the recognition of a legal or political right to parent consistent with GCP. My worry here is that conceptualizing the specification of parental rights according to GCP makes this a live possibility.

injustices toward Indigenous persons seriously seems to require that we step back from a view that treats all specifications of parental rights as normatively equal. That is, we seem to be led to modify GCP to give us the following:

> *Gheaus's Comparative Principle** (GCP*): "The right to parent is held by the person who, among those willing to parent, is going to advance the child's legitimate interests best" on the condition that specifying parental rights in such a way does not severely conflict with other very weighty duties of justice (esp. duties not to perpetuate historical injustices).[17]

The issue with GCP was that it assigned lexical priority to *X* over *Y*. But taking our duty not to perpetuate historical injustices toward Indigenous persons seriously requires that we place strong side constraints on what can be done to improve children's lives; that is, we ought to assign lexical priority to *Y* over *X*, which gives us GCP*.[18]

But perhaps there is a reply for Gheaus in the vicinity. Part of what Gheaus has been assuming is that we should treat all specifications of parental rights as normatively equal. That is, when we are thinking about specifying parental rights, we ought to think of the entire set of would-be parents as falling under the scope of GCP. Perhaps this is governed by some feasibility constraints—for example, that all the willing parents *can* actually exercise their parental rights—and this would require that the would-be parents and the child are both inhabitants of the same territory (e.g., Canada). Yet proceeding in this general manner yields an incorrect conclusion, for it seems wrong to specify the parental rights of Indigenous children to non-Indigenous persons who exist outside of the Indigenous community simply because they would be the "best" available parent. However, if GCP is democratically decided upon *within* a given Indigenous community, then it does seem *prima facie* tenable.[19] This is because there is salient a difference between

> *A*: specifying parental rights *between* a community

and

> *B*: specifying parental rights *within* a community.

17 Gheaus, "The Best Available Parent," 434. Note that I specify duties of justice, e.g., the duty not to perpetuate historical injustices, as being "very weighty." One could specify such duties with an infinite weight, yet this would likely run into cases where some evaluative fact(s) outweigh a very weighty fact about side constraints. Thus, I admit that they may be defeasible in some exceptional circumstances.

18 Cf. Rawls, *A Theory of Justice*, 42n23.

19 I would like to thank an anonymous referee for suggesting this point.

Whereas *A* seems to issue the wrong verdict, *B* might not. This is because *A* assumes that we can specify parental rights in such a way that they *can* severely conflict with our duty not to perpetuate historical injustices toward Indigenous persons. Yet *B*, if it were democratically decided upon by a particular Indigenous community, might avoid the worry that the entire set of would-be parents in Canada is included within the scope of GCP.

But this construal of GCP might just end up being another way of restating GCP*. For the thought behind GCP* was that we ought to take *Y* to be lexically prior to *X*, and then build that into GCP. Similarly, if we take *B* to be lexically prior to *A*, that gives us a democratic deliberative constraint, which would satisfy GCP* is most cases. While I think this is right, it is not clear that Gheaus would accept such a concession to her much more demanding GCP. This is because Gheaus seems to want GCP to serve as a universal monistic principle that tells us something about the justification of parental rights *as such*, rather than the justification of parental rights indexed to a particular community.[20] Perhaps such an account can work in some cases, but I hope to have shown that things are much more complicated when we consider applying GCP to Indigenous persons in Canada.

Before wrapping up, I should briefly note what does *not* follow from adopting GCP*. A critic might wonder whether GCP* permits interference with parental rights in cases of child abuse.[21] Fortunately, it does because GCP* only rules out interfering with (or specifying) parental rights to bring about optimific results. GCP* thus hinges on the following *asymmetry intuition*: we seem to be permitted to prevent some bad state of affairs from happening (e.g., preventing a murderer from killing an innocent) at the cost of compounding a historical justice, but we do not seem to be permitted to bring about some good state of affairs (e.g., optimifically specifying parental rights) at the cost of compounding historical injustice.[22] That the asymmetry seems to hold shows that my argument in this paper applies to Gheaus's view in particular and not to other

20 As an anonymous referee notes, Gheaus could have left room in her view for the moderated view I offer here. But Gheaus seems unwilling to promote other considerations of justice at the expense of letting children be parented by suboptimal parents. This is why Gheaus rejects the dual-interest view defended by Brighouse and Swift. See, e.g., Brighouse and Swift, *Family Values*.

21 I am grateful to an anonymous referee for suggesting that I say more on this crucial point.

22 You might also find the asymmetry intuition intuitive, but if you do not, then I suspect that you might disagree with me over a deeper question: Is the right prior to the good or is the good prior to the right? Throughout the paper, I have taken the familiar Rawlsian view that the right is prior to the good, e.g., in saying that we ought to assign lexical priority to *Y* over *X*, which gives us GCP*. But of course, some consequentialists might find this background intuition entirely uncompelling; unfortunately, I cannot address this deeper disagreement in this short note.

moral principles. In other words, I have only demonstrated that one particular optimific moral principle (i.e., GCP) ought to be revised, not that there is a general issue with all optimific moral principles in general.

Leaving that qualifying point aside, I want to make fully explicit what I take to be the root issue of this dialectic: the meaning of "best." It matters greatly what we think "bestness" means, for recall the trouble with GCP was that it specifies "bestness" comparatively, which, together with the duty to bring about optimific results, gives rise to the view that even if $P_{101}$ is only 1 percent better than $P_{100}$, we would be making a moral mistake if we specified the right to $P_{100}$. To see why this is such an odd result, consider two cases.[23]

> *Optimific Parent*: Due to the history of colonial oppression, child *X* would have vastly better outcomes if raised by $P_{1,000,000}$ as opposed to $P_1$.
>
> *Marginally Better Parent*: Due to the history of colonial oppression, child *X* would have marginally better outcomes if raised by $P_{1,000}$ as opposed to $P_{999}$.

In the Optimific Parent case, one might reasonably think that it is preferable for the child to be adopted by $P_{1,000,000}$, even if doing so exacerbates historical injustice. In contrast, in the Marginally Better Parent case, the fact that $P_{1,000}$ having parental rights over *X* would make things go best for *X*, comparatively speaking, does not seem to be a reason to specify parental rights in this way; such a reason seems outweighed by our very weighty duty not to perpetuate historical injustices. But this is precisely the problem with GCP: it issues the same verdict in both cases. Yet insofar as there is a salient difference between the two cases, we seem to be led to modify GCP to GCP*.

In closing, I would like to suggest that our very weighty duty not to perpetuate historical injustices should encourage us to seek different avenues for improving the lives of children. No doubt, GCP* should not be taken to exhaust our normative vocabulary, for there are other weighty duties of justice (i.e., reparations), and these duties might turn out to entail an even more radical view than the best available parent view. Thus, what I have argued for in this paper should be understood as a conciliatory attempt to modify Gheaus's original proposal, to think through its implications within a particular context.[24] For

23 Thanks to an anonymous associate editor for asking me to expand on this point about GCP and for asking me to consider the differences between the following two cases.

24 In a way, the motivation behind GCP* grows out of a more expansive reading of the fourth premise of Gheaus's argument for the best available parent view. When Gheaus says that "childrearing can have negative externalities" ("The Best Available Parent," 435), I think we ought to take "negative externalities" to capture not only rights violations, but also unfulfilled duties of justice. Thanks to an anonymous associate editor for helping me develop this point.

those attracted to Gheaus's view, I hope that GCP* ultimately allows us to recognize that, alongside our moral ideals, we also have political ideals to create a better world for our children, rather than merely better children for our world.[25]

*McGill University*
*jordan.walters@mail.mcgill.ca*

## REFERENCES

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

Eneas, Bryan. "Sask. First Nation Announces Discovery of 751 Unmarked Graves Near Former Residential School." *Canada Broadcasting Corporation*, June 24, 2021. https://www.cbc.ca/news/canada/saskatchewan/cowessess-marieval-indian-residential-school-news-1.6078375.

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

Migdal, Alex. "182 Unmarked Graves Discovered near Residential School in B.C.'s Interior, First Nation Says." *Canada Broadcasting Corporation*, June 30, 2021. https://www.cbc.ca/news/canada/british-columbia/bc-remains-residential-school-interior-1.6085990.

Miller, J. R. "Residential Schools in Canada." In *The Canadian Encyclopedia*. Historica Canada, October 10, 2012. https://www.thecanadianencyclopedia.ca/en/article/residential-schools.

Rawls, John. *A Theory of Justice*. Cambridge, MA: Harvard University Press, 1971.

Watson, Bridgette, and Courtney Dickson. "Remains of 215 Children Found Buried at Former B.C. Residential School, First Nation Says." *Canada Broadcasting Corporation*, May 27, 2021. https://www.cbc.ca/news/canada/british-columbia/tk-emlúps-te-secwépemc-215-children-former-kamloops-indian-residential-school-1.6043778.

Wilk, Piotr, Alana Maltby, and Martin Cooke. "Residential Schools and the Effects on Indigenous Health and Well-Being in Canada: A Scoping Review." *Public Health Reviews* 38, no. 8 (2017).

25 I would like to thank an anonymous associate editor and two anonymous referees at the *Journal of Ethics and Social Philosophy* for their extremely thorough and helpful comments on this paper. For helpful feedback and discussion on earlier drafts, I would also like to thank Chris Howard, Connor Kianpour, Khaleel Rajwani, Matthew Palynchuck, Em Walsh, and Daniel Weinstock. Finally, I would like to acknowledge support from the Social Sciences and Humanities Research Council of Canada.

*Journal of Ethics and Social Philosophy*
VOL. 23, NO. 2 · NOVEMBER 2022
*https://doi.org/10.26556/jesp.v23i2.1971*


# THE STABILITY OF THE JUST SOCIETY

## WHY FIXED POINT THEOREMS ARE BESIDE THE POINT

*Sean Ingham and David Wiens*

Political theorists often investigate the attributes of normatively desirable states of affairs. What are the important features of a just society? What makes a democracy desirable? These and similar questions characteristically preoccupy political theorists. In this vein, the question of stability has attracted recurring interest: Can some desirable state of affairs, if realized, persist over time, or are desirable states of affairs bound to degenerate? Alexander Schaefer argues that this focus on the stability of desirable states of affairs—"static political theory" as he calls it—is deeply misguided. The alleged problem is that the question of stability presupposes the existence of "equilibrium" states of affairs. The claim is roughly this: unless a theorist can show that a social-moral system contains an equilibrium state (in a sense to be defined below), there is no point investigating the relative stability of different states within that system. "Before discussing *stability*, the theorist must discuss *existence*."[1] Having exposed this presupposition, Schaefer presents an argument to challenge it, which we reconstruct below. His stated aims are modest. Rather than establish the general nonexistence of social-moral equilibrium states, he "aims to shift the burden of proof": political theorists cannot simply assume that social-moral systems will contain equilibrium states; they "must *prove this*, or at least provide some reason for us to believe it" (1, emphasis in original; also see 9). Yet Schaefer takes this shift to have far-reaching implications: absent an argument that social-moral systems are likely to contain equilibrium states, political theorists "may need to refocus [their] gaze," turning their attention from normatively desirable states of affairs to the "process[es] by which such states arise and are swept away" (1).

Schaefer's conclusion may be correct—perhaps political theorists should spend less effort examining static states of affairs and more time studying

1 Schaefer, "Is Justice a Fixed Point?" 1, emphasis in original; see also 4, 9. Parenthetical page references hereafter refer to the early online version of Schaefer's article.

dynamic social processes. But whatever the merits of this claim, his argument gives us no reason to accept it. The argument fails for two reasons. First, Schaefer's challenge to the existence of social-moral fixed points threatens not only the existence of equilibrium states of affairs but also of robust dynamic processes as he defines that concept. What goes for static political theory goes for his favored "dynamic political theory" too. Second, Schaefer is mistaken about the burden of proof borne by political theorists who are interested in the stability of certain desirable social-moral states. To wit, suppose a theorist claims that a particular state of affairs $s$ is an equilibrium within a larger social-moral system $S$. Schaefer seems to think that, before this theorist can investigate the stability of $s$, they must first establish a general existential claim: namely, that we can expect $S$ to contain at least one equilibrium (to be subsequently specified). But that is too strong. Since $s$ is the object of interest, the theorist need only establish *that s is an equilibrium*; if true, this can often be shown without establishing the general existential claim. So Schaefer's argument fails to give static political theorists a reason to rethink the burden of proof they bear; it remains the same as it ever was. Political theorists can continue analyzing social-moral states of affairs, safe in the knowledge that technically formidable fixed point theorems pose no threat to this enterprise.

## 1. WHAT GOES FOR STATES GOES FOR PROCESSES TOO

Schaefer defines an equilibrium state as a fixed point within a model system (1). Thus, to understand why we should doubt the existence of social-moral equilibria, we must first grasp the concept of a fixed point within a social-moral system (see 5). Suppose we describe social-moral states using a set of (real-valued) variables, of which there are an unspecified number $n$; thus, each possible social-moral state is identified by a vector of length $n$, and $\mathbb{R}^n$ defines the space within which possible states are located. Let $A$ be the set of points that contains all possible states. Let a *social-moral system* be a function $f{:}A \rightarrow A$ that takes a social-moral state as an input and returns a social-moral state as an output. A state $x^* \in A$ is a *fixed point* if and only if $f(x^*) = x^*$. Schaefer presents several examples to illustrate this definition. To apply this idea to social-moral systems, Schaefer interprets $f$ as a dynamic transition function: if $x$ describes the state realized by a social-moral system at time $t$, then $f(x)$ describes the state that emerges from $x$ at $t + 1$ (see 5).[2] Given this interpretation of $f$, the definition of

2 This temporal quality is added by Schaefer; it is not part of the general definition of $f$ as used in Kakutani's theorem.

a fixed point implies that, if $x^*$ is a fixed point, then, once $x^*$ is realized at some time $t$, $x^*$ is realized at every time thereafter (2).

We can now briefly restate Schaefer's skeptical argument:

1. Kakutani's fixed-point theorem enumerates four conditions that are jointly sufficient for the existence of an equilibrium state (4–5).
2. For each of Kakutani's four conditions, we can construct a plausible social-moral scenario for which that condition is violated (6–9).
3. "If these counterexamples . . . sound like plausible descriptions of our own social-moral systems, then we have reason to doubt that there exist any fixed points of justice" (5).
4. "It is reasonable, therefore, to suspect that social-moral systems may also resist equilibration" (1).

Below, we will show why, contrary to premise 3, examples of social-moral systems that violate Kakutani's conditions provide no reason to doubt the existence of social-moral fixed points. For now, we grant this premise for the sake of argument. A key upshot, according to Schaefer, is that political theorists should turn their attention from investigating the stability of desirable social-moral states to investigating the robustness of desirable social-moral processes (10–11). Whereas (equilibrium) states are static, like a snapshot of a social-moral system frozen in time, processes are dynamic, associated with "continual flux" and constant evolution. To capture the distinction, we might think of processes as collections of mechanisms by which social-moral states "arise and are swept away" (1).

A problem arises for Schaefer here: his argument can be used to raise doubts not only about the existence of social-moral equilibrium states but also about the existence of robust social-moral processes. Schafer raises doubts about the former by interpreting a mathematical object, $\mathbb{R}^n$, as defining the space of possible social-moral states. To do this, he interprets the dimensions of $\mathbb{R}^n$ as corresponding to variables we might use to describe the attributes of social-moral states. But we are not required to interpret $\mathbb{R}^n$ in this way, and we might just as well interpret the dimensions of $\mathbb{R}^n$ as corresponding to whatever variables we might use to describe the attributes of social-moral processes. Schaefer's distinction between "*process desiderata* and *state desiderata*" (12) cues us to this alternate interpretation. Just as states can be described as realizing, say, more or less social equality or more or less material welfare, so processes can be described as being better or worse at mitigating violent conflict or providing more or less protection for individuals' rights. So we can think of $A \subseteq \mathbb{R}^n$ as the set of possible social-moral processes. We can also reinterpret $f$ in a similar

way: let $x$ describe the operative process at $t$; then $f(x)$ describes the operative process at $t + 1$. Now, a process $x^* \in A$ is a fixed point if and only if $f(x^*) = x^*$.

Schaefer defines a "robust process" as a process that "maintains some of its qualitative features, even as … society shifts between distinct states" (11). In other words, given a dynamic interpretation of $f$, a robust process is a fixed point in the space of possible processes: assuming $x^*$ is a fixed point, once $x^*$ becomes operative at $t$, it will remain operative at $t + 1$.

We can now repurpose Schaefer's argument to raise doubts about the existence of robust social-moral processes, using his counterexamples to Kakutani's conditions as templates for producing our own counterexamples. Just as it "is not difficult to imagine that we could approach full equality or full despotism without ever completely realizing either" (6), so it is not difficult to imagine that we could tinker with a social process so that it continually gets better at mitigating violent conflict or protecting individuals' rights without ever completely eliminating violent conflict or perfectly securing individuals' rights. This gives us an analogue for Schaefer's counterexample to Kakutani's first condition, which requires that the space of possible processes be compact (see 5).[3] Similarly, just as we can imagine how a "public conception of justice might change slowly and continuously, like a stick gradually bending into an arc, until it reaches a critical point where the stick snaps, disrupting a continuous trend that preceded this new state" (7), so we can readily imagine a historical trend in which the social processes of production and wealth accumulation undergo small changes—for example, as inequalities of wealth and political power increase, bequests from parents to children and transfers of resources from the politically powerless to the powerful come to predominate—culminating in a social revolution that sweeps away the old processes of production and replaces them with something entirely different. This gives us an analogue for Schaefer's counterexample to Kakutani's third condition, which requires that $f$ be closed.

We could go on, but we trust we have made our point: if Schaefer has given us reasons to doubt the existence of social-moral equilibrium states, then we can use a slight reinterpretation of the mathematical objects on which his argument depends to generate reasons to doubt the existence of robust social-moral processes. If doubts about the existence of equilibrium states are enough to unsettle the case for doing static political theory, so too doubts about the existence of robust processes must be enough to unsettle the case for doing dynamic political theory.

3 That is, we can think of these hypothetical "perfect" social processes as the unattainable limit points to which sequences of feasible social processes converge, but since they lie outside the set of feasible social processes, that set is not compact.

## 2. MISPLACING THE BURDEN OF PROOF

No one should be unsettled by the preceding arguments, however, neither ours nor Schaefer's. A theorist who wants to establish that a particular social-moral state is an equilibrium is required to do nothing more (and nothing less) than demonstrate that the state in question is an equilibrium. To take Schaefer's example, if Rawls wants to establish that a society well-ordered by his principles of justice is an equilibrium, then he is required to do nothing more (and nothing less) than show that this state is an equilibrium. Contrary to what Schaefer argues, Rawls is not required to demonstrate that, in general, we can expect (unspecified) equilibrium states to exist.

To consolidate this point, let us consider some examples from applied game theory, which we can treat as a collection of models of limited social-moral systems. Suppose a theorist is studying the factors that foster social cooperation and uses Rousseau's "stag hunt" as a model for the relevant kinds of social interactions.[4] In a "stag hunt," players can either hunt stag together (cooperate) or hunt hare alone (go their separate ways). Suppose our theorist wants to show that the state in which the players cooperate is an equilibrium. Do they first need to demonstrate the general claim that there exists an equilibrium for the stag hunt? Of course not. They need only show that the state in which the players cooperate is an equilibrium. This can be shown directly, without establishing the general existential claim.

Indeed, this is the standard method of argument in applied game theory: the analyst explicitly identifies a particular profile of strategies and directly verifies that it satisfies the conditions for an equilibrium, rather than relying on theorems, like Kakutani's, to first establish that some (unspecified) equilibrium exists and only afterward identifying it explicitly. Very often, the assumptions of those theorems are not met in any case. Consider, for example, a seminal model in political science, the Hotelling-Downs model of electoral competition.[5] In this model, two candidates for political office compete for votes by choosing a "policy platform." The set of possible platforms is represented by some interval on the real number line—say, [0,1]—and the candidates can choose any platform within this interval. Each candidate prefers winning to tying and prefers tying to losing. Because the election outcome and thus candidates' payoffs are not continuous functions of the candidates' strategies, the game's best-response correspondence does not satisfy the continuity ("closed

4 Rousseau, "Discourse on the Origins and Foundations of Inequality among Men"; see also Skyrms, *The Stag Hunt and the Evolution of Social Structure*.

5 Hotelling, "Stability in Competition"; Downs, *An Economic Theory of Democracy*.

graph") assumption used in Kakutani's theorem.[6] By Schaefer's reasoning, we should now doubt that this model system has any equilibria and, so, we should hesitate to investigate the properties of any states within this system. Yet it is relatively straightforward to demonstrate that if voters have single-peaked preferences over the set of possible platforms, then the situation in which both candidates choose the median voter's preferred policy platform is an equilibrium; indeed, it is a unique equilibrium.[7] This can be proved directly, without first demonstrating the general claim that there exists an equilibrium for the game.

Schaefer claims that his aim is to "shift the burden of proof" onto static political theorists to "provide some reason to believe" the general claim that "fixed points of justice exist" (5). By this, he seems to mean that they must provide some reason to believe the social-moral system under examination satisfies general conditions ensuring the existence of an equilibrium state. But that is too much to require of a theorist who simply claims that some particular state $s$ is an equilibrium. To be sure, our theorist must show that $s$ is indeed an equilibrium (if that is what they claim). But they have always borne that burden, and Schaefer is wrong to argue that they must bear anything heavier.[8]

One might be persuaded by what we have said here yet struggle to see precisely where Schaefer's reasoning goes astray. In a diagnostic spirit, then, let us think about the general form of his argument:

6 To illustrate, let $m \in (0,1)$ be the location of the median voter's ideal point, and fix $x < m$. For each $n \in \mathbb{N}$, let

$$\hat{x}_n := x - \frac{1}{2n},$$

and

$$x_n := x - \frac{1}{n}.$$

Let

$$\mathbf{x} := (x, x),$$

$$\mathbf{x}_n := (x_n, x_n),$$

and

$$\hat{\mathbf{x}}_n := (\hat{x}_n, \hat{x}_n).$$

Given their payoff functions, for each player, $\hat{x}_n$ is a best response to the other player's choice of $x_n$ for all $n$. We can see that the sequence $(\mathbf{x}_n, \hat{\mathbf{x}}_n)$ converges to $(\mathbf{x}, \mathbf{x})$, yet choosing $x$ is not a best response when the other player chooses $x$, as required by the closed-graph assumption.

7 For details, see Gehlbach, *Formal Models of Domestic Politics*, 2–5.

8 If one's goal is merely to establish that a function $f$ has a fixed point, then demonstrating that a particular candidate $s$ (i.e., a point in the domain of $f$) is a fixed point is as good as any proof that $f$ satisfies general conditions like Kakutani's. Either approach demonstrates that $f$ has a fixed point.

1. Theorist *R* claims that a specific object *s* has property *F*.
2. By an abstruse mathematical theorem, we know that for any object *x* in the relevant domain, *x* has property *F* if *x* satisfies condition *C*.
3. Theorist *R* has given us no reason to believe that *s* satisfies condition *C*, and it is easy to imagine how objects that are similar to *s* might fail to satisfy condition *C*.
4. Thus, we should wonder whether any *x* in the relevant domain has property *F* and, in particular, whether *s* has property *F*, as Theorist *R* claims.

The conclusion does not follow, of course, for two reasons. First, the fact (if it is one) that a few objects that are similar to *s* violate condition *C* does not imply that all objects in the relevant domain violate condition *C* nor even that *s* violates *C*. Second, even if every object in the relevant domain fails to satisfy condition *C*, it could still be that some objects have property *F* since the theorem merely states that *C* is sufficient for *F*.

So theorists who claim that specific social-moral states are equilibria should be unperturbed by Schaefer's argument. If one wants to raise doubts about, say, Rawls's claim that a society well-ordered by his principles of justice (*s*) is an equilibrium (property *F*), then one should engage Rawls's argument for that specific claim rather than raise doubts about the existence of social-moral systems that satisfy Kakutani's conditions (condition *C*). After all, Rawls makes no use of Kakutani's theorem (nor do other theorists), and the assumptions of the theorem are not necessary for the existence of equilibrium states (5). More generally, if one wants to raise doubts about the whole enterprise of analyzing the properties of desirable social-moral states, one needs to do more than sketch a handful of cases that violate a set of conditions that are not necessary for the existence of fixed point equilibria.[9]

*University of California, San Diego*
*singham@ucsd.edu*
*dwiens@ucsd.edu*

9 We thank Zeynep Pamuk for discussion that helped to clarify our thoughts on Schaefer's article. We also thank Alexander Schaefer and Paul Weithman for comments on a previous draft.

## REFERENCES

Downs, Anthony. *An Economic Theory of Democracy*. New York: Harper and Row, 1957.

Gehlbach, Scott. *Formal Models of Domestic Politics*. New York: Cambridge University Press, 2013.

Hotelling, Harold. "Stability in Competition." *Economic Journal* 39, no. 153 (March 1929): 41–57.

Rousseau, Jean-Jacques. "Discourse on the Origins and Foundations of Inequality among Men." In *The Basic Political Writings*. 2nd ed. Translated and edited by Donald A. Cress, 27–92. Indianapolis: Hackett, 2011.

Schaefer, Alexander. "Is Justice a Fixed Point?" *American Journal of Political Science* (forthcoming). https://doi.org/10.1111/ajps.12631.

Skyrms, Brian. *The Stag Hunt and the Evolution of Social Structure*. Cambridge: Cambridge University Press, 2004.