Moral Diversity and Moral Responsibility

Forthcoming in *Journal of the American Philosophical Association*

Brian Kogelmann, University of Maryland, bkogel89@gmail.com

Robert H. Wallace, University of Arizona, rhwallace@email.arizona.edu

Abstract: In large, impersonal moral orders many of us wish to maintain good will towards our fellow citizens only if we are reasonably sure they will maintain good will towards us. The mutual maintaining of good will, then, requires we somehow communicate our intentions to one another. But how do we actually do this? The current paper argues that when we engage in moral responsibility practices – that is, when we express our reactive attitudes by blaming, praising, and resenting – we communicate a desire to maintain good will to those in the community we are imbedded in. Participating in such practices alone won't get the job done, though, for expressions of our reactive attitudes are often what economists call *cheap talk*. But, when we praise and blame in cases of moral diversity, expressions of our reactive attitudes act as *costly signals* capable of solving our social dilemma.

Key words: moral responsibility, Strawson, cheap talk, costly signal, moral disagreement

1. Introduction

Althea and Bertha inhabit the same society. They are not particularly close – they are neither friends, family, nor are they lovers – but still occupy the same community nonetheless. Since both Althea and Bertha are beings with a sense of morality, they wish to maintain relations of good will towards one another. Yet they are neither saints nor martyrs. As a result, they do not wish to display good will towards one another at all costs. More specifically, Althea and Bertha both wish to display good will towards one another *conditionally* – that is, Althea wishes to display good will towards Bertha only if she is reasonably sure that Bertha will display good will towards her, and vice versa. The worst-case scenario is to display good will towards one's fellow citizens, while one's fellow citizens harbor ill will towards one. This makes one a sucker, and no one wants to be a sucker.

Althea and Bertha thus find themselves being pulled in two opposing directions. On the one hand, they both most desire to live in a community where all display good will towards all. This suggests that they should display good will, in hopes of achieving the best possible outcome. But on

the other hand, both desire to avoid a society in which they display good will towards others, yet others display ill will towards them. This suggests that they should prudentially display ill will, so that they may avoid the worst possible outcome. How is this dilemma resolved? In order to secure the most desired outcome, Althea and Bertha need some way of communicating to one another that they both desire to maintain good will, thereby assuring one another. This is more difficult to do than one might initially think, however. A fundamental lesson of game theory is that (under certain kinds of incentive structures) when communication is not costly to engage in – that is, when talk is cheap – we cannot rationally infer facts about the intentions of others from their communicative acts. But on the flipside, when communication is costly – that is, when communication acts as a costly signal – we *can* rationally infer facts about the intentions of others from their communicative acts.

This paper proposes a mechanism – but certainly by no means the *only* mechanism – by which Althea and Bertha can solve their dilemma by appealing to this notion of costly signaling. The basic idea is that when we disagree with others about important moral questions, we incur costs from such disagreement. The fact that we willingly incur such costs – that is, the fact that we willingly engage in moral disagreement when we could otherwise simply remain silent – allows others to rationally infer that we wish to maintain relations of good will with them. The same applies to our ability to rationally infer the intentions of others when they similarly engage in moral disagreement. Hence, our central thesis: engaging in moral disagreement is sufficient but not necessary to maintain relations of good will within a community.[1]

The next section offers a more careful analysis of Althea and Bertha's dilemma and sketches in the abstract what a solution might look like (§2). Again, the general idea is that the two can assure one another of their desire to mutually maintain good will if they communicate their intentions to one another, which is possible so long as such communication is sufficiently costly. After this, we explore how Althea and Bertha might communicate their intentions to one another by turning to the moral responsibility tradition that follows the work of P.F. Strawson (§3). In short, when we express

[1] We thus join a growing body of literature in political philosophy that points out the positive benefits diversity and disagreement can bring about. As a few examples see Gaus (2016); Muldoon (2016); Kogelmann (2017). That said, we do not wish to suggest that there is something problematic about homogenous societies. It may be that deeply homogenous societies do not face the same incentives that Althea and Bertha face and thus do not play assurance games with one another. Or, if they do, there might be some other mechanism that solves their dilemma besides the one we highlight in this paper. Again, the claim is that moral disagreement is sufficient but *not* necessary to maintain mutual relations of good will.

our reactive attitudes and so blame and praise when such reactions are called for, we communicate our desire to maintain good will towards others. Problematically, though, expressions of our reactive attitudes can be quite cheap. Yet expressions of our reactive attitudes are not cheap – that is, expressions of our reactive attitudes are costly – when we blame and praise in an environment of moral heterogeneity (§4). Hence our central thesis: engaging in moral disagreement is sufficient to secure a community in which all display good will towards all. We end by further refining this thesis, highlighting conditions that must hold for the mechanism we identify to successfully function. Not only must we have the *right kind* of moral diversity for this solution to work, but we must also have the *right attitude* towards moral diversity as well (§5). This latter point, we believe, fails to characterize much contemporary public discourse.

2. The Possibility and Fragility of Mutual Good Will

Let us make more precise the problem highlighted in the introduction. Althea and Bertha both wish to maintain good will towards one another but do not wish to do so at all costs. Neither wants to display good will if the other harbors ill will. We can model this situation game theoretically, as shown in Figure 1. Here, Althea and Bertha both most prefer to maintain good will towards one another. Althea's second most preferred outcome is where she harbors ill will towards Bertha but Bertha maintains good will towards her – here, Althea can gain advantages in her social interactions with Bertha for Bertha is being a sucker. The same applies for Bertha's second most preferred outcome. The third most preferred outcome for Althea and Bertha is where both harbor ill will towards one another. Finally, the least preferred outcome for both parties is where that party maintains good will towards the other but the other harbors ill will – here the party harboring good will is a sucker. Basically, in Figure 1 the desire to display good will towards others is *conditional* – the players wish to maintain good will towards others only if they are reasonably sure they will receive good will in return.[2]

		Bertha
		Maintain good will	Harbor ill will

[2] Compare to John Rawls's definition of reasonable persons, defined such that they "are ready to propose principles and standards as fair terms of cooperation and to abide by them willingly, *given the assurance that others will likewise do so*" (Rawls 1993/2005: 49) (emphasis ours).

		Bertha: Maintain good will	Bertha: Harbor ill will
Althea	Maintain good will	3, 3	0, 2
	Harbor ill will	2, 0	1, 1

Figure 1

The technical term for the game illustrated in Figure 1 is an *assurance game*. Assurance games present an interesting problem. Both the outcome (Maintain good will, Maintain good will) and the outcome (Harbor ill will, Harbor ill will) are pure Nash equilibrium solutions. Roughly, they are both outcomes to the game that we could expect rational parties to settle on. But, clearly, the mutual maintaining of good will is preferred to the mutual harboring of ill will by both parties. So what, then, is the problem? Why wouldn't Althea and Bertha simply both harbor good will towards one another? Even though the mutual harboring of good will is an equilibrium outcome in this game, and even though the mutual harboring of good will yields the Pareto dominant equilibrium when compared to the risk-dominant equilibrium (where both parties harbor ill will), it is not obvious that rational players will actually achieve such an outcome. The slightest doubt that Bertha will harbor good will can force Althea to harbor ill will so as to avoid the sucker payoff, where she harbors good will and Bertha harbors ill will (Aumann 2000). Brian Skyrms nicely summarizes the problem: "rational players are pulled in one direction by considerations of mutual benefit and in the other by considerations of personal risk" (Skyrms 2004: 3).

So the dilemma highlighted in Figure 1 presents a tricky social problem. But to what extent does this dilemma actually obtain? That is, when do persons actually face the incentives highlighted in Figure 1? We typically want to maintain good will towards our friends, family, and lovers *unconditionally*. Even if our children harbor ill will towards us, we still want to harbor good will towards them – Figure 1 is thus not applicable in such a case. But Figure 1 plausibly models our interpersonal relationships with *some*. When it comes to strangers on the street, other drivers on the highway, and fellow imbibers at the bar, we want to maintain good will towards them only if we are reasonably sure they will maintain good will towards us. If others maintain ill will towards us it is foolish – and, in some cases, dangerous – to continue harboring good will back towards them.

Indeed, Figure 1 plausibly models the relations between those persons constituting what F.A. Hayek called "the Great Society": "a large-scale moral order in which we often confront others as strangers." In such an environment we "know little about the other, except that she is a moral

person" (Gaus 2011: 268). Despite a lack of interpersonal closeness our relationships to strangers are still moral relations: we still wish to offer them *some* modicum of good will, even if this good will is conditional.[3] It is these kinds of impersonal relationships and the problems they raise that we shall be interested in for the remainder of the paper, and not those deeply personal relationships that constitute close friends, families, and small groups. To put it another way: we are interested in those relationships among persons who desire to maintain good will conditionally, and, we believe, such a set of persons is coextensive with those persons in large, impersonal orders. If one believes that persons in large, impersonal orders wish to maintain good will unconditionally (as we grant is likely the case among friends and family) then the forthcoming analysis will not be compelling. We, however, believe this to be implausible.

So assuming Figure 1 models a nontrivial subset of our interpersonal relationships, we thus face an assurance problem: how can we be sure that other persons in our community will maintain good will towards us, and how can they be sure that we will maintain good will towards them? One *prima facie* plausible solution is that assurance can be achieved through acts of communication. Althea says: "I promise that I will maintain good will towards you if you promise to maintain good will towards me." This seems like it should assure Bertha to harbor good will rather than ill will. And if Bertha says something similar then this should assure Althea to display good will as well. As a result, both parties display good will and the most desirable equilibrium attained.

But though intuitively plausible, this will not do. Suppose Althea, through an act of communication as suggested above, tries to signal to Bertha that she will harbor good will in order to induce Bertha to harbor good will as well. This is indeed Bertha's best response to Althea's harboring good will. But what should Bertha infer from Althea's communicative act? Given knowledge of Althea's preferences, Bertha knows that it is in Althea's interest *no matter what* to induce Bertha to harbor good will. If Bertha harbors good will, the *worst* outcome Althea can achieve is a payoff of 2 – if Althea harbors ill will and Bertha harbors good will. If, on the other hand, Bertha harbors ill will, the *best* outcome Althea can achieve is a payoff of 1 – if both players harbor ill will. Given this, Bertha cannot infer from Althea's communicative act that Althea will actually harbor good will. Whether or not Althea actually plans to harbor good will, it is in Althea's interest

[3] Indeed, much work in evolutionary psychology and evolutionary accounts of moral norms suggest that we are such conditional cooperators. See, for instance, Bowles and Gintis (2013).

for Bertha to believe that Althea will harbor good will (so Bertha then harbors good will as well), and thus in Althea's interest to say she will in fact harbor good will.

The problem here is that Althea's communicative act is *cheap talk*. Cheap talk is defined as communication that does not affect the payoffs of a game (Crawford and Sobel 1982). When Althea tells Bertha that she will harbor good will it doesn't cost her anything; likewise, when Bertha tells Althea that she will harbor good will as well. Now suppose that, while telling Bertha that she will harbor good will, Althea pulls out a knife and cuts her hand open. Why? Because cutting her hand open isn't cheap.[4] Althea is attempting to show Bertha that she really is committed to harboring good will. Bertha cannot infer a commitment on Althea's part from cheap talk. But why, if she plans to harbor ill will anyways, would Althea cause herself pain? If she plans to harbor ill will, cutting her hand open would likely be irrational. She would be imposing a loss on herself with seemingly no compensating benefit. But, if Althea plans to harbor good will then the behavior is rational. Althea is willing to lower her utility to show Bertha that she is committed to harboring good will. The pain from the knife wound is a loss, but a loss that Althea believes will be more than compensated if both parties inhabit a society where mutual good will is maintained. In other words, Bertha can infer that the knife cutting is an honest *costly signal* by Althea.[5]

Let us try to make this notion of costly signaling more precise. In the above Figure 1 we used ordinal utilities – all we know is that higher numbers are more preferred to lower numbers. Figure 2 now represents the same assurance game with cardinal utilities – now utilities represent *intrapersonal* intensity of preferences, such that the gap between the numbers now represents how much more preferred an outcome is. Importantly, cardinal utilities typically do not contain information about *interpersonal* intensity of preferences (though they can be defined in a manner such that they do contain such information).[6]

	Bertha	
	Maintain good will	Harbor ill will

[4] Here we follow Kogelmann and Stich (2016: 723-724).
[5] By "cost" we follow the standard meaning in game and decision theory, where to impose a cost on some individual *i* is to reduce *i*'s utility, which just means that *i*'s preferences are not as well satisfied as they otherwise would be. There are thus no actions that are costly *as such*, as costs depend on preferences. When we thus make judgments in the forthcoming analysis about what kinds of actions are costly we are implicitly making judgments about how preferences typically look. It is of course logically possible there exists preferences such that the kinds of actions we say are costly are not costly given these preferences. We are not interested in logical possibility, but rather empirical likelihood.
[6] For more on the distinction between ordinal and cardinal utilities see Kogelmann and Gaus (2017).

Althea	Maintain good will	10, 10	0, 3
	Harbor ill will	3, 0	2, 2

Figure 2

Suppose cutting her hand open costs Althea a utility of 0.5. This behavior could be rational no matter how Althea plans to act. If she plans to harbor good will and successfully convinces Bertha to harbor good will then she gains a utility of 9.5 compared to the outcome that otherwise would have obtained, where she harbors good will, but Bertha does not (because her payoff in this case is zero). But, if Althea plans to harbor ill will and successfully convinces Bertha to harbor good will then she has gained 0.5 utility compared to the outcome that otherwise would have obtained, where both Althea and Bertha harbor ill will (because her payoff in this case is two). No matter what, Althea gains utility, so sending such a signal could be rational regardless her intentions. But now suppose that cutting her hand open costs Althea utility s, where $1 < s < 8$. In this case, cutting her hand open is rational only if Althea plans to harbor good will. For suppose Althea plans to harbor ill will. Then, Althea has caused herself more pain than she gains from Bertha's switch from harboring ill willing to harboring good will. Given that this would be irrational, Bertha should conclude that Althea's cutting her hand open is a sincere signal when doing so costs Althea utility s. It can easily be checked that any time Althea and Bertha play an assurance game with the ordinal structure of Figure 1 or Figure 2 then there will *always* be some costly signal Althea can send that would *only* be rational to send so long as she plans on harboring of good will. Likewise for Bertha. It is the mutual sending of such costly signals that can assure both parties to harbor good will, rather than harbor ill will. This achieves the desired outcome, solving our dilemma.

§3 Communicating Good Will

The last section presented a tricky social problem and sketched in broad, abstract terms what a potential solution might look like. When citizens engage one another in large, impersonal orders they often face the incentives of an assurance game. Because they are moral beings they wish to mutually maintain good will towards one another – this is the most desired outcome, and an equilibrium solution to the game. Since both want to avoid being suckers they could quite easily end

up in a situation where all harbor ill will towards all – this is the undesirable equilibrium they could very well settle at. To make sure mutual good will is maintained, persons must assure one another. But how? One way is through communication, but *only if* communication is sufficiently costly.[7] Yet how can we communicate our desire to maintain good will towards our fellow citizens? And, when we do this, will such communication be costly enough? This section answers both questions in turn.

In terms of our first question – how can we communicate our desire to maintain good will towards our fellow citizens? – there already exists an answer in the moral responsibility literature. Moral responsibility takes as its subject those who are apt candidates for moral demands, and thus those who are liable to be the target of moral sanctioning practices from the community they are embedded in. Though an important question, we are not interested in what it means to be an apt candidate for moral demands – whether one must act freely and with sufficient knowledge or something of the like. Rather, we are interested in the moral sanctioning practices themselves. That is, we are interested in how it is we hold persons morally accountable, and what it is we are doing when we do hold persons morally accountable. Given Althea and Bertha's assurance problem, we assume that the moral standard to which we hold others accountable in large-scale impersonal orders is a standard of *reasonable good will*. Failure to meet this standard makes blame fitting. Typically, exceeding this standard makes praise fitting.

For many in the moral responsibility literature, our moral sanctioning practices are fundamentally (or paradigmatically) public and communicative (e.g., Watson 1987/2008: 116-121; Darwall 2006: 70-74; McKenna 2012: 174-175). Strawson's seminal "Freedom and Resentment" (1962) first articulated a theory of moral responsibility that highlights this feature of praise and blame. On this view, praise and blame are expressed *reactive attitudes*, emotions that are particularly reactive to the intentions of other persons. According to Strawson's view, for Althea to blame Bertha is for her to express to Bertha that she believes Bertha has acted with ill will (and so acted wrongly) by resenting Bertha. For Althea to praise Bertha is for her to express that she believes Bertha has acted with good will by showing Bertha gratitude. Fundamental for our purposes,

[7] As an anonymous reviewer pointed out, this claim might sound false if taken too literally. We do not mean to say that persons are actively calculating when they pass each other on the street. Instead, we offer a model detailing how a tricky social problem might be overcome that may explain features of our shared world, even if it fails to be descriptive of individual cognitive processes. Compare: teenagers who eat Tide Pods are probably not consciously rationally calculating the benefits of such behavior. Nevertheless, a costly signaling model may help explain why such an action might be individually rational for a teenager to perform. As a dangerous and risky action, it acts as a costly signal of one's status and fitness (Murphy 2018).

though, is the fact that the moral responsibility practices between Althea and Bertha are communicative – when the two hold one another responsible, they are communicating things to one another via their expressed reactive attitudes.

The Strawsonian model is a well-developed view of moral responsibility and is particularly useful for articulating a potential solution to Althea and Bertha's problem. To make things more precise, we will say that when we blame someone we communicate to them through expressed reactive attitudes that we think they have done a wrong and so acted from ill will. Likewise, when we praise, we commend an action as performed with good will. Call instances of this kind of communication through expression of our reactive attitudes *accountability signaling*, for such acts of communication signal to others that we hold them accountable for their good or ill will. Accountability signaling is primarily about communicating to other members of our community what we believe the quality of their will to be.

It would be wrong to think, though, that when we engage in moral responsibility practices we *only* signal to those we praise and blame. For when engaging in such practices we also communicate something about ourselves. When we blame those who ought to be blamed we communicate that we have good will to others who bear witness to the expression of our reactive attitudes; likewise when we praise those who are deserving of such approval. As Angela Smith points out, blame has the important secondary function of asking for "moral recognition" of the wrong done "by the wider moral community" (Smith 2012: 44). By responding to this call for recognition, we signal our own quality of will not only to the person wronged but also to the broader community at large. And if we fail to express indignation on some stranger's behalf when they are wronged, then we have indicated that we lack a reasonable degree of good will towards strangers generally. Call this kind of communication – not done to the perpetrator of a wrong or right action alone but to the community we are imbedded in at large – *quality signaling*, for we are signaling the quality of our will, be it good or ill.[8] There are thus two kinds of signaling and two different ways we communicate to others when we engage in moral responsibility practices. We signal to others that we believe they

[8] More generally, quality signaling occurs whenever one acts intentionally, and not only when one engages in moral responsibility practices. Consider, for instance, recycling and polluting. The fact that one goes out of one's way to recycle signals (not always honestly) that one cares about the environment and one's fellow inhabitants of the world, thus implying that one has good will. And, the fact that one throws one's water bottle on the side of the road signals (again, not always honestly) that one lacks care and respect for the environment and one's community, thus implying that one has ill will.

have acted with either good or ill will (this is accountability signaling), and, in doing so, also signal to others around us the quality of our own will in the process (this is quality signaling).

Although we proceed with the Strawsonian account of moral responsibility in mind, we note that the framework we develop extends to other accounts of our moral responsibility practices. Often, even those who reject the Strawsonian view take interest in the public and communicative aspects of moral responsibility, thus implying that our notions of accountability and quality signaling plausibly extend. T.M. Scanlon, for instance, significantly modifies the basic ideas articulated above, arguing that blameworthy actions "show something about the agent's attitudes towards others that impairs the relations that others can have with him or her," including the basic moral relationship all persons have towards one another (Scanlon 2008: 128). To express blame, then, precisely points this fact out – in other words, it *communicates* something to those one is blaming.[9] Others, like Pamela Hieronymi (2001), Matthew Talbert (2012), and Angela Smith (2012), see blame as a form of moral protest. As before, such blaming *communicates* to others beliefs about their actions, attitudes, or character. Shared among a diverse array of views about our moral responsibility practices is thus the idea that when we engage in the relevant practices we also engage in a communicative enterprise.

So, we have an answer to our first question (how can we communicate our desire to maintain good will to our fellow citizens?): by participating in moral responsibility practices. Following our analysis of Strawsonian accounts of moral responsibility, we can say that expressions of our reactive attitudes communicate a desire to maintain good will towards others. When we are indignant at acts that display ill will we signal that we have good will to others around us, through acts of what we have called quality signaling. And when we praise acts that display good will we again signal that we have good will towards those around us, through acts of quality signaling. This quality signaling can assure others we wish to maintain good will towards them. And if others quality signal through expression of their reactive attitudes then they communicate to us that they wish to maintain good will towards us.

But this solution works *only if* such instances of quality signaling are not cheap talk. This leads into our second guiding question for this section: will such instances of signaling be sufficiently costly? Problematically, expressions of our reactive attitudes are often quite cheap. If Althea says:

[9] Importantly, Scanlon's account is such that it is possible to *privately* blame others, in which case blame is not communicative (Scanlon 2008: 187-188). Our paper is only concerned with the public expression of blame (and other reactive attitudes as well), remaining agnostic on whether blame can indeed be private.

"Hey! What's the matter with you?" then plausibly such a display of indignation does not cost Althea much in terms of making her preferences less well satisfied than they otherwise would be, were she not to express indignation. And if Althea says, "Well done!" then such an act of praise does not cost her much either. But if these expressions of Althea's reactive attitudes cost her nothing then others around Althea cannot rationally infer that she wishes to maintain good will. Likewise applies to Althea's ability to infer that those around her wish to maintain good will. We have not solved our assurance dilemma quite yet.

Now surely, it might be objected, some expressions of our reactive attitudes are costly. Indeed, think of how uncomfortable it can be to tell one's friend or partner that they have screwed up – such discomfort is certainly a cost, making the expression of resentment difficult. But recall the central target of our inquiry: we are focusing on those interpersonal relationships that arise in communities among relative strangers, for it is *these* sorts of relations that we believe give rise to the assurance problem we are concerned with. So, though resentment among those we are close with might be quite costly, it intuitively seems like expressions of our reactive attitudes among persons we lack deep interpersonal relationships with will be cheap. This suggests that expressions of our reactive attitudes such as praise and blame are not performing the communicative function we wish for them to perform – such expressions are not allowing others to rationally infer the quality of our will when we engage in quality signaling. We thus still do not have a way of assuring those who inhabit large-scale impersonal orders that we wish to maintain good will towards them. And, they have no way of assuring us.

4. When Blaming is Costly

The last section argued that (*i*) engaging in moral responsibility practices through expressions of our reactive attitudes communicates one's desire to maintain good will towards others through acts of what we have called quality signaling, but (*ii*) many times expressions of our reactive attitudes are quite cheap, implying that such instances of quality signaling cannot solve our assurance problem, for others cannot rationally infer anything about our intentions from such acts. But on the flipside, if expressions of our reactive attitudes were sufficiently costly then we would have a plausible mechanism by which our assurance problem can be solved. In these sets of cases, persons

can draw rational inferences about our intentions. To this end, the current section proposes a mechanism by which our quality signaling acts as a costly signal.[10]

To get a feel for the mechanism that induces costs on our quality signaling, consider a case.

Honor Killing. Dupree resides in a community where honor killings are performed against women who violate tradition sexual norms. Cassidy, a young woman in the relevant community, engages in premarital sex. As a result, the community kills Cassidy according to the relevant honor killing standards. In response, Dupree is indignant at the members of his community: "What is wrong with you? Are you all depraved? This is one of the most morally disgusting things I have ever witnessed!" Dupree says this despite not knowing Cassidy personally.

What is the likely effect of Dupree's expression of indignation? Since the community, by and large, thinks that honor killings should be done, those in the community that are subject to Dupree's indignation will likely push back: they might judge Dupree, shun Dupree, or engage in some kind of formal or informal sanction against him. Perhaps Dupree is now no longer able to shop at their stores. As such, Dupree's expression of his reactive attitudes in this case is *costly*. It costs him something to engage in quality signaling.

Intuitively it does seem clear that Dupree's actions are costly. But what, exactly, induces the costs? The key here is that when quality signaling in cases of moral diversity Dupree opens himself up to sanction and criticism from the community at large. It is not easy to be told that one is a moral monster or even simply just misguided about some crucial moral issue; suffering backlash from those who hold differing moral standards than one does can be a difficult thing indeed. As such, the *psychological costs* of being criticized, reprimanded, and sanctioned are the relevant costs that one undergoes when one quality signals in cases of moral diversity, though there might be other possible costs as well: at the limit, bodily harm done by those one has disagreed with. We are primarily interested in the relevant psychological costs. One may suffer backlash when one praises and blames

[10] We by no means suggest this is the only mechanism. Hence our central thesis: that moral diversity is *sufficient*, but not *necessary*, to secure a community in which all display good will towards all. One mechanism that we do not explore is opportunity costs: that, when one's quality signaling is particularly extensive and requires great time and effort to carry out, one has cost one's self. We believe that acts of "moral grandstanding" as discussed by Justin Tosi and Brandon Warmke (2016) are ways of inducing such costs, thereby serving a critical assurance function even if, as the authors argue, such acts are generally morally suspect.

in cases where others disagree that praise and blame are what is called for. This is the cost one undergoes when one quality signals in environments of moral heterogeneity.

Because Dupree's quality signaling is costly in *Honor Killing*, though the members of the community might be upset with Dupree for issuing judgments that call into question their moral norms, there *is* something they can be sure of: that Dupree wishes to maintain a good will towards them. For if Dupree did not wish to do this, then imposing this cost on himself through quality signaling would likely be irrational. It is only because Dupree wishes to genuinely maintain good will towards others that he is willing to engage in costly signaling. If he wished to maintain ill will in the community then the rational thing to do would be to stay silent. Or perhaps he joins in with the community even though he disagrees, with the hope that doing so tricks others into thinking he wishes to maintain a good will. From there he may harbor ill will and achieve the payoff where he exploits and others are suckers.

Here, it might objected: why should those Dupree blames and expresses indignation towards interpret him as harboring good will? For from their perspective, Dupree has a flawed moral view and, if he had things his way, would replace the current (correct) moral norms with incorrect ones. Though it is true those Dupree blames disagree sharply with the norms Dupree propagates, this does not mean that they will be able to rationally infer from this point of disagreement that Dupree harbors ill will towards them. Indeed, the fact that Christians and Muslims disagree over important moral questions does not allow all those in one group to rationally infer that all those in the other group harbor ill will.[11] Instead, those persons Dupree blames must ask themselves: given the costs Dupree faces for his disagreement, what reason does Dupree have, *other than that he cares about what is right*, for speaking up? If no alternative answer can be given here,[12] those around Dupree should interpret his costly signaling as evidence of him seeking to harbor good will, as being motivated by

[11] Though the inference is not rational, some do in fact draw such an inference. We think drawing such inferences is indicative of one of the problematic attitudes towards moral disagreement – what we call *hubristic realism*, discussed in §5.2 below – that, when it obtains, makes it such that the costly signaling mechanism we identify no longer successfully functions.

[12] Note that it *is* possible to give other answers here. For instance, a white man publicly complaining about the unfairness of affirmative action hiring practices will likely face backlash for his indignation. This is costly. Does he express indignation because he genuinely cares about what is right (thus implying he has good will)? Perhaps. But he could also be doing so because a change in norms would benefit him. Thus, in cases where self-interest aligns with the position one costly signals in favor of, drawing rational inferences about good will from costly signaling becomes difficult. Notice, though, that we have stipulated that Dupree and Cassidy do not have a close interpersonal relationship. Thus, it is rational for members of his community to infer his concern for others generally from his costly disagreement.

the right kinds of reasons (even if one argues for the incorrect moral view) is indicative of having a good will (Arpaly 2003: 79, 84-115).

Honor Killing is an extreme case, but we believe the lesson of this case generalizes to all cases where there is significant moral diversity: cases where persons disagree about the relevant moral norms governing the circumstances in question. Consider, for instance, students who shut down a debate at a liberal arts college because the speaker is controversial. A rogue student who signals her indignation at her fellow classmates typically does so in a costly way; there may be significant repercussions from those who disagree with her. As such, her quality signaling is costly. Other examples abound: being indignant on Twitter towards Black Lives Matters protesters after they have destroyed property if one has many left-leaning Twitter followers; displaying resentment towards privileged white people who voted for Trump when one lives in a community full of Trump supporters; and standing up and arguing that young girls ought to be given the same education as their male peers when one lives in Pakistan. Here, there is no doubt among anyone that Malala Yousafzai wishes to maintain good will, and for good reason: Malala's quality signaling was costly indeed.

Importantly, we are not arguing that Dupree and others in similar circumstances are morally required to express their reactive attitudes in cases of moral disagreement. We are merely highlighting the consequences of Dupree's doing this: when Dupree decides to express indignation at those who radically disagree with him, such quality signaling acts as a costly signal capable of solving our assurance problem. Of course, given the potential benefits of Dupree's actions, one could perhaps argue that Dupree is indeed morally required to do so. We shy away from such arguments for space constraints, though we do note that such arguments might exist. We are merely highlighting what happens when Dupree does engage in moral responsibility practices in an environment of moral heterogeneity.

So here is where we stand. Normal displays of quality signaling when engaging in moral responsibility practices likely cannot assure our fellow citizens that we wish to maintain good will towards them and vice versa. Such quality signaling is often too cheap. As a result, we can end up in a situation where everyone displays ill will towards everyone else, even though everyone wishes to maintain good will towards everyone else. There is, however, at least one way in which quality signaling can achieve its intended communicative function: when one quality signals in cases of

moral diversity. For when one does quality signal in cases of moral diversity, one imposes costs on one's self. This allows others to rationally infer one's intentions.

5. Refining the Moral Diversity Solution

The last section argued that in environments of moral heterogeneity expressions of our reactive attitudes can act as costly signals capable of solving the assurance problem characterizing large, impersonal orders. The final section of this paper refines this thesis. First, we note that only *certain kinds* of moral diversity will allow expressions of our reactive attitudes to act as effective costly signals. Moreover, even when this right kind of diversity obtains it is still not sufficient for expression of our reactive attitudes to be costly. We also must take seriously the possibility of genuine moral disagreement, as well as take seriously the reprimand from those who disagree with the moral standards we propagate. In other words, we must foster the right kinds of *attitudes* towards moral diversity.

5.1 *The Right Kind of Moral Diversity*.

In cases of moral diversity instances of quality signaling act as costly signals assuring others we wish to maintain good will towards them. Disagreeing with others over deep moral questions is costly to the point that it often only makes sense to do so when one wishes to achieve the (Maintain good will, Maintain good will) outcome from Figure 1. From this it might be thought that a simple corollary follows: the more diversity, the better. For the more disagreement there is among the relevant community the greater the likelihood that our quality signaling is costly. And, when expressing our reactive attitudes is costly, persons are capable of assuring their fellow citizens that they wish to maintain good will towards them.

This inference is incorrect. To see why, consider a case.

Soles. Esau and Franklin live in the same community, even though they have radically different backgrounds. Esau is from an Arabic country, Franklin from a Western liberal democracy. A third party wrongs Franklin in some way. Esau, who sees this happen, takes his shoe off and points the sole towards the offender, in an effort to signal his indignation at

> the perpetrator through an act of accountability signaling, along with the quality of his will towards Franklin through an act of quality signaling. Franklin is confused by Esau's actions, so he does not interpret Esau's actions as Esau wishing to maintain good will.

What has gone wrong here? In many Arab cultures displaying the soles of one's shoes is a particular way of signaling indignation. When Esau stands witness to Franklin's being harmed by some third party, Esau tries to signal his indignation towards this third party. If it all goes well, Esau should successfully signal to the perpetrator that he thinks the perpetrator has shown ill will through an act of accountability signaling, and he also should have successfully signaled to Franklin the quality of his will through an act of quality signaling. But he has at least failed to do the latter, for Franklin, when he sees Esau take off his shoes and point the sole towards the perpetrator, is simply confused by Esau's behavior. Here, diversity has caused problems rather than offered solutions. The diversity between Esau and Franklin seems too great in *Soles*, causing a breakdown of communication. We are thus left with a somewhat pessimistic and trivial conclusion: diversity is sometimes capable of solving our assurance problem, except when it cannot.

We now wish to make more nuanced our claim that moral diversity is capable of solving our assurance problem by inducing costs on our quality signaling. Consider two ways in which diversity might manifest itself.

> *Token diversity*. Token diversity characterizes Esau and Franklin's relationship if and only if they disagree about *how it is* they ought to engage in accountability and quality signaling.
>
> *Felicity diversity*. Felicity diversity characterizes Esau and Franklin's relationship if and only if they disagree *under what conditions* they ought to engage in accountability and quality signaling.

With this distinction in hand we now refine the central thesis of our paper: in order to solve the assurance problem, there needs to be significant felicity diversity but very little token diversity. In other words, when it comes to *how it is* we engage in accountability and quality signaling we seek homogeneity; when it comes to *under what conditions* we should engage in accountability and quality signaling we seek heterogeneity.

Why is this so? The problem with token diversity is that there may be confusion concerning when a signal is a signal. This seems to be what went wrong in *Soles*. Here, Esau tried to signal his indignation towards the perpetrator and his quality of will towards Franklin by displaying the soles

of his shoes. But because Esau and Franklin disagree about how it is they ought to express their reactive attitudes – Franklin thinks a simple "Hey buddy what's the matter with you?" is what's called for – Esau's signal fails to achieve its intended outcome. Franklin does not interpret it as a signal in the first place due to the pair's token diversity. Hence successful communication has broken down.

Note that many of the instances of successful costly signaling we have discussed thus far in the paper can fail as successful instances of communication when token diversity obtains. Consider the first case we introduced in §2 above: where Althea and Bertha must assure one another that they will mutually display good will. The example of costly signaling we used was Althea cutting her hand open – something that most agree is costly. But now suppose Althea and Bertha are from radically different cultures: Althea is from a wealthy Western liberal democracy and Bertha is from a pre-agricultural foraging society. To convince Bertha that she really intends to maintain good will, Althea burns a pile of money, which is the common example of costly signaling used in the technical game theory literature.[13] But in this case Bertha does not know what money is. To her, Althea sets aflame funny looking pieces of paper. As such, Bertha cannot interpret Althea's costly signal as a genuine costly signal because token diversity obtains between the two. If there is token homogeneity, though, then these problems dissipate.

But this same homogeneity cannot characterize the conditions under which we hold people responsible for the moral diversity solution to work. For as we have seen, it is precisely in cases of moral diversity – in terms of felicity conditions specifically – that expressions of our reactive attitudes can become costly, for we open ourselves up to sanction from those whose moral norms we criticize. It is this costliness that assures others we wish to maintain good will towards them. Hence, when it comes to moral diversity broadly construed we face a double-edged sword: diversity in terms of how we express our attitudes can induce noise, but diversity in terms of the moral standards we hold each other accountable to is one mechanism that can help facilitate a community in which all maintain good will towards all.

[13] Indeed, the title of Bernheim and Redding (2001)'s paper is "Optimal Money Burning: Theory and Application to Corporate Dividend Policy."

5.2 *Taking Disagreement Seriously*.

Unfortunately, significant felicity diversity along with high levels of token homogeneity within a community are still not sufficient to ensure that costly signals are sent capable of solving our assurance problem. Along with felicity diversity and token homogeneity, persons in a shared community must have the right kinds of attitudes towards the diversity they confront. There are two ways of responding to disagreement that, if they obtain, cause problems for our proposal. We now wish to highlight these problematic attitudes.

To highlight the first problematic attitude, consider the following variation on *Honor Killing*.

> *Silent Honor Killing*. Dupree resides in a community where honor killings are performed against women who violate tradition sexual norms. Cassidy, a young woman in the relevant community, engages in premarital sex. As a result, the community kills Cassidy according to the relevant honor killing standards. Though Dupree disagrees sharply with the community's standards he does not express any indignation, for he has no reason to think his moral standards are superior to the community's or vice versa.

What went wrong in *Silent Honor Killing*? In the case as stated Dupree does not express his indignation towards the community and thus fails to send any kind of signal at all, costly or not. Clearly this is a problem, for it is the sending of costly signals that solves our assurance problem, and Dupree sends no such signal in the current case.

The more general worry with *Silent Honor Killing* is that Dupree embraces one common – all too common among our undergraduate students – way of responding to the fact of moral disagreement, what we shall call *naïve relativism*. Far from actually having any interesting meta-ethical commitments, the naïve relativist reasons in something like the following manner: because there is no obvious way to judge between competing moral demands, there is no point in engaging in argumentation over these demands. Naïve relativists like Dupree believe they lack the moral standing to blame, given that his view is no better than anyone else's. So he fails to publically express his reactive attitudes. It should be clear why this sort of attitude towards moral disagreement is a problem. On the view advanced in this paper, it is precisely *public* disagreement with conflicting moral demands that acts as a costly signal capable of solving our assurance problem. But if persons are naïve relativists then they see no point in disagreeing with others at all and thus will send no

signal in the first place. So the first presumption our proposal makes when it comes to attitudes towards moral disagreement is that persons are not naïve relativists. They think there is at least *some* point to engaging in substantive moral debate with others.

The importance of avoiding naïve relativism seems rather clear. But there is another kind of attitude towards moral diversity that also causes problems for our proposal. To highlight this second problematic attitude, consider another case.

> *Trump*. Guyute blames a Trump voter for supporting what Guyute believes to be an unconscionable candidate. The Trump supporter fights back, telling Guyute that he ought to be ashamed for passively standing by and watching his country slowly but surely be torn apart by cheap immigrant labor while simultaneously more and more jobs get shipped overseas. Far from being taken aback by such reprimand, Guyute finds the Trump supporter's indignation amusing. He believes Trump supporters are not the kinds of people worth taking seriously and finds it quite funny whenever one expresses indignation towards him. Nevertheless, Guyute is often indignant at what Trump supporters say and do.

Now what has gone wrong in this case? Perhaps at first glance nothing, for maybe Guyute is correct that it is not worth engaging with certain kinds of persons who hold certain kinds of views, thus ignoring the reprimand aimed at him. But on a second pass there is some cause for concern. Guyute intuitively bears no costs for his actions in the current case, for, in not taking seriously the Trump supporter's expressions of indignation, Guyute does not suffer the psychological costs we argued are typically associated with moral disagreement. Since Guyute predictably suffers no costs in these cases, the expression of his own reactive attitudes towards others is cheap. This raises the worry that others in the community (especially the Trump supporter herself) might not think that Guyute is committed to maintaining a good will, leading us back to our original assurance problem. Perhaps counter-intuitively, taking seriously the Trump supporter would be costly for Guyute and thus assure others (and especially the Trump supporter herself) that he has a desire to maintain good will.

Guyute harbors the second problematic attitude towards moral disagreement, what we call *hubristic realism*. Also free from any kinds of interesting meta-ethical commitments, the hubristic realist reasons in something like the following way: even though people disagree with me when it comes to the appropriate moral standards for governing social life, I need not care about such

disagreements, for it is my view that it is the correct view and everyone else is deeply misguided.[14] The problem here is that when one is a hubristic realist several of the costs associated with moral disagreement simply vanish.[15] If one thinks that one's interlocutors are fools then one bears little psychological costs when one is reprimanded by them. Indeed, at the limit, moral reprimand from those one disagrees with can actually provide net utility gain, for it can be amusing to take cheap shots from what one deems to be the cheap seats. But when this is true expressing the reactive attitudes in environments of felicity diversity will not be costly; instead, such expressions will be quite cheap. This leaves our assurance problem unresolved.

Note that it is an implicit assumption of many of the cases we have given thus far that those expressing the reactive attitudes in cases of moral disagreement are not hubristic realists. Returning to the original *Honor Killing*, if Dupree does not care that the rest of his community thinks he is deeply misguided and also does not care about the sanctions imposed on him then the reprimand he receives for his expressions of blame will not be costly. What was a case of costly signaling is now cheap talk. The same applies to other cases: if the student who is indignant at her fellow classmates for shutting down a campus speech simply does not care about what her peers think of her then such indignation is no longer costly. If one does not care that one's Twitter followers rake one over the coals for blaming Black Lives Matter protesters, then, again, such an instance of blame is now cheap. And finally, as highlighted by *Trump*, if one thinks the backlash one receives from Trump supporters is funny or amusing after one blames them then, again, one's blaming them is no longer costly.[16]

---

[14] Another way of thinking about hubristic realism is in terms of whether disagreement counts as a form of evidence that one is misguided in one's own moral convictions. For the hubristic realist, disagreement is often seen as an insidious effort to undermine community, rather than as genuine evidence that one might be mistaken. This is what may happen in the aftermath of *Honor Killing*. Dupree's community may fail to see his blame as a costly signal because, of course, someone who believes that honor killing is wrong must be wicked to their core. Those who avoid hubristic realism, though, often view disagreements as evidence that one could be in some sense mistaken.

[15] Several, but not all. Sometimes there are more than psychological costs one bears when one vocally expresses moral disagreement – indeed, one could suffer serious bodily harm from such an act. Being a hubristic realist does not make these costs go away.

[16] To continue drawing connections between our work and that of Rawls's (see footnote 2 above), the avoidance of hubristic realism is related to one of the characteristics of a reasonable person. For Rawls, the burdens of judgment are those factors that give rise to genuine good-faith disagreements among persons (Rawls 1993/2005: 56-57). Part of being a reasonable person (according to Rawls's definition) is to accept the burdens of judgment as an explanation as to why persons disagree (Rawls 1993/2005: 54). But in accepting this it is hard to see how one could also be a hubristic realist. For if one accepts that a set of factors lead to good-faith disagreements then it is hard to see how one could also think others are so deeply misguided to the point that they are not worth taking seriously. This just means, though, that one is not a hubristic realist.

So, engaging in moral responsibility practices in cases of felicity diversity acts as a costly signal *only if* we have the right kinds of attitudes towards moral diversity. First, we must avoid being naïve relativists when confronted with disagreements for, if we are naïve relativists, then we likely will not send any signal at all in the first place. And second, we must avoid being hubristic realists for, if we are hubristic realists, then we may not bear sufficient costs when we express our reactive attitudes in an environment of moral heterogeneity. To sum up: we must engage in moral discourse as if there is something important at stake to be resolved, but also with the humility that our positions on these moral questions may be wrong, and that those we disagree with – sometimes radically disagree with – may in fact be right. When a community is constituted by persons so described in which there is also sufficient felicity diversity with high levels of token homogeneity then we have reason to believe that all will be assured to continue displaying good will towards all.

6. Conclusion

It is by no means the case that our communities will be well-functioning. Indeed, though many of us are fortunate enough to live in Western liberal democracies where we are reasonably sure that others will display good will towards us and where others are reasonably sure that we will display good will towards them, much of the world lives in chaos rather than harmony. Beyond contributing to the literature on moral responsibility, we hope that our paper offers one small insight concerning why this disparity might be so. In communities where differences are stark the intentions of our fellow citizens may be more transparent than in cases where we are characterized by sameness. Though perhaps a bit counterintuitive, such a fact might help explain why some of us do in fact live better together.

Works Cited

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

Aumann, Robert. (2000) "Nash Equilibria Are Not Self-Enforcing." In *Collected Papers, Vol. 1*: 615-620. Cambridge: The MIT Press.

Bernheim, Douglas B. and Lee S. Redding. (2001) "Optimal Money Burning." *Journal of Economics and Management Strategy* 10: 463-507.

Bowles, Samuel and Herbert Gintis. (2013) *A Cooperative Species*. Princeton: Princeton University Press.

Crawford, Vincent P. and Joel Sobel. (1982) "Strategic Information Transmission." *Econometrica* 50: 1431-1451.

Darwall, Steven. (2006) *The Second-Person Standpoint*. Cambridge: Harvard University Press.

Gaus, Gerald. (2011) *The Order of Public Reason*. Cambridge: Cambridge University Press.

Gaus, Gerald. (2016) *The Tyranny of the Ideal*. Princeton: Princeton University Press.

Hieronymi, Pamela. (2001) "Articulating an Uncompromising Forgiveness." *Philosophy and Phenomenological Research* 62: 529-555

Kogelmann, Brian. (2017) "Justice, Diversity, and the Well-Ordered Society." *The Philosophical Quarterly* 67: 663-684.

Kogelmann, Brian and Gerald Gaus. (2017) "Rational Choice Theory." In *Research Methods in Analytic Political Theory*, edited by Adrian Blau: 217-242. Cambridge: Cambridge University Press.

Kogelmann, Brian and Stephen G.W. Stich. (2016) "When Public Reason Fails Us: Convergence Discourse as Blood Oath." *American Political Science Review* 110: 717-730.

McKenna, Michael. (2012) *Conversation and Responsibility*. Oxford: Oxford University Press.

Muldoon, Ryan. (2016) *Social Contract Theory for a Diverse World*. New York: Routledge.

Murphy, Ryan. Working Paper, April 2018. "The Rationality of Literal Tide Pod Consumption." Available at SSRN: https://ssrn.com/abstract=3160438

Rawls, John. (1993/2005) *Political Liberalism*. New York: Columbia University Press.

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

Skyrms, Brian. (2004) *The Stag Hunt and the Evolution of Social Structure*. Cambridge: Cambridge University Press.

Smith, Angela. (2012) "Moral Blame and Moral Protest." In *Blame: Its Nature and Norms*, edited by Justin Coates and Neal Tognazinni: 27-48. Oxford: Oxford University Press.

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

Talbert, Matthew. (2012) "Moral Competence, Moral Blame, and Protest." *Journal of Ethics* 16: 89-109.

Tosi, Justin and Brandon Warmke. (2016) "Moral Grandstanding." *Philosophy & Public Affairs* 44: 197-217.

Watson, Gary. (1987/2008) "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme." In *Free Will and Reactive Attitudes*, edited by Michael McKenna and Paul Russell: 115-142. Burlington: Ashgate Publishing.