The Meta-Wisdom of Crowds[1]

Justin Sytsma (Victoria University of Wellington)
Ryan Muldoon (University at Buffalo)
Shaun Nichols (Cornell University)

It is well-known that people will adjust their first-order beliefs based on observations of others. We explore how such adjustments interact with second-order beliefs regarding universalism and relativism in a population. Across a range of simulations, we show that populations where individuals have a tendency toward universalism converge more quickly in coordination problems, and generate higher total payoffs, than do populations where individuals have a tendency toward relativism. Thus, in contexts where coordination is important, belief in universalism is advantageous. However, we also show, across a range of simulations, that universalism will enshrine inequalities and eliminate diversity, and in these cases it seems that relativism has its own advantages.

1. Introduction

Some judgments seem to be universally true. This expresses a second-order belief about these judgments, which we'll label "universalism." Beliefs about many mathematical judgments, e.g. *that 3×4 = 6+6*, provide familiar examples. Other judgments seem to be true only relative to some context. This expresses an opposed second-order belief, which we'll label "relativism." Beliefs about many claims concerning date and time, e.g., *that it's 3PM*, provide familiar examples.[2] There has been a swell of work trying to chart which kinds of judgments people tend to regard as universal and which they tend to regard as relative (Goodwin and Darley 2008, Wright et al. 2013). There's also been work trying to determine how people make inferences about which judgments are universal and which relative (Wright et al. 2014, Goodwin and Darley 2012; Ayars and Nichols 2020). Here we want to explore the social effects of adopting a universalist or relativist stance about some matter.

One of the central ideas we will explore is how a tendency toward universalism might facilitate coordination. There are important precedents on the matter. In his classic book *Ethics,* Mackie maintains that our ordinary concept of morality carries metaethical presuppositions that align with our treatment of universalism, although he puts this in terms of morality being objective (see Footnote 2). Famously, Mackie maintains that there is no objective morality and

[1] Forthcoming in *Synthese*. We would like to thank James Beebe, Justin Bruner, Jerry Gaus, Daniel Simpsonbeck, an anonymous referee for *Synthese*, and the audience at the 2019 Formal and Experimental Workshop at Northeastern University for their helpful comments and suggestions.

[2] The terminology in the empirical literature on folk metaethics is not yet systematic. Relativism is the central notion of interest for us. According to relativism about morality, there is no single true morality (Harman 1985). We use "universalism" as the contrary of relativism, such that universalism about morality holds that there *is* a single true morality (e.g., Wong 2006). "Objectivism" is another term that is contrasted with relativism (including in Nichols 2004 and Mackie 1977, 36-38). But "objectivism" is naturally treated as the contrary for subjectivism rather than relativism (see, e.g., Finlay 2007), so we prefer "universalism" as the closest to an antonym for "relativism."

so ordinary judgments about morality are false. But he maintains that the belief there is a single fact of the matter about moral questions might play an important role in organizing social life. He writes:

> We need morality to regulate interpersonal relations, to control some of the ways in which people behave towards one another, often in opposition to contrary inclinations. We therefore want our moral judgments to be authoritative for other agents as well as for ourselves: objective validity would give them the authority required (1977, 43).

In a similar vein, Kyle Stanford has recently argued that the metaethical belief that morality is objective (or "externally imposed" [2018, 2]) functions to facilitate cooperation, as reflected in situations like the prisoner's dilemma. He writes:

> experiencing moral demands and obligations as externally imposed simultaneously on both ourselves and others ensures that if I myself come to be motivated to conform to a particular norm or standard of behavior that I experience as distinctively moral in character, I automatically demand that others conform to it as well, judging them to be less attractive potential partners in social interaction generally if they do not (2018, 8).

If I regard moral demands as externally imposed (as opposed to generated by something like self-interest), then I will be less inclined to cooperate with those who defy the moral demands.

Mackie and Stanford have substantive proposals for the potential function of metaethical beliefs, and the effects of such beliefs on social life. We will examine the effects of meta-evaluative beliefs as well. But our approach will diverge from theirs in a few ways. First, we want to explore second-order beliefs concerning universalism and relativism more generally. We don't intend to restrict the investigation to moral topics. Neither do we restrict the investigation to beliefs about externally imposed demands. Second, our investigation will not be focused narrowly on issues of cooperation but on broader issues about coordination, which can occur in the absence of competing interests. Third, we will take a more quantitative approach to evaluating these issues by conducting simulations in which we vary the extent to which members of a population favor universalism or relativism. Finally, unlike some recent work that is more evolutionary in focus, such as that by Brian Skyrms (1996) and Cailin O'Connor (2019), we adopt an epistemic approach in which people take themselves to be trying to find out what's the case, rather than what is strategically advantageous. One benefit to our approach is that it relies on domain general learning strategies that people are known to employ.

1.1 Background on folk metaethics

Extant work on beliefs about universalism and relativism have revealed significant variation. First, and perhaps least surprisingly, there is variation by domain. People tend to be universalists about scientific claims and relativists about claims of taste (Goodwin and Darley 2008). There is also variation by individuals and groups. For instance, younger people are more likely to express relativistic views about morality than are older people (Beebe & Sackris 2016). Most surprisingly, there is variation about universalism/relativism *within* domains. For instance, with regard to the ethical domain people in the US tend to be universalists about the moral status of bank robbery, but not about the moral status of culturally contentious issues like abortion or euthanasia (Goodwin & Darley 2008, Wright et al. 2014).

As these examples might suggest, one feature that seems to influence whether a claim is treated as universally true or only relatively true is the degree of perceived consensus regarding the claim. For instance, there is a correlation between the extent to which a moral claim is regarded as universal and the degree of perceived consensus surrounding the claim. In addition,

consensus information about a claim directly influences people's beliefs about whether the claim is universal or relative (Goodwin & Darley 2012, Ayars & Nichols 2020). People are more likely to treat a moral claim as universal when they are told that the vast majority of people agree about the issue as compared to when they are told that a bare majority of people agree about the issue.

2. Modeling Universalism

2.1 From consensus to universalism and back again

Studies on consensus show that people's beliefs about others' first-order beliefs (e.g., whether most people make the same judgment about some matter) affect second-order beliefs (e.g., whether the judgments on this matter are universally or only relatively true) (Goodwin and Darley 2012; Ayars and Nichols 2020). But it's also the case that second-order beliefs should have implications for first-order beliefs. In particular, if one has a second-order belief of universalism, then, ceteris paribus, one should pay more attention to the crowd, being more likely to change one's first-order belief to conform with the consensus. By contrast, if one has a second-order belief of relativism, then there would be no push to conform, since a relativist can countenance that different first-order beliefs can be equally correct.[3]

The basic idea here can be illustrated by considering domains with different prior expectations about universalism versus relativism. Let's say that there is a disputed claim, with 75% of people on one side and 25% of people on the other. If the disputed claim is a scientific claim, ceteris paribus, one would be inclined to think that the 25% are making a mistake. But if instead the disputed claim is about whether a piece of music is beautiful, one would be inclined to think that the minority has a different aesthetic sensibility (such that relative to the minority's sensibility the music is beautiful but relative to the majority's sensibility it isn't), and that no mistake is being made. The important point is that relativists and universalists about a given judgment will tend to have different responses to minority opinions. Universalists presuppose that there is a single fact and so will tend to treat minority opinions as noise. Relativists do not presuppose that there is a single fact and so will be less inclined to treat minority opinions as noise and more inclined to treat minority opinions as reflecting legitimate (i.e., non-mistaken) alternative perspectives.

2.2 Operationalizing universalists and relativists

With these considerations in place, we will characterize agents as *universalist* or *relativist* based on how they treat their first-order beliefs in response to evidence about consensus. Universalists will believe that there is a fact of the matter scoped such that disagreements mean that one side is making a mistake. Relativists, on the other hand, will assume that lack of consensus is simply consistent with no universally-scoped fact of the matter. In these circumstances, a universalist will want to be on the right side of a factual dispute, potentially giving her a reason to change her mind about her first-order belief. A relativist, on the other hand, will just take note of the diversity of belief, comfortable in the thought that no one need be mistaken.

Where this approach is made somewhat more complex is that agents can update these second-order beliefs. A universalist who sees even splits in first-order beliefs will start believing that perhaps the relativists have a point and will ultimately switch to relativism. A relativist who

[3] We treat this as a conceptual claim here, taking these responses to follow from the way we've characterized universalism and relativism. It's an open empirical question, however, whether people's second-order beliefs affect their first-order beliefs in this way.

encounters lopsided disagreements will start believing that perhaps the universalists were on to something. How one updates one's first-order beliefs depends on one's second-order beliefs, but second-order beliefs can also be revised based on the evidence.

Thus, to model the effect of universalism on coordination, we need agents with two kinds of belief about a given claim, each of which could be revised in response to new information about the beliefs of others:

i. A first-order belief φ

ii. A second-order belief about universalism/relativism that applies to the belief φ

There are different ways to operationalize these notions. But as a first pass, we will propose the following as characterizing how someone holding a second-order belief of universalism (UNI) adjust their *first-order beliefs* based on consensus.[4] For some consensus threshold *X* indicating majority belief (e.g., notably greater than 50%):

- If first-order belief is ~φ and second-order belief is UNI and consensus >= *X*% believe φ, switch to φ
- If first-order belief is φ and second-order belief is UNI and consensus >= *X*% believe ~φ, switch to ~φ

That is, if a person with the second-order belief of universalism registers that a suitable majority of people believe φ, then that agent will either persist in that belief or update it. People holding the second-order belief of relativism (REL), as we will treat them, do not adjust their first-order beliefs based on consensus directly, but they might change their second-order belief based on a clear enough consensus, becoming universalists, and now holding the second-order belief of universalism can subsequently update their first-order belief as laid out above.[5] Thus we also need to characterize how agents adjust their *second-order beliefs*. For some threshold *Y* indicating lack of strong consensus (suitably close to 50%), and some threshold *Z* indicating high consensus (suitably close to 100%), where $50\% < Y < X < Z < 100\%$:

- If UNI and consensus < *Y*% φ and < *Y*% ~φ, switches to REL
- If REL and consensus >= *Z*% φ or >= *Z*% ~φ, switches to UNI

That is, if a universalist sees that consensus is suitably close to an even split, that agent will revise their second-order belief to relativism. And if a relativist sees that consensus is suitably close to full agreement, that agent will switch their second-order belief to universalism.

The values for the three consensus thresholds (*X*, *Y*, and *Z*) will need to be explored, and might vary between groups or individuals. Following Simpsonbeck and Sytsma (ms), we'll use a threshold value of 70% for revision of first-order belief (*X*). We will then explore the values for revision of second-order beliefs to relativism (*Y*) and to universalism (*Z*).

3. Initial Simulations

All simulations were run in R, extending the work of Simpsonbeck and Sytsma (ms) to include changes in second-order beliefs. Since our focus here is on how changes in second-order beliefs in response to consensus affect convergence on first-order beliefs in a population, we otherwise

[4] Note that this is distinct from the "conformity bias" approach first found in Boyd and Richerson's *Culture and the Evolutionary Process*. That model, and others in that tradition, focuses on first order beliefs only.

[5] Of course, this is a simplification. For one thing, on our models, the only way for consensus to lead a relativist to change her first-order view is for her first to become a universalist. But a relativist might change her view simply because her view is in a tiny minority, without giving up on her relativism. To take a homey example, if 49% of the people think that it's summer and 49% of the people think that it's winter and 2% of the people think that it's fall, then if I find myself in that 2%, I'm likely to change my first-order belief without becoming a universalist about seasons (see, e.g., Ayars & Nichols 2020, experiment 3).

adopted the default model they developed. This includes that in each simulation, there is a population of 1000 individuals. Further, at the start of each simulation, each individual has a randomly assigned first-order belief (φ, ~φ) and a randomly assigned second-order belief (UNI, REL). Individuals assess consensus in the population by sampling the first-order beliefs of a random subset of the population, with their assessment being updated over time: at each timestep, a random 1% of the population assesses the first-order beliefs of a random 1% of the population (excluding themselves) and do so with perfect accuracy.[6] Each individual remembers up to 100 evaluations, but shows a recency bias, favoring their more recent evaluations.[7]

After assessing consensus, it is possible for individuals to update their beliefs. Individuals first look at whether they should update their second-order belief, then their first-order belief. In our first set of simulations, the threshold for changing their second-order belief from universalist to relativist (*Y*) was set at 55%: if an individual holds UNI and their consensus estimate is less than 55% φ and less than 55% ~φ (i.e., greater than 45% φ), then that individual will switch to REL. The threshold for changing their second-order belief from relativist to universalist (*Z*) was set at 95%: if an individual holds REL and their consensus estimate is at least 95% φ or at least 95% ~φ (i.e., 5% φ or less), then that individual will switch to UNI. After assessing their second-order belief, individuals assessed their first-order belief. The consensus threshold for change in first-order beliefs (*X*) was set at 70%: if an individual holds UNI and their consensus estimate is 70% or greater for φ they will switch to φ (if they held ~φ); if that individual holds UNI and their consensus estimate is 70% or greater for ~φ they will switch to ~φ (if they held φ).

In our first set of simulations, the probability that an individual would hold φ was varied across the simulations from 0.2 to 0.8 in 0.02 increments. The probability that an individual would hold UNI was independently varied from 0 to 1 in 0.25 increments. This makes for a total of 155 simulations in the first set. Each simulation was run over 250k timesteps, plotting time to convergence (how long it took for the entire population to converge on one first-order belief; that is, the population coming to 100% φ or 100% ~φ). The results are shown in Figure 1.

The first set of simulations indicates that when there is high initial consensus, you get rapid convergence on a single first-order belief regardless of the starting second-order beliefs in the population. But as initial consensus decreases, the higher the initial probability of an individual holding UNI in the population, the more likely it is that the population will converge and converge more quickly. For instance, when the population starts out 100% universalist, rapid convergence occurred except when the starting probability of φ was 0.48 or 0.50. In contrast, when the population starts out 100% relativist, convergence was only seen when the probability of φ was less than 0.32 or greater than 0.64. Inside this range, the population did not converge in the 250k timesteps.

This first set of simulations suggests that the tendency for a population to converge on a first-order belief becomes much stronger as the tendency toward universalism about that belief goes up, even if there is relatively little first-order consensus to begin with. This makes sense: to believe in universalism is to believe that there is a correct first-order belief, and in a population where people call on the wisdom of the crowd to help determine what that correct belief is, universalism will naturally promote convergence. Although we expected universalism to

[6] See Simpsonbeck and Sytsma (ms) for simulations involving error, including both sampling error and recall error.
[7] Following Simpsonbeck and Sytsma (ms), we used a 2x recency multiplier: the first set of evaluations by a given individual was given a base weighting, then each subsequent set of evaluations by that individual was weighted two times the previous weighting.

promote convergence given the operationalization, the results vividly illustrate how strongly a universalist presupposition facilitates convergence.

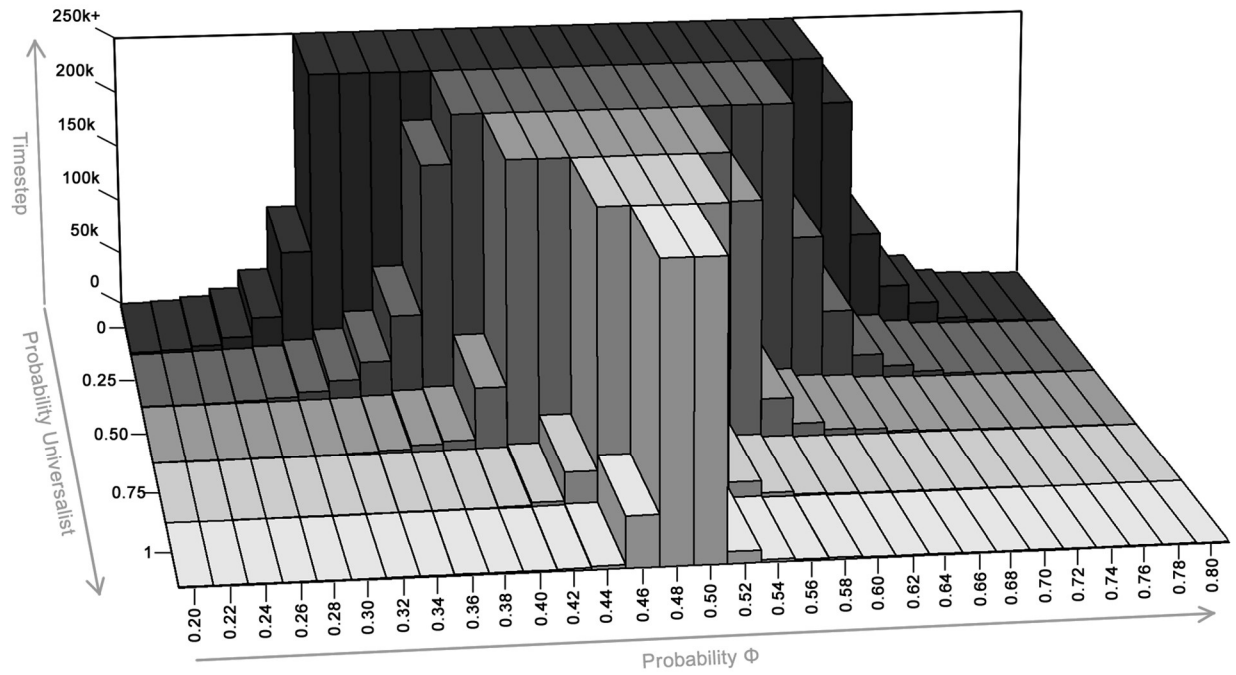

Figure 1: Time to convergence for each simulation in Set 1 up to 250k timesteps, varying the starting probability of φ and the starting probability of UNI.

3.1 Emphasizing universalism

To further explore how universalist presupposition facilitates convergence on a single first-order belief in a population, in our second set of simulations we varied the thresholds for changing second-order belief to favor more readily switching to universalism (lower *Z*) and to disfavor switching to relativism (lower *Y*).

In our initial set of simulations, we set *Z* to 95%, which means that individuals would only switch from relativism to universalism if they had reason to believe that 19/20 people in the population held the same belief. In the present simulations, we lowered this threshold from 95% to 55% in one percentage point increments. Similarly, in our initial simulations, we set *Y* to 55%, which means that individuals would switch from universalism to relativism if up to 11/20 people in the population held the same belief. In the present simulations, we lowered this threshold from 55% to 51% in one percentage point increments. The result is a total of 205 simulations in the second set. In each simulation, the probability that an individual would initially hold φ was set to 0.5, and all individuals started out as universalists. All other details of the simulations matched those from the first set. The results are shown in Figure 2.

The second set of simulations indicates that even when the initial consensus is evenly split between two first-order beliefs, a strong enough preference toward universalism can lead to rapid convergence on a single belief in a population. Thus, we find that convergence occurs if individuals in our population are marginally less likely to switch second-order beliefs from universalism to relativism and/or marginally more likely to switch second-order beliefs from

relativism to universalism from what was specified in the first set of simulations. For instance, in these simulations we find that if we lower the threshold for switching to universalism by just one percentage point from what we used in the previous simulations (shifting *Z* from .95 to .94) and the threshold for switching to relativism by just two percentage points (shifting *Y* from .55 to .53), the population now converges on one belief.

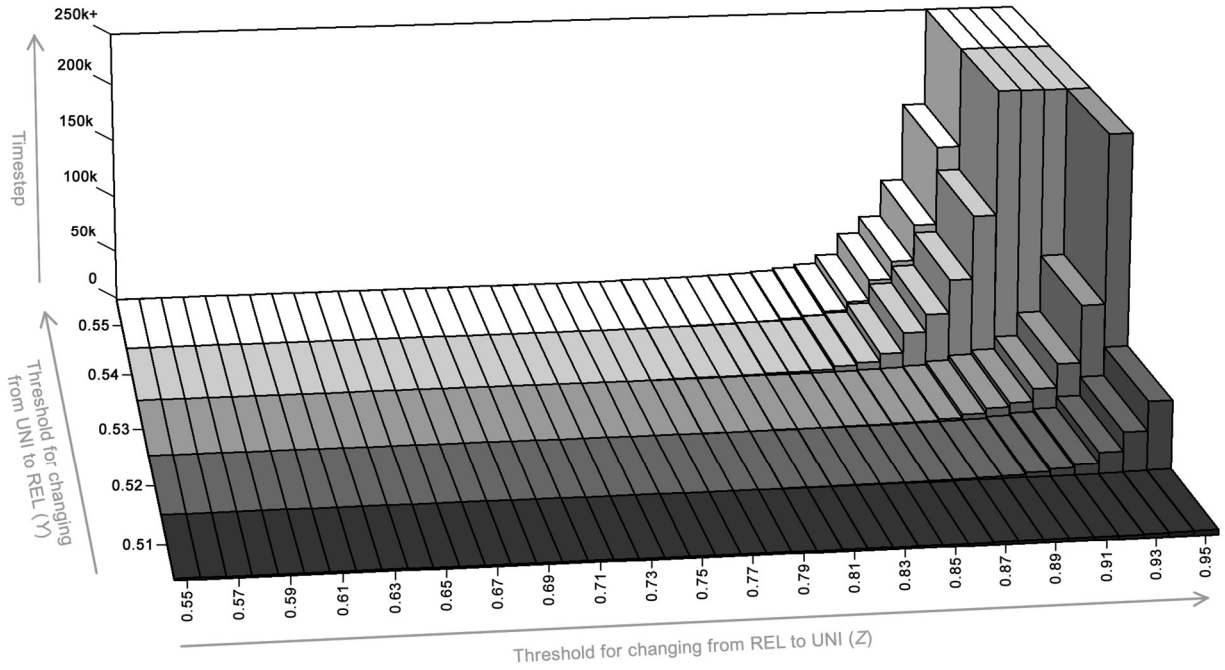

Figure 2: Time to convergence for each simulation in Set 2 up to 250k timesteps, varying the threshold for changing from UNI to REL (*Y*) and the threshold for changing from REL to UNI (*Z*).

4. Universalism as a tool for coordination

Thus far we've seen how evidence from consensus will lead populations favoring universalism to converge on a single view more quickly than populations favoring relativism. We can naturally extend these findings to the context of coordination games that specify the payoffs of people's beliefs in pairwise interactions.

Coordination games are usually construed in terms of strategic interactions of agents, where players seek to maximize their self-interest, and players coming to the same belief will maximize benefits. While this is a useful and common framework, this is not what we do. Instead, we show that coordination can be achieved just with agents trying to determine the correct thing to believe. Doing so, we might then construe coordination games as problems in social learning. Universalists think that there is a single right answer that they're trying to discern. And this needn't involve strategic thinking. That is, there need not be any deliberation about how to directly maximize one's own benefits. Rather, people are simply applying some principles of social learning to arrive at beliefs. This is not unlike trying to uncover social or natural regularities. One benefit to this approach is that such social learning is ubiquitous and domain general. And most of the time this won't involve any sort of dilemmas where issues of

acting strategically come into play—people are just trying to find a good way to wash potatoes, or to keep the roof from leaking, and so on.

4.1 Coordination game 1: Choosing Greetings

In some games, there is a common coordination interest among all players but there are two options that are equally good, as in choosing a greeting when handshakes are no longer safe (Figure 3). In this case, the benefits come from coordinating on either of these two choices, not on which of the choices is arrived at.

		Other(s)	
		Bow	**Wave**
Self	Bow	1, 1	0, 0
	Wave	0, 0	1, 1

Figure 3: Payoff matrix for Choosing Sides game

In our third set of simulations, we extended the first set to track the population-wide benefits of beliefs when individuals periodically participate in a game structured like Choosing Greetings. To do this, the setup from the first set of simulations was modified so that at the end of each timestep, a random 2% of the population was paired up and those pairs played a coordination game with the payoffs shown in Figure 3. Here, "Bow" was equated with φ and "Wave" was equated with ~φ. This means that when two individuals both holding φ played the game, they would each get a payoff of 1. And the same for two individuals both holding ~φ. But if one individual held φ while the other held ~φ, neither would get a payoff. The same 155 simulations were run as shown in Figure 1, but this time we tracked total payoffs across the games, which was plotted as the percentage of maximum possible total payoff for the simulation. The results are shown in Figure 4.

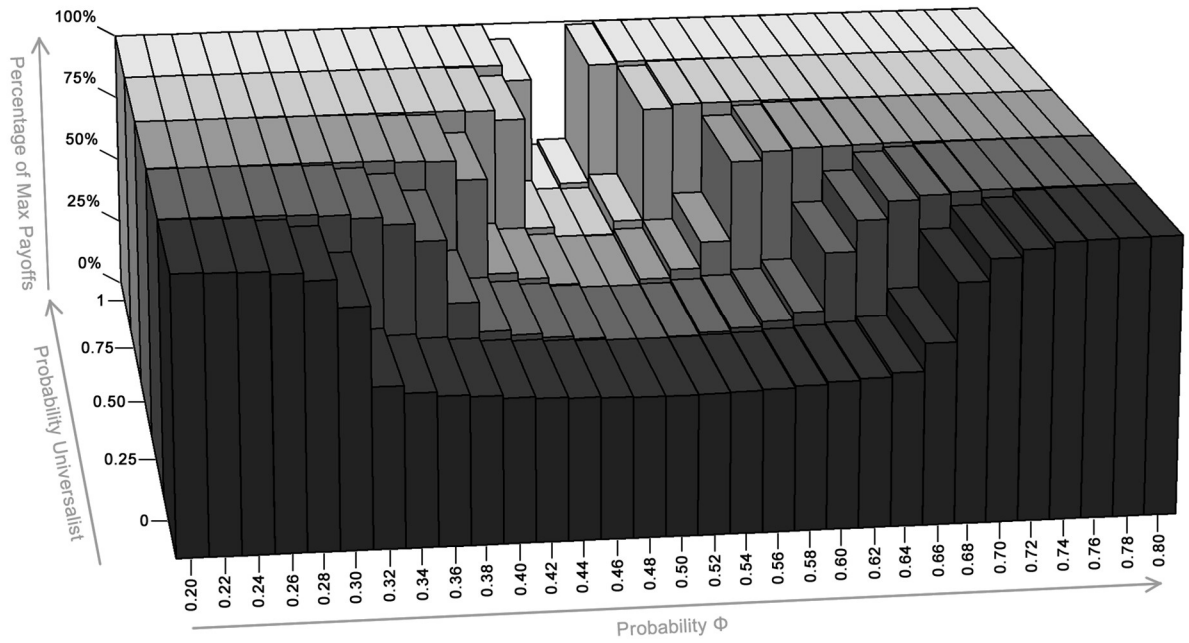

Figure 4: Percentage of maximum payoffs according to matrix for Choosing Sides for each simulation in Set 3 up to 250k timesteps, varying the starting probability of φ and the starting probability of UNI.

Since the individuals did not adopt their beliefs strategically, the convergence effects are the same as in our initial set of simulations (see Figure 1). However, the payoffs for the individuals depend on whether the "players" hold the same first-order belief or not. As such, we would expect that the more quickly population converges to a single belief, the greater will be the total payoffs for the population in the Choosing Greetings game. And this is exactly what we see in the graphs: populations with a greater preference for universalism do much better than populations with a greater preference for relativism. Thus, not only does a preference for universalism promote convergence compared to a preference for relativism, but in basic coordination games, it also leads to much better outcomes for the population.

4.2 Coordination game 2: Stag Hunt

There are, of course, a wide variety of coordination games. We will consider just one more for present purposes, an assurance game represented by the Stag Hunt (Figure 5). In this game, if we coordinate and both go for the stag, we will both do well, but if you go for a stag and I go for a hare (the lower left cell in Figure 5), my reward will be meager and you will go hungry as a result of our failure to coordinate.

		Other(s)	
		Stag	**Hare**
Self	Stag	2, 2	0, 1
	Hare	1, 0	1, 1

Figure 5: Payoff matrix for Stag Hunt game

In our fourth set of simulations, we used the same specifications as we did for the third set involving the Choosing Greetings game, changing only the payoff matrix to reflect that shown in Figure 5. The results are shown in Figure 6. Although the trend is not as dramatic, once again we find that a preference for universalism tends to lead to better overall outcomes. This makes sense: a preference for universalism promotes convergence, but in the absence of a feedback mechanism between the outcomes of the game and individual beliefs (a development that is beyond the scope of the present paper, but see Section 7), this can lead to either convergence on φ (which can be equated with "Stag"), and with it an optimal result, or convergence on ~φ (which can be equated with "Hare"), which gives a sub-optimal result, as is seen in the lower values on the left hand side of Figure 6. Further, quicker convergence on ~φ when beliefs are relatively evenly split has the potential to slightly depress total payoffs, as can be seen in the couple of spots in the graph where a preference for relativism generates slightly higher payoffs than a preference for universalism (e.g., when the probability of φ is 0.48). Still, overall a preference for universalism again produces better results than a preference for relativism.[8]

[8] We also ran a simulation using stag hunt with higher payoff ratios (namely, getting a stag yielded 10 units for each agent, while the rabbit remained at 1 unit for each). This made it clearer that a preference for relativism produces somewhat better results than a preference for universalism when the simulation starts with an initial probability of φ that is slightly lower than 0.5. But again, overall preference for universalism produces better results overall.

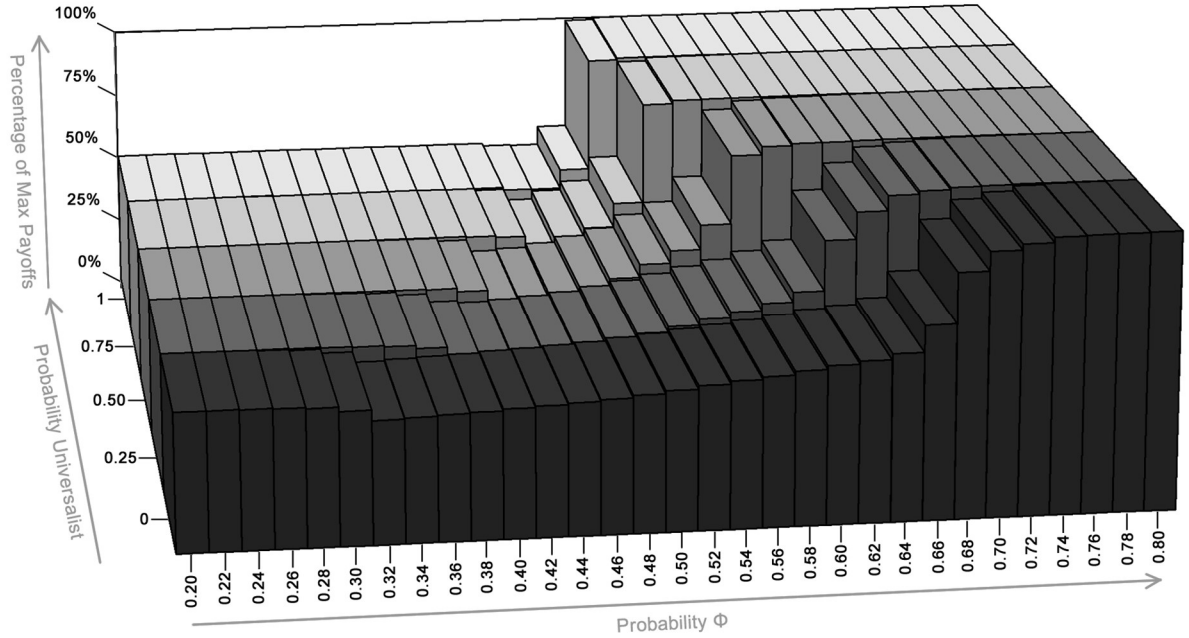

Figure 6: Percentage of maximum payoffs according to matrix for Stag Hunt for each simulation in Set 4 up to 250k timesteps, varying the starting probability of φ and the starting probability of UNI.

4.3 Implications

The upshot of our third and fourth sets of simulations is that, where coordination is important, it's better to err on the side of universalism. In effect, if you're a relativist you're "immune" from changing your first-order belief based on the wisdom of the crowd. You would think your minority belief is just as correct as the consensus belief. So coordination wouldn't occur in a population of die-hard relativists (relativists who won't change to universalism regardless of the consensus) who start with mixed first-order beliefs.

In human society, coordination is often essential, and the greater the importance of coordination, the greater the advantage of presuming universalism. Insofar as coordination problems are central to human social life, this is a significant practical advantage of a preference for universalism.

Note that the kind of universalism that facilitates coordination need not be an especially strong form of universalism. For instance, when learning a new norm, one needn't have any thoughts about whether the norm applies to rational aliens.[9] What matters instead is just that one expects that there is a single right answer about what the norm is *for people like us,* and this doesn't require thought or commitment regarding far-flung populations. For the gains of universalism to accrue, we only need to think of how the rules apply with those with whom we cooperate. So driving on the right-hand side is something that I care about when considering

[9] By contrast, Stanford's claim that "externalizing" moral demands facilitate cooperation relies on a stronger notion, according to which "we regard such demands as imposing unconditional obligations not only on ourselves, but also on any and all agents whatsoever, regardless of their preferences and desires" (2018, 1).

those who I might encounter on the road. Likewise, I may believe that there are professional norms that apply to anyone *in my profession*, but not to those outside of it. The model presented here is compatible with the idea that the "universe" of universalism extends only out to the set of people who interact with each other on areas where a particular belief is relevant. It is also compatible with something stronger still, but our model is silent on these distinctions. The important result is just that in all of these contexts, the presumption of universalism, even in the minimal form that we are considering, will facilitate coordination.

At the same time, at least for many norms and practices, the truth lies with relativism. The relativist will often be right to think that there is no single fact (this is surely true in many coordination games). Nonetheless, the benefits of coordination will often make it preferable for people to be naïve universalists.

One way of thinking about this more minimal conception of universalist beliefs is simply in terms of whether you think someone is making a mistake for failing to agree with the majority. Universalists maintain that someone who fails to agree with the majority is making a mistake; they might go on to judge that people who do not correct their mistakes are blameworthy. Relativists can't (and won't) make such judgments. Failure to coordinate, from the relativists' perspective, is just a manifestation of different views that happen to not line up. Some people learn about hunting hare, some people learn to hunt stag. That they might sometimes go hunting together is just something that happens on occasion.

5. Sometimes coordination has costs

While coordination is often held up as an abstract good in models of social dynamics, there are lots of instances of coordination that are costly in one way or another.[10] Coordination is better thought of as a social tool to achieve some end, rather than a good in and of itself. Just because hammers are useful when we need to drive nails into boards, that does not mean we want to use them to get rid of headaches. Likewise, while there are clearly situations in which coordination serves to solve particular problems, like safely driving on two-way streets, there are other situations in which it creates costs, such as if everyone in a city coordinated on eating at the same restaurant at the same time. Insofar as universalism facilitates coordination, it might produce undesirable outcomes in certain cases by creating coordination where none is wanted.

We will consider a few different types of situation to explain why coordination is not always desirable. First, as with the restaurant case, we are frequently better off if people choose to anti-coordinate. Restaurants can only serve so many people at once, and so we are better off if people come at different times to smooth demand given the fixed supply of seats. Likewise, and perhaps more canonically, we anti-coordinate when we make sure that what we contribute *complements* what others contribute in some joint endeavor. If I cook dinner, we're better off if you bring dessert instead of a second main course. A second, related type of situation occurs when for one reason or another we coordinate on a belief that produces suboptimal outcomes, such as when convenience generates a situation where only fast food restaurants can survive.

A third situation to consider is that we can often coordinate on morally *bad practices*. We might think here that there is some reason to coordinate, but the mechanism or practice by which

[10] Classic presentations of coordination games, such as Lewis (1969) or Schelling (1960), or cooperation games in Axelrod (1969) present coordination or cooperation to be desirable. However, there have long been conflictual coordination problems, such as Battle of the Sexes. More recently, there is a growing literature that uses coordination models to describe how undesirable or unjust states can be emerge and be stable. O'Connor (2019) offers an in-depth treatment, in part building from Skyrms (1996).

we do it is deeply unjust. We are all familiar with instances of societies that coordinate on unequal social hierarchies, or unfair divisions of household labor, or coordinating mechanisms that rely on arbitrary properties of the participants.[11] There may be some reason to coordinate in these cases (dividing tasks makes sense), but there are many coordination alternatives that would have been fairer. Coordination equilibria are nonetheless equilibria, which can make them very difficult to dislodge.

A final situation involves scenarios where we coordinate but no coordination was needed. In these situations there is no underlying problem that needed solving, and yet we have still found ourselves in coordination games that we unintentionally created from nothing. This is most easily seen in rules of etiquette or rules of dress, where we have settled on social rules where we just didn't particularly need any. The social world is rife with social rules that we follow, including many where there is no notable gain from having coordinated on those rules. Not only are the coordination outcomes arbitrary, but the very fact of coordination carries a cost, inhibiting individual freedom and autonomy. Once a situation becomes understood as one in which there is a social rule to follow, individuals are no longer free to just follow their preferences. These last two cases describe different aspects of what Mill called the tyranny of the majority, but as our model can show, that there is a majority to tyrannize might be an artifact of social belief dynamics rather than stable preferences. It is the tyranny of updating rules. While there has been notable work on the costs of unjust coordination, much less has been said on this last category.

5.1 Coordination game 3: Anti-coordination

The first and most obvious case where coordination would produce suboptimal outcomes overall is in anti-coordination games. For instance, if two of us are getting together to watch the game and we're each going to bring snacks, it would be suboptimal for us to both bring chips or to both bring dips. What we want to do is to anti-coordinate, one bringing chips and the other bringing dip, as shown in Figure 7.[12]

		Other(s)	
		Chips	**Dip**
Self	Chips	0, 0	1, 1
	Dip	1, 1	0, 0

Figure 7: Payoff matrix for an Anti-coordination game

To confirm the costs of a preference for universalism in such games, in our fifth set of simulations we used the same setup as for the previous games, changing only the payoff matrix

[11] O'Connor (2019) has an in-depth game-theoretic analysis of the gendered division of labor.

[12] Astute readers might note that anti-coordination games can be converted to coordination games if one introduces the notion of *roles* (say, the "chips" role and the "dips" role) and redefine coordinating as "doing what your role requires." While this is indeed a way of constructing a model, this is much more difficult to evolve without putting one's thumb on the scale. We would have to ask how the roles emerged and were coordinated on (most naturally with a correlated equilibrium concept) before we could get to how they are used in the newly-created coordination context. Universalism could help under these circumstances, but not without extra conceptual moves and coordination on those concepts, especially given that there are many possible role-pairs that could emerge in such a setting.

to match that shown in Figure 7. The results are shown in Figure 8. As expected, a preference for relativism now produces better outcomes overall than a preference for universalism. While this result was expected, it does help to highlight the problematic assumption that coordination is an unalloyed good. The simulations illustrate that coordination is valuable in certain social contexts but not others. A preference for universalism puts a thumb on the scales, encouraging coordination. And we have seen that this is often a good thing. But a *general* preference for universalism would put a thumb on the scales in *all* contexts, not merely those in which coordination is beneficial, and this can also lead to bad outcomes. Once we notice that there is important context-sensitivity to the value of universalism, its defense becomes much more complex.

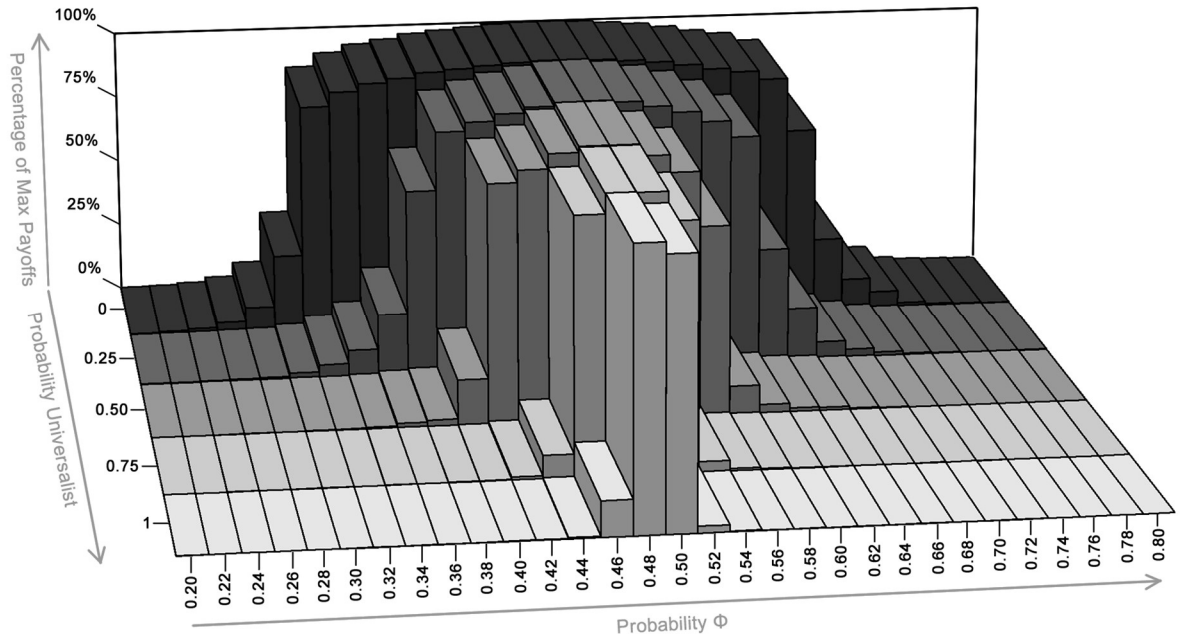

Figure 8: Percentage of maximum payoffs according to matrix for the Anti-coordination for each simulation in Set 5 up to 250k timesteps, varying the starting probability of φ and the starting probability of UNI.

5.2 Coordination game 4: Modified Stag Hunt

A second type of case where coordination can produce suboptimal outcomes involves situations where for one reason or another, if we coordinate we are likely to end up coordinating on a belief that produces a suboptimal outcome. Recall the Stag Hunt game discussed in Section 4.2. There, coordinating on one practice (hunting stags) produces better outcomes than coordinating on another (hunting hares). But we saw that a preference for universalism, coupled with a starting preference for hunting hares, can lead to coordination on a suboptimal outcome. For this game, the payoff matrix is such that either coordination point produces a better outcome than not coordinating. This is not always the case, however. Consider the variation on the Stag Hunt game shown in Figure 9. In the Modified Stag Hunt game, the greatest overall reward is still for coordinating on hunting stags, but not coordinating now produces a better overall outcome than

coordinating on hunting hares. One can imagine this occurring when the hunters would get in each other's way while individually hunting for hares.

		Other(s)	
		Stag	**Hare**
Self	Stag	2, 2	0, 3
	Hare	3, 0	1, 1

Figure 9: Payoff matrix for Modified Stag Hunt game

The payoff matrix for the Modified Stag Hunt game is the same as is standardly given for a prisoner's dilemma. In typical discussions of a prisoner's dilemma, the dilemma arises for the prisoners because they are acting strategically in a one-off game without communication. We are currently looking at a lower level mechanism than this, however: in our simulations individuals are not acting strategically, but simply following the meta-wisdom of the crowd across an extended series of interactions.[13] And in this case, when the population is likely to converge on the suboptimal outcome, overall rewards will be increased by slowing down the coordination process, as is seen in the results for our sixth set of simulations shown in Figure 10.

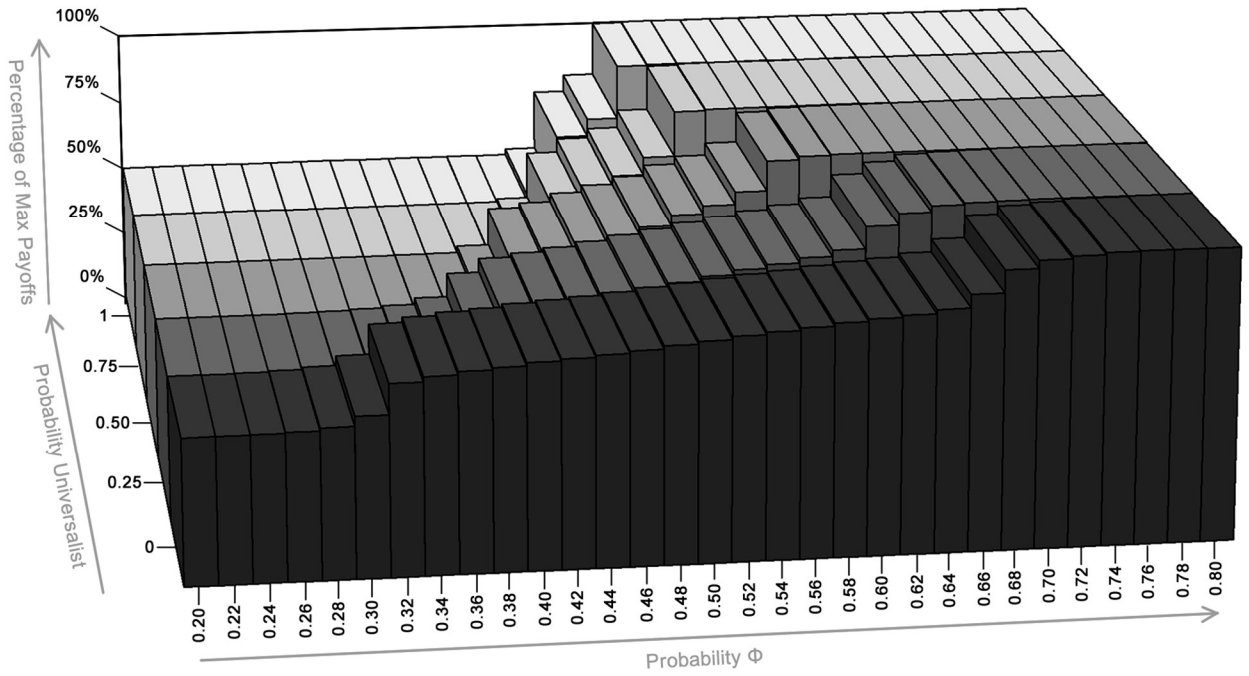

Figure 10: Percentage of maximum payoffs according to matrix for Modified Stag Hunt for each simulation in Set 6 up to 250k timesteps, varying the starting probability of φ and the starting probability of UNI.

[13] While it is beyond the scope of the present paper, the introduction of strategic interactions is planned for future work.

As expected, when the population is likely to converge on hunting hares—when the starting probability of holding φ is below 50%, as on the left half of the graph—we find that the overall payoffs are higher when the population has a preference for relativism. The preference for relativism slows down convergence on the suboptimal outcome, increasing the overall payoffs in the meantime. But it should be noted that this increase in overall payoffs is not distributed equally. In this case, the added benefits that accrue from slower convergence on hunting hares are obtained by those who hunt hares getting a greater reward when they pair up with someone who continues to hunt stags. For instance, when 60% of the population hunts hares, the average payoff for individuals who hunt hares will be 1.8 compared to 0.8 for individuals who hunt stags. Here this inequality is transient and disappears as the population converges on one belief. In our final set of simulations, however, we consider a case where coordination entrenches inequality.

5.3 Coordination game 5: Unequal Rewards

Our third case of undesirable coordination highlights the issue with priors that favor universalism just noted: in some cases they can entrench inequality. Consider a different kind of game—an *unequal* coordination game. In these games, each player benefits from coordination, but the coordination points confer advantages to one player over the other. For example, if two friends, a sculptor and musician, would enjoy spending the evening together but one prefers going to a concert while the other prefers going to a gallery, this generates an unequal coordination game, as shown in Figure 11.

		Sculptor	
		Concert	**Gallery**
Musician	Concert	2, 1	0, 0
	Gallery	0, 0	1, 2

Figure 11: Payoff matrix for Unequal Coordination game

To demonstrate the consequences of a preference for universalism in unequal coordination games, in our seventh set of simulations we used the same setup previously, but added in a division in the population that is relevant to the payoffs, randomly dividing the population into 500 "musicians" and 500 "sculptors." As seen in Figure 12a, a preference for universalism produces better *overall* payoffs. But these payoffs are highly unequal, as can be seen if we separately plot the total payoffs for musicians (as in Figure 12b) and the total payoff for Sculptors (as in Figure 12c). Since a preference for universalism promotes convergence, and whichever belief the population converges on differentially favors one group, while a preference for universalism produces greater overall payoffs, those payoffs tend to be less evenly distributed. This is clearly seen in Figure 12d, which plots inequity (the percentage of maximum possible difference in payoffs). Coordination here involves agreeing to a particular outcome that makes inequality permanent.[14] While this may be efficient in the sense that it maximizes overall payoffs, it is difficult to justify to those on the losing end, and there may be out-of-model

[14] In a more complex model, one could introduce the possibility of choosing a mixed strategy rather than pure strategies, or alternatively instead of relying on a Nash equilibrium solution concept, we could employ a correlated equilibrium concept, which could allow for efficient "taking turns" outcomes. However, we note that it is strikingly easy to find examples of unequal coordination games in the real world that have settled into a Nash solution with pure strategies. The gendered division of labor offers a large class of examples of this.

consequences, like feelings of resentment, loss of trust, and other social effects that can undermine future efforts at cooperation where it is more needed.

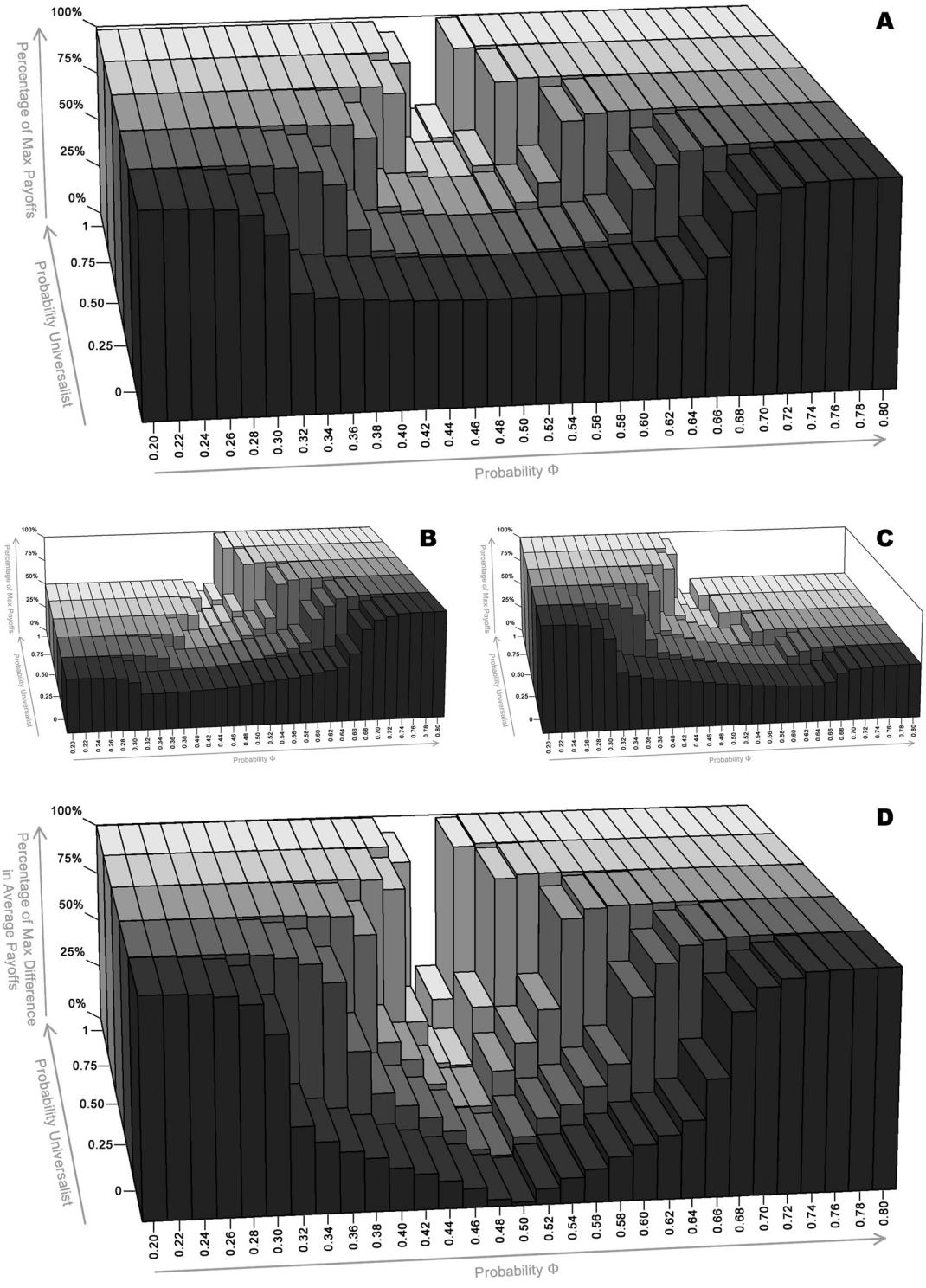

Figure 12: (A) Percentage of maximum payoffs according to matrix for Unequal Rewards for each simulation in Set 7 up to 250k timesteps, varying the starting probability of φ and the starting probability of UNI; (B) percentage of maximum payoffs

for “musicians”; (C) percentage of maximum payoffs for “sculptors”; (D) percentage of maximum difference in average payoffs between “musicians” and “sculptors.”

5.4 Implications

While there are benefits to a preference for universalism, our fifth, sixth, and seventh sets of simulations illustrate that there can also be clear costs. There are cases where the best thing is not to coordinate. There are cases where rigidifying a coordination point will rigidify inequality. And, of course, there cases where a preference for universalism would rigidify belief when there isn’t a coordination problem at all.[15] This threatens to be stifling for diversity of thought and comes at a cost to individual freedom and autonomy if these beliefs are linked to sets of allowable or forbidden actions.[16] A distinct source of human wellbeing is the capacity to choose one’s plan of life and to be able to endorse one’s own choices. A life that is mostly governed by arbitrary social rules that are enforced by one’s community renders our individual capacity for choice inert. There are two costs to this: most obviously, there are costs imposed on people with minority views, who have made considered judgments on what a good life looks like, but are not allowed to act on these judgements; but, there are also costs for those in the majority if they hold their views unreflectively—if they have taken up the view because an already-established social rule is present, then they do not get to exercise their autonomy or their capacity for judgment.

6. Making room for relativism

As we have just seen, there are some social situations in which coordination is not desired. In some domains, we prize our liberty to go our own way (see, e.g., Muldoon 2015). More disturbingly, rigid universalism would entail a tyranny of the majority and promote prejudice against minority opinion. As a result, even though there are considerable advantages for members of a population to have a prior in favor of universalism, there are also benefits to populations that embrace relativism for many norms and beliefs.

We have seen that preference for universalism in a population creates powerful dynamics that push toward coordination. In fact, in our simulations, a suitably strong preference for universalism leads to rapid convergence on a single belief even when the population starts with an even split in beliefs. And, as we have noted, this is often a good thing. But, as we’ve also seen, for some beliefs, convergence can be a bad thing. We conclude by showing that just as a strong preference for universalism in a population leads to convergence, a strong preference for relativism can maintain diversity. Diversity of belief can be valuable for a number of reasons, but we will highlight three: first, as we have discussed, there are many areas in life where coordination is not valuable, and indeed can be harmful. Second, it may be the case that coordination points were settled in the past, and payoffs to the various options adjusted as technology or society changes. So what was once an optimal choice is now inferior to some other option. Third, as Mill discusses in *On Liberty*, the presence of other (even wrong) beliefs gives us the opportunity to sharpen our arguments for our beliefs. Merely holding true beliefs does not help us understand why they are true, or how they might apply in a new context. Having an environment where there are a variety of competing beliefs provides the context for better understanding our own.

---

[15] This last class of problems with coordination – coordination where none is needed – is not something that we have modeled for this paper, as it requires a different modeling approach, but is planned for future work.

[16] See, for instance, Muldoon (2016) for an in-depth account.

6.1 Emphasizing relativism

In our second set of simulations, we showed that lowering the thresholds for changing second-order belief on the basis of consensus information favored universalism, promoting convergence. In our final two sets of simulations, we do the reverse: we increase the thresholds to favor more readily switching to relativism (higher *Y*) and to disfavor switching to universalism (higher *Z*). In our initial set of simulations, we set *Y* to 55% and Z to 95%. In the present simulations, we increased *Y* from 55% to 95% in one percentage point increments, and we increased *Z* from 95% to 99% in one percentage point increments. To illustrate how a suitably strong prior for relativism could maintain diversity, these simulations were run with a large majority of the population initially holding the same first-order belief. Specifically, in our eighth set of simulations, the probability that an individual would initially hold φ was set to 0.1, and in the ninth set of simulations φ was set to 0.2. And in both sets of simulations all individuals started out as relativists.

Finally, we added an extra wrinkle to the models, including personal preferences for first-order beliefs. As noted above, for some claims we would expect people to tend to have a personal preference: all else being equal, we would expect them to prefer that a given belief were true. For instance, I might have a preference for bowing as a greeting as it conveys more respect than other options. If I am a universalist about the issue, then I will adopt an opposing belief, that one should wave to greet people, if there is high consensus. (For instance, I might think that the majority is probably tracking some feature of the situation that is more important than my preference.) But that doesn't mean that my preference disappears. I still have the preference to drive on the left side of the road based on my vision, I just think that preference doesn't match with the correct answer about what one should do. Of course, for some claims, individuals might not have a personal preference at all.

To implement personal preferences, in our final sets of simulations each individual in the population was assigned a preferred first-order belief that was fixed across the duration of the simulation. This is in addition to the first-order belief that they happen to hold at a given time during the simulation. Each individual's preferred first-order belief was set to be the same as their first-order belief at the start of the simulation. In the previous simulations, when an individual changes from universalism to relativism on the basis of their consensus information, their first-order belief stays the same. With the addition of personal preferences, however, when an individual changes from universalism to relativism, if their first-order belief differs from their preferred belief, they will now change their first-order belief to their preferred belief. In other words, when individuals have personal preferences, relativists will always hold their preferred first-order beliefs. We expected that personal preferences would help maintain diversity in first-order beliefs in the population, decreasing the likelihood that the population would converge on one first-order belief. The reason is that as a population converges on one first-order belief, there will be some back-and-forth, with individuals shifting their beliefs based on the consensus information. Some of those individuals will shift from universalism to relativism. Without personal preferences, those individuals would not change their first-order belief, and thus would not alter the trend toward convergence. But with personal preferences, those who preferred the minority belief would switch back to it, pulling against the trend toward convergence.

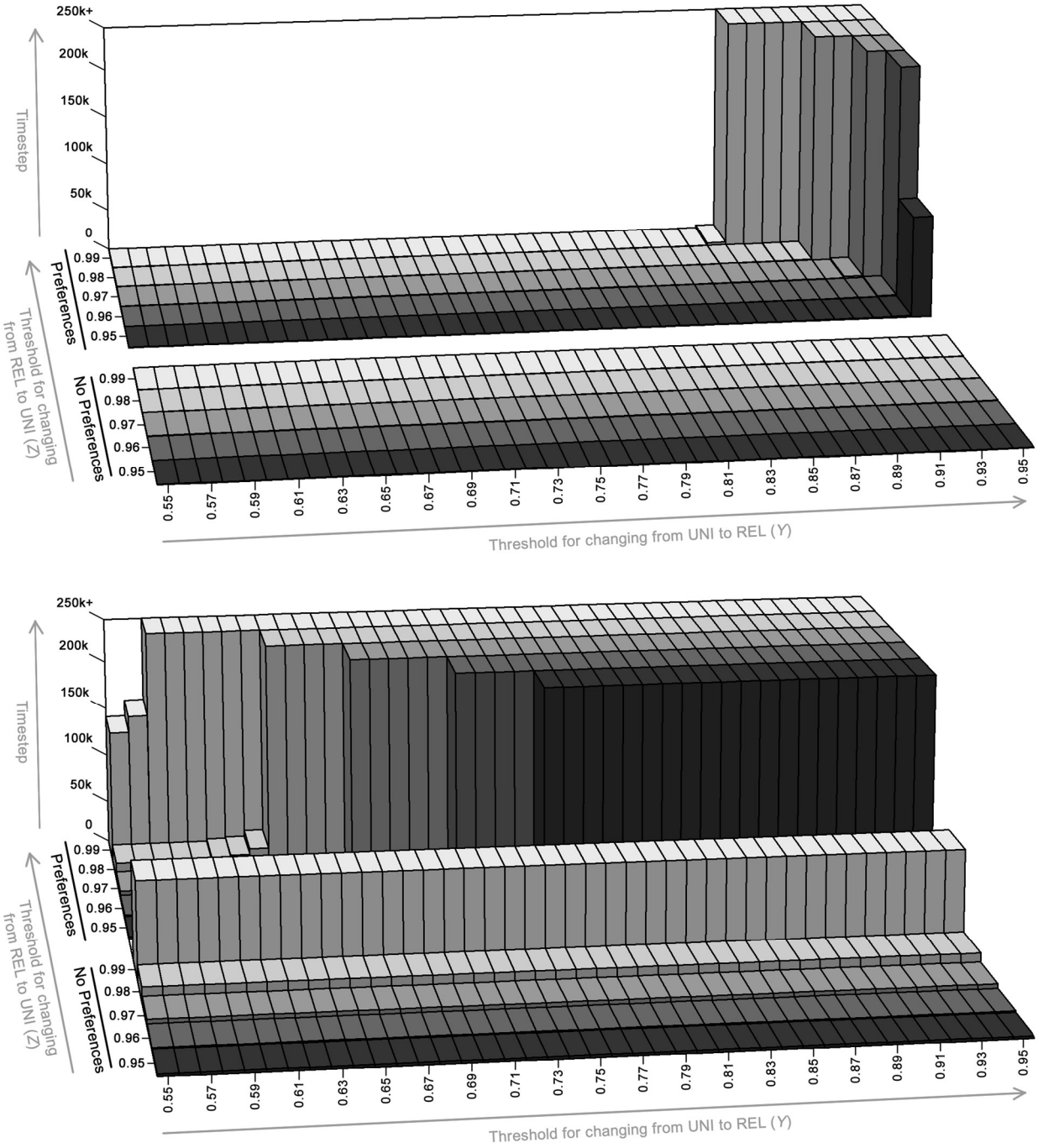

Figure 13: Time to convergence for each simulation in Set 8 (top; initial probability of φ set to .1) and Set 9 (bottom; initial probability of φ set to .2) up to 250k timesteps with all individuals starting out as relativists, varying the threshold for changing from UNI to REL (*Y*) and the threshold for changing from REL to UNI (*Z*). Each set of simulations was run both with and without personal preferences.

To examine both the impact of increasing the thresholds (Y, Z) to favor relativism and the introduction of personal preferences, we ran the present sets of simulations both with and without personal preferences. The results are shown in Figure 13. What we find is that even when there is just a small chance of an individual initially holding an alternative belief from the bulk of the population (.1 in Set 8, .2 in Set 9), this minority position is maintained if the members of the population have a personal preference for their starting beliefs and if there is a suitably strong prior in favor of relativism. Here we find that if individuals require evidence of a large majority belief before switching to universalism and allow a small enough level of minority opinion to lead them to switch to relativism, the population does not converge on the majority belief.

These final simulations help us capture something useful about the interplay between individual belief and our responsiveness to social evidence, especially when it comes to moral concerns. We may take morality, rather than just the social rules that comprise things like etiquette or social convention, to be an area where robust agreement is desirable. It facilitates coordination, can help us have more stable societies, and potentially can aid in promoting social trust. But moral progress is only possible when there is room for people to hold minority views. Civil rights movements start small and do the work of convincing the majority. Vegetarianism and veganism are minority views in the West, but may eventually be taken to be the correct view. And so on. Making room for relativism is not merely for the areas where we think we are beyond the scope of universalist beliefs, but also for those areas where we admit the possibility of error or the possibility of future change.

7. Conclusions and Future Work

In this essay we aimed to explore the ways in which an explicit model of first-order and second-order beliefs can help shape an agent's response to disagreement. Second-order beliefs about universalism and relativism clearly inform how we should understand social evidence about first-order beliefs, but likewise, we argue that this first-order evidence can inform what our second-order beliefs ought to be. Across a family of related models, we explore the dynamics of updating these two kinds of belief based on social evidence and the consequences for social coordination.

We have found that belief in universalism is a powerful tool for social convergence and coordination. This is a robust result, and our model demonstrates that the notion of universalism that generates this coordination result does not need to be all that demanding. Nothing like an externally imposed moral demand (e.g., Stanford 2018) or a universal truth of reason (Clarke 1728) is required. A preference for universalism across interactors in a community is sufficient. Where we differ from others in the universalist literature most strongly, however, is in challenging the idea that coordination is an unalloyed good. While there are important areas of our social lives where coordination is invaluable, it is also easy to find cases where coordination imposes real costs on individuals. Coordination can generate socially inferior outcomes (as in anti-coordination situations), it can generate efficient but deeply unfair outcomes (as in unequal coordination games), and as a more novel consideration it can impose rules where none were needed (in situations that simply don't present a coordination problem). In each of these cases, coordination imposed by a universalist tendency leads to normatively inferior outcomes.

In our final sets of simulations, we show that our basic model can also describe the conditions under which we might be able to avoid some of these outcomes: that is, when we might be able to preserve diversity of belief. As we argue, diversity of belief can be valuable not

only to try and prevent these immediate harms, but also because it can keep an epistemic community flexible. This allows them to better respond to changes in underlying conditions, or to make room for positive moral change by not snuffing out minority views too quickly.

While this more nuanced take on the value of universalism and the coordination it enables helps advance the discussion, there is more work to be done. In future work, we hope to explore mechanisms for eliciting conditions-responsive universalism. There is a very interesting discovery problem at the base of our metaethical positions: there is clearly a useful domain for universalism, and clearly a useful domain for relativism, but it is very much not clear where the borders are or how people come to agree about those borders. Our current models assume, as does the rest of the literature, that the world transparently provides people with a clear strategic situation. But an implication of our framework is that where universalists see a coordination problem, relativists might see a situation that provides no coordination benefit from matched strategies. Likewise, it is possible for one or the other to be wrong. A more robust model would allow agents to come to learn more about the world as they interact with each other, and have the possibility, though not the guarantee, of learning when to coordinate when it is valuable, and when to let people go their own way. Focusing on this problem of discovery can help us understand how people draw the boundaries they do between universalism and relativism, and perhaps inform how they should.

References

Ayars, A., & Nichols, S. (2020). Rational learners and metaethics: Universalism, relativism, and evidence from consensus. *Mind & Language*, *35*(1), 67-89.

Beebe, J. R., & Sackris, D. (2016). Moral objectivism across the lifespan. *Philosophical Psychology*, *29*(6), 912-929.

Boyd. R. & Richerson, P. J. (1985) *Culture and the Evolutionary Process.* Chicago: University of Chicago Press.

Clarke, S. (1728). *A Discourse concerning the Unchangeable Obligations of Natural Religion, and the Truth and Certainty of Christian Revelation*. London: James & John Knapton, 7th Edition.

Finlay, S. (2007). Four faces of moral realism. *Philosophy Compass*, 2(6), 820-849.

Goodwin, G. & Darley, J. (2008). The psychology of meta-ethics: Exploring objectivism. *Cognition, 106,* 1339-1366.

Goodwin, G. & Darley, J. (2012). Why are some moral beliefs perceived to be more objective than others? *Journal of Experimental Social Psychology, 48,* 250-256.

Harman, G. (1985). Is There a Single True Morality? In D. Copp and D. Zimmerman, *Morality, Reason and Truth: New Essays on the Foundations of Ethics*. Totowa, NJ: Rowman & Allanheld.

Mackie, J. (1977). *Ethics: Inventing Right and Wrong*. London: Penguin.

Muldoon, R. (2015) Expanding the Justificatory Framework of Mill's Experiments in Living. *Utilitas,* 27(2), 179-194.

Muldoon, R. (2016) *Social Contract Theory for a Diverse World: Beyond Tolerance*. Routledge.

Nichols, S. (2004). After objectivity: An empirical study of moral judgment. *Philosophical Psychology*, 17, 3-26.

O'Connor, C. (2019). *The Origins of Unfairness*. Oxford University Press.

Simpsonbeck, D., & Sytsma, J. (ms). Simulating metaethics: Consensus and the independence of moral beliefs. http://philsci-archive.pitt.edu/16461/

Skyrms, B. (1996). *Evolution of the Social Contract*. Cambridge University Press.

Stanford, P. K. (2018). The difference between ice cream and Nazis: Moral externalization and the evolution of human cooperation. *Behavioral and Brain Sciences*, *41*.

Wong, D. B. (2006). *Natural Moralities: A Defence of Pluralistic Relativism*. New York: Oxford University Press.

Wright, J. C., Grandjean, P. T., & McWhite, C. B. (2013). The meta-ethical grounding of our moral beliefs: Evidence for meta-ethical pluralism. *Philosophical Psychology*, 26(3), 336-361.

Wright, J., McWhite, C. & Grandjean, P. (2014). The cognitive mechanisms of intolerance: do our meta-ethical commitments matter? In T. Lombrozo, J. Knobe, & S. Nichols (eds.), *Oxford Studies in Experimental Philosophy, volume 1.* Oxford, UK: Oxford University Press.