HARVEST MOON
AN UNDERGRADUATE JOURNAL OF PHILOSOPHY

created by students at the
UNIVERSITY OF CALIFORNIA AT BERKELEY
2011 — VOLUME 8
If I have exhausted the justifications, I have reached bedrock and my spade is turned. Then I am inclined to say: “This is simply what I do.”

*Ludwig Wittgenstein*
This journal was made possible by generous grants from the UC Berkeley Philosophy Department and the ASUC. Our deepest thanks for their support.

Harvest Moon is a sponsored publication of the Associated Students of the University of California. The contents and opinions herein are those of the writers (those whose names appear before their papers) and not of the ASUC or University of California.
Contents

EDITOR’S NOTE ix

PAPERS:

Cooperation in the Presence of Diverse Ethical Theories: Dynamic Second-Order Strategies and Moral Convergence
by Justin Bruner 1

On the Key to Negative Judgments of Beauty in Kant’s Theory of Taste
by Matthew Clark 27

The Utility of Maximal Properties and the Possibility of Mereological Gunk Both as Sufficient Reasons for Preferring Mereological Universalism over Nihilism
by Charles Goldhaber 45

Counterfactuals and Expected Utility
by Eduard Meleshinsky 63

NOTES ON CONTRIBUTORS 81
Editor’s Note

I am pleased to present the eighth edition of the Harvest Moon: Berkeley’s Undergraduate Philosophy Journal. I think this volume testifies to the talent, rigor, and commitment of Berkeley’s philosophy students. With essays ranging in topic from aesthetics to game theory, these selections not only circumscribe both analytic and continental traditions but demonstrate careful exegesis as well as inspired and original thought. The papers contained in this journal represent some of the best philosophical work written by undergraduates in the department and I think they speak to the diversity of subject matter and style of philosophy which goes on here at Berkeley. The Harvest Moon, unlike any other undergraduate journal of its kind, is entirely written, edited, and published by Berkeley undergraduates. It is an important and long standing practice of our journal to not only allow our peers to see their work in print but to engage in philosophical dialogue with them through the editing process. As all good philosophers know, the best way to practice philosophy is through discussion and both editor and author grow throughout production of this journal.

Of course none of this would be possible without my editorial staff and especially my production editors. I would like to thank all who have contributed to the journal giving their time and sharing their love of philosophy. I am excited to pass the baton to a new and promising group of eager young philosophers. It waxes and wanes but Harvest Moon shines on!

Laura Davis
April 2011
JUSTIN BRUNER

Cooperation in the Presence of Diverse Ethical Theories

Dynamic Second-Order Strategies and Moral Convergence
1. Introduction

Evolutionary game theory has been employed by philosophers, biologists and social scientists to understand the emergence of such phenomena as cooperation and signaling. Although the emergence of cooperation and morality has been explored profusely by biologists, social scientists and philosophers alike, the discussion has focused primarily on how an evolutionary framework can assist us in understanding how moral norms can infiltrate and dominate a population. For instance, Alexander and Skyrms (1999) are primarily concerned with whether or not individuals will converge upon a particular strategy while on a lattice. This paper investigates how we converge on moral acts while all along holding various different ethical systems. The basic question is: can moral convergence occur in the presence of diverse moral reasons?

It is a truism that although all individuals hold certain basic moral facts as true, there exists a variety of distinct ethical systems employed to support the agreed-upon moral claims. For example, take the uncontroversial claim that murdering the innocent for sadistic pleasure is wrong. An individual may abide by this rule for a number of reasons. She could respect the command for it is God’s will. She could refrain from this behavior because it does not maximize aggregate utility. She could refrain from the behavior for it violates the formula of humanity. It appears that whether one is a divine command theorist, consequentialist, or deontologist, there exists a large set of moral acts that are uniformly condemned or supported.

This fact has been overlooked: individuals act, but act for various reasons. While current research has primarily focused upon explaining the emergence of moral uniformity—the notion that a moral act such as cooperation will be adopted by the majority of a population—there has been little to no work explaining how a population can sustain multiple ethical systems while attaining moral convergence.

While a significant amount of research has relied upon replicator dynamics, there exist a number of models representing individuals on a social network. For example, Skyrms and Alexander have investigated how individuals positioned on a lattice will interact, and whether co-
operation will emerge from said interactions. Other social structures such as circles have been studied and various other structures are beginning to be thoroughly scrutinized for the purposes of tracking strategies and cooperation over generations. Ellison (1993) observed that individuals playing a stag hunt with neighbors on a circle will result in hare hunting.

This paper will proceed in the following manner. Section 2 will set up the machinery and game theoretic tools necessary to explore this topic. Section 3 will flesh out and formalize what is meant by “ethical systems”—for us to investigate the phenomena we must have a means to incorporate it into our model. Section 4 will be the payoff. We will run through a number of simulations that Section 2 and Section 3 helped to set up and motivate. Section 5 will discuss the various limitations of the overall approach, in addition to attempting to analyze the results of Section 4.

2. The Stag Hunt and Evolutionary Games: An Overview

The stag hunt was first formulated by Rousseau in his *Discourse on the Origin and Basis of Inequality Among Men*. The details are as follows: a group of individuals are on a hunting expedition. There are hares and stags in the forest. Hares, while easy to catch, are small and provide little sustenance. Stags are large beasts that can feed a grown man for days on end. Unfortunately, stags are more difficult to hunt and thus cannot be captured by one man alone—a number of hunters must cooperate with each other to ensure that the stag is secured. While hares can definitely be caught by a lone individual, a stag requires a team of hunters. Although Rousseau under-describes the situation—lots of vital details are left out for us to fill in, such as how will the stag be split between the members of the party or by how much does the stag offset the meat from the hare—we can nonetheless take away a great deal from this scenario.

What is interesting about the stag hunt is the fact that the coordinating equilibrium (the state of the world in which all hunt stag) is superior to the hare equilibrium (the state of the world in which all hunt hare). Everyone would do better to coordinate on hunting stag. Yet, the game is not as simple as just playing the strategy that leads to

---

3 Alexander and Skyrms. 1999.
4 For work on social structures, see Ellison (1993). For an analysis of more complex and nuanced social structures, see Alexander (2007).
the optimal equilibrium. If one were aware that all of his fellow hunters were set on acquiring stag, it would then be reasonable to engage in the activity as well. Yet, because a lone stag hunter will reap no benefits from his toil, there is an inherent risk in deciding to hunt stag. For one to be motivated to hunt stag, one must believe a sufficient number of other hunters will hunt stag as well, and it is this contingency that plagues stag hunting with risk. The hare hunter faces no such problem. Although the equilibrium in which all hunt hare is sub-optimal, the hare hunter will always get the same payoff regardless of what those around him do. While the stag equilibrium may be optimal, the hare equilibrium is risk dominant. There is no risk involved in hare hunting, while the stag hunter could come back empty-handed if there are too few willing to cooperate with him.

This paper will employ a two-person version of the stag hunt. For stag hunting to be successful, both individuals will have to agree to hunt stag. If either turns to hare hunting the benefits of stag hunting are eradicated. Our payoff matrix takes the following form:

<table>
<thead>
<tr>
<th>Player 1/Player 2</th>
<th>Stag</th>
<th>Hare</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stag</td>
<td>3,3</td>
<td>0,2</td>
</tr>
<tr>
<td>Hare</td>
<td>2,0</td>
<td>2,2</td>
</tr>
</tbody>
</table>

Table 1: The two-person stag hunt.

The game gets more complicated when we expand this to an N-person game for which there exists a threshold of players, “k” (where k ≤ N), necessary to facilitate successful stag hunts. In this N-person game, as “k” increases in value, the risk of hunting stag increases as well and it is easy to see how a group of hunters could get stuck in the unfortunate sub-optimal equilibrium of hunting hare. However, things seem more optimistic in the case of the two player stag hunt. Given that one needs to rely only on a single additional person, it seems reasonable that we both may conclude that it is best to hunt stag.

Skyrms (2004) demonstrated that given the framework I will describe in the next paragraph, a population of individuals on a lattice playing the stag hunt game with their neighbors have on average a two-thirds chance of converging on stag hunting (meaning that all individuals in the population hunt stag). This social network setup has been very popular and numerous other researchers have manipulated the lattice or evolutionary dynamics in order to show cooperation arises
This paper will presuppose much of the framework developed by Skyrms (2004). For one, we will imagine a population of individuals situated on a lattice, where each can interact with her neighbors. Neighborhoods are defined as either von Neumann neighborhoods or Moore-8 neighborhoods. The former include neighborhoods that consist of individuals directly north, south, west and east of the individual in question. Moore-8 neighborhoods consist of the cells N, NE, NW, E, W, S, SE, and SW. For example, if we are using von Neumann neighborhoods, then each individual plays the stag hunt game with his neighbor to the north, south, east and west. Each individual has only one strategy (hunt hare or stag) and plays this strategy against all of his neighbors. In other words, an individual cannot customize his response to particular neighbors—if he hunts stag, he must hunt stag with everyone. In each round, all individuals interact with all of their neighbors by playing the stag hunt game together. This leaves everyone with a total payoff for a given round; i.e., the sum of the payoffs acquired from all four interactions with neighbors. For example, consider the following setup, with Alice on the left and Barbara on the right, and where S and H represent individuals on the lattice hunting stag and hare, respectively:

A:

S
H H S
H

B:

S
H S S
H

Figure 1: Two examples of a von Neumann neighborhood.

Alice and Barbara are represented by the H and S, respectively. Furthermore, notice that we are using a von Neumann neighborhood—Alice will only interact with neighbors directly adjacent to her. Examining Figure 1, Alice will get two utils from each neighbor for a total payoff of eight. Barbara, on the other hand, will get a payoff of zero from her neighbor to the left and below. She will, however, get three from each of the remaining two neighbors, making her total payoff six.

I will define a meta strategy, or second-order strategy, as a choice mechanism or function that determines whether one will hunt stag or

5Abramson and Kuperman (2001) investigate a population playing the prisoner’s dilemma on a number of randomly generated social structures. Santos, Pacheco, and Lenaerts (2005) utilize a social network in which individuals have more contacts than others.
hare in the following round. Skyrms (2004) introduces two possible choice mechanisms, imitation and best response. Second-order strategies determine what one’s first-order strategy will be at time \( t+1 \). As mentioned previously, second-order strategies can be conceived of as functions, and can take as their inputs nearly any fact about the model, from neighborhood demographics to histories of the strategies utilized. Let \( R \) be an input of the second-order function; \( R \) contains all the information that the function will take into account. A perfectly acceptable second-order strategy is:

\[
\text{For all } R, \; \text{SOS}(R) = \text{Stag}.
\]

Thus, an individual would always hunt stag no matter what. I label these strategies as second-order for they guide the more basic first-order strategies (i.e. whether one should hunt stag or hare).

Taking an example from Skyrms’ work, let us examine the best response second-order strategy. Skyrms defines the best response function as that which “chooses the action that is [a] best response when played against each of the actions your neighbors played last time.” Thus, the individual attempts to maximize utility given that the others will play the same way they did on the last round. With a lattice social network, the best response would be the following (where \( R \) is the number of stag hunters in an individual’s von Neumann neighborhood):

\[
\text{SOS}(R) = \begin{cases} 
\text{Stag if } R > 2 \\
\text{Hare if } R < 2
\end{cases}
\]

Another second-order strategy, imitation, demands that individuals adopt whatever first-order strategy the most successful individual within

\footnote{In the literature, these are referred to as “learning rules.” See Rory (2009).}
\footnote{Skyrms, p. 34}
\footnote{Note that this is nowhere near a true “best response” in the expected utility sense of the phrase. Best response in the limited sense used in this paper is analogous to static expectations in macroeconomics—individuals simply assume that what happened in the last period will flow over to the next. This of course is not an accurate representation of how real individuals behave, and is definitely not the optimal strategy.}
\footnote{If an individual is in a neighborhood which consists of less than three stag hunters, the payoff of himself being a stag hunter is six, while the payoff of him being a hare hunter is eight. If he is in a neighborhood which consists of three or more stag hunters, the payoff of stag hunting is at least nine, which is obviously better than hunting hare. This is the logic of the best response second-order strategy.}
Cooperation and Diverse Ethical Theories

their neighborhood is employing. Skyrms runs simulations of hunters with best response as well as simulations of imitators. We run these simulations and find similar results as Skyrms (2004). We add the nuance of allowing for second-order dynamics. In other words, in addition to first-order strategies, second-order strategies can be imitated by individuals as well. The following simple example illustrates this.

Individuals are placed on a circle where each individual interacts with his neighbor to his direct left and right (call this a flatlander neighborhood). Each individual is randomly assigned a second-order strategy of either best response or imitation and adjusts her first-order strategy accordingly. However, over time we allow the second-order strategies to change as well. Thus if individual A is in B’s flatlander neighborhood and notices that B’s second-order strategy is different from her own and B is doing better than her, it is permissible for A to jettison her current second-order strategy to adopt B’s meta strategy.

Using Matlab, a number of simulations were created to determine how this meta dynamic would influence stag and hare hunting within a population on a circle. The first two sets of simulations we will focus on are those in which only one second-order strategy is employed and our social structure will be that of a circle. Each individual was assigned a random strategy (hunt hare or stag), but all shared a common meta strategy (imitation or best response). We find the following statistics for best response and imitation: of the populations in which imitation is the second-order strategy on average about 24.37% of their agents ended up cooperating on stag hunting by the end of the simulation, while populations in which the second-order strategy is best response have, on average, no stag hunters. These findings were derived from 200 simulations of 100 individuals on a circle for 50 generations. In all of these simulations, the population had stabilized by the 50th generation.

Notice that best response almost always converges to hare hunting. The explanation for this is simple: the only situation in which it is a best response to hunt stag is the case in which you are surrounded by a stag hunter on both sides. Even if we have a near homogenous population of stag hunters, the existence of just a single hare hunter will

---

10If they themselves are the most successful, or even if they are tied for most successful, they retain their second-order strategy.

11A note on terminology: meta strategy and second-order strategy, throughout this paper, are used synonymously. Also, meta dynamics refer to dynamic second-order strategies—i.e. second-order strategies that change over time (dynamic, as opposed to static).

12Code for all simulations ran can be found at http://brunerthesisproject.wordpress.com/.
in time poison the entire population, resulting in uniform hare hunting. Consider the following in which each line represents the population during a single generation:

\[
T = 1 \ldots S \ S \ S \ S \ H \ S \ S \ S \ S \ S \ S \ S \ldots
\]
\[
T = 2 \ldots S \ S \ S \ H \ H \ H \ S \ S \ S \ S \ S \ldots
\]
\[
T = 3 \ldots S \ S \ H \ H \ H \ H \ S \ S \ S \ S \ S \ldots
\]

etc.

Figure 2: The spread of hare hunting on a circle.

The stag hunters bordering the lone hare hunter would do best to switch to hare (four utils > three utils). This creates a ripple effect, which results in hare hunters taking over the population. The only initial condition that would lead to stag hunters dominating the population is a circle of all stag hunters. Additionally, the only other distribution that could sustain any stag hunters is the following: \ldots S H S H S H S H S \ldots As the reader can infer, the next round will look like \ldots H S H S H S H S H \ldots and so on forever.

Imitation on a circle does not consistently lead to everyone hunting stag, but it does do a better job at fostering cooperation than our best responder simulation. In general, this is due to the fact that while a line of stag hunters bordered by a hare hunter in the best response population were doomed to be overtaken by hare hunters, an enclave of stag hunters are impervious to attack when imitation is the primary second-order strategy. A stag hunter on the edge of a cluster of stag hunters is bordered by both a stag hunter and a hare hunter. Although this border guard only gets a measly three utils, his neighboring stag hunter gets the benefit of having two stag hunters as neighbors, and thus gets a grand total of six utils, whereas the border hare hunter only gets four utils. Being an imitator, the border stag hunter continues to mimic his stag-hunting neighbor. Groupings of stag hunters are protected from invasive hare hunters. Nonetheless, the prospect of all individuals converging to the hunt-stag equilibrium in the imitation population are quite bleak. Hare hunters can never be persuaded to hunt stags in this simple model—any hare hunter bordering a stag hunter will always find his utility to be superior to that of his neighboring stag hunter, and thus a hare hunter will never think to imitate a stag hunter. Ergo, it is not surprising that the number of stag hunters is well below 50%—stag hunters cannot persuade any of their hare hunting brethren
to join them on the stag hunt, while hare hunters can surely entice stag hunters to go for hare!

I found similar bleak results on a two-dimensional lattice (using the von Neumann neighborhood) for best response, but much more hopeful findings for imitation. A simulation employing just best response meta strategies resulted in an end population of only hare hunters (20 of the 20 simulations ran resulted in all hare hunters after 50 generations with 10,000 individuals). On a lattice, all simulations utilizing the imitation meta strategy resulted in universal stag hunting (20 of 20 with 10,000 individuals for 50 generations). Next, 100 simulations of 400 individuals on a lattice was constructed to which individuals were randomly assigned both first-order and second-order strategies. Second-order strategies were permitted to change.\textsuperscript{13} We find that with meta dynamics (i.e. non static second-order strategies), on average, a whopping 88.64\% of a population converges to stag hunting. Only 25.76\% of the individuals were best responders while the remaining majority had imitation as their second-order strategy.

\section*{3. The Phenomena of Second-Order Strategies}

As mentioned above, evolutionary models have been previously utilized to demonstrate the possibility of a population naturally taking up moral and cooperative practices. However, it is questionable how much this proves, for many of the simulations only assume one meta strategy a population can utilize. As demonstrated in the above section, choice of meta strategy is vital to whether we attain social cooperation or not. Kirchkamp (1999) and more recently Moyano and Sanchez (2009) both permit meta strategies to change over time for a population on a lattice playing the prisoner’s dilemma. What is worth noting is that both Kirchkamp and Moyano utilize second-order strategies that seem to be derivations of imitation.\textsuperscript{14} Although they are allowing for meta strategies to change, this is of little interest to my project for the various meta strategies they employ are too narrow in scope to capture the variety of moral reasons and theories which motivate and guide action.

\textsuperscript{13}As of now, how these second-order strategies change over time is purposefully left vague. This will be thoroughly addressed in Section 3 and Section 4.

\textsuperscript{14}For example, the three meta strategies that Mayano and Sanchez employ are: (1) imitate the best in the neighborhood, (2) randomly select an individual and imitate if more successful than yourself and (3) assign probabilities of adopting a neighbors strategy based on their relative success.
What is novel about this paper is that I attempt to employ a number of very distinct second-order strategies that correspond to various ethical theories. People have a range of ethical commitments and moral reasons, and these in turn dictate their behavior. The question that has been glossed over by previous investigations is: can we attain an equilibrium in which stag hunting, a strategy that corresponds to the action that is socially cooperative, is taken up by a population in the face of a plurality of ethical theories? The importance of the stag hunt game is to formally represent social cooperation in a population. By tracking the frequency of stag hunters in a population, we can get a sense of how cooperative this society is.\(^{15}\)

It is easy to imagine that individuals actually do employ second-order strategies in decision-making. Which meta strategy they use, however, is difficult to predict. The assumption that all individuals in a particular population will employ the exact same second-order strategy is not at all intuitive or realistic. All of us are not opportunists completely obsessed with determining which strategy we should use to best take advantage of our current position in the population (like the best responder). Some individuals would rather look to the most prosperous in their vicinity and mimic their traits in the hopes of attaining similar levels of greatness. Both meta strategies seem completely plausible ways of leading one’s life. Additionally, it seems completely reasonable that an individual could change her second-order strategies sometime in her life. An individual switching from an imitation to best response second-order strategy may have this to say, “I have constantly been doing what others more successful than me do—perhaps it is time I try to figure things out myself for once!”

I argue that this more flexible use of second-order strategies is good for a number of reasons. First, meta dynamics more accurately capture what I take to be the state of the human condition. We are not simply creatures who decide between stag and hare hunting. We have deeper ethical commitments (whether we are aware of them or not). In other words, what action we take is not primary. What is primary is our commitment to an ethical theory, which in turn determines whether we behave in a cooperative manner or not. Furthermore, to deny that people cannot update their second-order strategy is to deny them the ability to significantly change and grow.

Second, the changing of second-order strategies is descriptively accurate—\(^{15}\)Rory (2009) sets up a simulation in which the stag hunt is played with evolving meta strategies. Yet once again, the meta strategies he utilizes are not representative of the plurality of moral theories and theories of motivation we find in the world.
it is a phenomenon we actually observe in the world. For example, individuals convert from selfish opportunists to selfless Christians all of the time. They do so for a number of reasons, of course. The main reason we are going to give for such a conversion is that they observe others around them prospering more than they themselves are. If I am the lone atheist in a community of Christians who are all thriving much more than I am, it seems reasonable to believe that there is a high chance that I will eventually convert. This switch is not a calculative or even a deliberate act but occurs, nonetheless. For the purpose of this paper, imitation will be my meta-meta strategy (i.e., it will determine how second-order strategies change). We could of course propose that individuals have a plurality of third-order strategies, and observe what occurs with this level of complexity, but for the purpose of this paper, we will stop at individuals possessing second-order strategies with a third-order strategy of imitation.

I will identify eight basic second-order strategies I will utilize: Aristotelian, Kant, Best Response, Martyr, Psychopath, Imitation, Emotional Deliberator, and Norm Follower. Each attempts to capture a particular intuitive ethical theory. It should be noted that these eight obviously do not exhaust the realm of normative ethics—there are many other ethical theories one could attempt to model. However, for simplicity’s sake, we will concern ourselves only with these eight. In section 5.2, I will look at why it is reasonable, from a modeling perspective, to presuppose that these eight second-order strategies are the primary second-order strategies employed by individuals. Until then I will take it on faith that all individuals can only be represented with one of the eight second-order strategies.

Each second-order strategy is selected because it mimics how we expect actual individuals to go about making moral decisions. The easiest to see is the Martyr. The Martyr is a strategy that encourages people to always “do the right thing.” In the case of the stag hunt, Martyrs will always hunt stag despite what their neighbors are engaged in or, for that matter, any other facts about the world. A population of all Martyrs will all do very well for themselves, while a lone Martyr may not do well amongst less cooperative individuals. Psychopaths use another intuitive, simple strategy. Like their name suggests, they are anti-social and will always hunt hare, no matter what their neighbors engage in. Both of these strategies have in common their independence—the output of each of these second-order strategies is not contingent on the state of the world. Of course, the majority of moral theories are not so
clear-cut. For example, it is easy to imagine an individual whose notion of good and evil stems from what the majority of individuals within his vicinity hold to be true. If all are engaged in cooperation, then he will cooperate; if the majority hunts hare, he will do likewise. This Norm Follower will follow the majority. Of course, there are multiple ways to spell this out. An individual could survey an entire community to determine which strategy is most popular. An individual could restrict his view to a neighborhood of individuals close to him on the lattice. Likewise, how an individual picks a strategy is also a subtle affair—an individual could simply pick the strategy that is employed by a majority of the population, or he could adopt a particular strategy with the frequency of the probability it is found in the population. I will adopt the former method and restrain my view to the von Neumann neighborhood. Thus, individuals will assess their four neighbors and pick the strategy that is done by the majority. In cases of ties, a coin will be flipped to determine the proper course of action.

For the next two meta strategies, I look to Aristotle and Kant. Aristotle’s writings on ethics are extensive, but for the purposes of my second-order strategy, I will attempt to condense a facet of his ethics to a few lines. Ethics for Aristotle is not merely a theoretical endeavor. Developing good moral character requires practice. Taking the correct action consists of not only realizing what is right, but acquiring the propensity to continually engage in that behavior. “Virtue of character,” Aristotle claims “results from habit […] virtue comes about, not by a process of nature, but by habituation.”\textsuperscript{16} While the typical individual who attempts to cooperate may do so with some difficulty, cooperation is easier for the Aristotelian—he has practiced doing the right thing! Although it may usually take three stag hunting neighbors to entice the best responder to hunt stag, Aristotle will hunt stag if one or more neighbors are engaging in it. Cooperating is easier for him, and thus he requires less of an incentive to do so.\textsuperscript{17}

Kant, on the other hand, characterizes most moral decisions as at their core possessing a deep conflict between the “moral” action and the action the individual desires. In essence, people realize that the moral thing to do is cooperate, but find it difficult to do so for it goes against one’s natural inclinations (to not cooperate). Thus, for example, a Kantian hare hunter would require much more incentive than the usual hunter would to switch to stag hunting. In the model, Kant would

\textsuperscript{16}Aristotle, p.33.

\textsuperscript{17}It should be noted that modern virtue ethics is also ascribed to Aristotle. In this sense our imitation second-order strategy is also Aristotelian in flavor.
require all four neighbors to be engaged in stag hunting before he will switch from hunting hare.

Last is a second-order strategy I identify as the “emotional deliberator.” It does not correspond to a traditional ethical theory but it is quite intuitive to imagine a number of individuals having such a second-order strategy. In short, individuals become “attached” to their particular first-order strategy. To entice an emotional deliberator who is currently hunting stag to hunt hare will require only one or less of her neighbors to be hunting stag. To entice an emotional deliberator currently hunting hare to hunt stag would require all of her neighbors to hunt stag. This seems intuitively plausible. It is typical for individuals to form positive affects toward their behaviors. For example, an individual who is arbitrarily assigned to be a stag hunter may eventually come to identify with cooperation and thus cooperate even when it is against his own self-interest. A similar story can be told of a hare hunter who after hunting hare for a long length of time may believe there is some virtue in self-reliance and independence (i.e., not having to worry about others and what they do when hunting). As a result, it will take more than what is typically necessary to entice the rogue hunter to engage in stag hunting.

Now that we have established the basic meta strategies we will be utilizing, we discuss the individual dynamics these meta strategies take on a lattice.

### Table 2: List of eight intuitive second-order strategies.

<table>
<thead>
<tr>
<th>2nd-order Strategy</th>
<th>1st-order Strategy Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Martyr</td>
<td>Always Stag</td>
</tr>
<tr>
<td>Psycho</td>
<td>Always Hare</td>
</tr>
<tr>
<td>Aristotle</td>
<td>Stag if at least one neighbor is Stag; Otherwise, Hare</td>
</tr>
<tr>
<td>Best Response</td>
<td>Stag, if at least three neighbors are Stag; Otherwise, Hare</td>
</tr>
<tr>
<td>Imitator</td>
<td>Identity subset of neighbors doing best and: If self, then continue current strategy; If neighbor A, then adopt A’s 1st-order strategy; If multiple neighbors, then flip coin—adopt strategy</td>
</tr>
<tr>
<td>Emotional Deliberator</td>
<td>Stag, if Stag and at least one neighbor is Stag; or if Hare and all four neighbors are Stag Hare, if Hare and at least one neighbor is Hare; or if Stag and all neighbors are Hare</td>
</tr>
<tr>
<td>Kant</td>
<td>Stag, if Stag and two or more neighbors are Stag; or if Hare and all neighbors are Stag Hare, if Stag and have less than two neighbors that are Stag; or if Hare and not all neighbors are Stag</td>
</tr>
<tr>
<td>Norm Follower</td>
<td>Stag, if at least three neighbors are Stag Hare, if at least three neighbors are Hare Flip a coin, if two neighbors are Stag</td>
</tr>
</tbody>
</table>
Aristotelians are the easiest to start with, and we run a number of simulations of them on a lattice. Given how generous they are, it comes as no surprise that all simulations of a population of Aristotelians converge to uniform stag hunting. The Aristotelian second-order strategy is simple enough for it to be analyzed analytically. For example, we can easily see how within a sea of hare hunters, two neighboring stag hunters can easily turn the whole entire population into stag hunters. An Aristotelean individual hunts stag if at least one of his neighbors is a stag hunter. Thus the two original stag hunters will remain stag hunters and all of their surrounding neighbors will likewise hunt stag. This will spread until the whole population is hunting stag. Thus, the only structure that can sustain hare hunting of any sort are the following: all hare hunting, or the following structure:

\[
\begin{array}{cccccc}
H & S & H & S & H & S \\
S & H & S & H & S & H \\
H & S & H & S & H & S \\
\end{array}
\]

Figure 3: Aristoteleans on a lattice.

In this case, each individual will alternate between stag and hare hunting:

\[
\begin{array}{cccccccccc}
H & S & H & S & H & S & H & S & H & S \\
S & H & S & H & S & H & S & H & S & H \\
S & H & S & H & S & H & S & H & S & H \\
H & S & H & S & H & S & H & S & H & S \\
\end{array}
\]

Figure 4: A possible structure which can sustain hare hunting in an Aristotelian population.

This blinking light formation and a pure hare hunting population are the only populations that can sustain hare hunting among Aristotelians.

Kantians favor hare hunting more. Twenty simulations of populations with Kantian second-order strategies of 10,000 individuals going for 50 generations, and we find on average about 26.03% of the individuals hunt stag. Additionally, stag hunting tends to be clustered—there
exists rectangular “islands” of stag hunters in a sea of hare hunters. Each rectangular region is self-sustaining, for along the edge of the rectangle all stag hunters are in contact with at least two other stag hunters (the minimum needed for a Kantian to continue to hunt stag).

Additionally, regions of stag hunting can be connected by a chain of stag hunters. Each member in the chain is able to sustain his stag hunting for he is connected to two other stag hunters. The final population of Kantians will be a network of islands of stag hunters.

\[
\begin{array}{cccccccccc}
S & S & S & H & H & H & H & H & H & H \\
S & S & S & S & S & S & S & S & H \\
H & H & H & H & H & H & S & S & H \\
H & H & H & H & H & H & H & H & H \\
\end{array}
\]

Figure 5: A stable structure—stag hunters are oriented such that they will not be converted.

Emotional deliberators come to very similar results as Kant. A simulation of 20 populations of 10,000 individuals reveal that on average a population has about 27.68% stag hunters after 50 generations. In all of the simulations, we found that by the fiftieth generation the population had become stable. Like the population of Kantians, emotional deliberators tend to form a network of connected rectangular communities.

Finally, Norm Follower tended to have different patterns altogether. First, Norm Followers tended to not stabilize nearly as soon as the majority of the other meta strategies did. This is mainly due to the fact that the initial population was randomly assigned, and there was no obvious norm to follow, for one-half of the population was hunting stag and the other half was hunting hare. Running 20 simulations of 100 individuals for 200 generations we find that on average Norm Followers neither converge to stag or hare hunting, but tend to vacillate between the two.

4. Dynamic Meta Strategies

We now turn to positioning all of these various meta strategies together on a lattice with a third-order strategy of imitation. While populations of best responders and imitators were previously compared in Section 2, we will now run a simulation in which a population will consist of
eight distinct second-order strategies.

Allowing second-order strategies to change over time may sound reasonable, but nonetheless comes with a number of new problems. One obvious detail we would have to spell out is the strategy that guides the second-order strategy changes; i.e., the third-order strategy or the meta-meta strategy. It is possible to imagine, although very unlikely, that some keen individuals may use a best response third-order strategy. After observing the first-order and second-order strategies of their neighbors, they decide to choose a second-order strategy that will be optimal. This deliberative and explicitly conscious selection of the second-order strategy seems quite contrary to the intuitive definition of second-order strategy I gave earlier. Prima facie, imitation is the most plausible third-order strategy (although I will not go as far to rule out all others), and for this reason I will assume that individuals all employ imitation as their third-order strategy.\(^{18}\)

However, by no means is this argument truly closed. There exists an array of reasonable third-order strategies that may be modeled. For example, an individual could only change his second-order strategy when he is doing sufficiently bad relative to his neighbors. This individual would fail to change, however, when the difference in outcomes between him and his more successful neighbors are beneath a particular threshold. For the sake of simplicity, this paper will not engage in these more nuanced approaches to third-order strategies.

A second major problem with meta dynamics is setting the frequency at which individuals decide to update their second-order strategies. If these second-order strategies are to correspond to anything remotely similar to what we actually have in reality, they need to be semi-permanent strategies, which are not updated perennially. A naïve approach would be to set a constant frequency of turnover of second-order strategies for all individuals. Every five generations, individuals imitate the second-order strategy of their most successful neighbor. A slightly more sophisticated version of this would be to realize that individuals have varying levels of plasticity of their second-order strategies—individuals will receive different turnover frequencies. An even fuller approach would be to not only give each individual a different turnover frequency, but also set the starting of their second-order

\(^{18}\)The third-order strategy of imitation requires that individuals imitate a neighbor’s first-order and second-order strategies if said neighbor has greater fitness than the individual. In cases in which multiple neighbors have equal fitness which is nonetheless greater than the individual’s, a coin flip will decide who will be imitated.
frequency at random places.

This paper will for the most part utilize the rudimentary “naïve” approach. As evidenced by the table below (Table 3), there is actually very little difference in outcomes between the naïve, sophisticated and fully sophisticated approaches.

We will set up the major simulation as such: individuals will be randomly assigned first-order and second-order strategies. We run a simulation with both naïve and sophisticated meta dynamics: 50 simulations of each naïve, sophisticated and fully sophisticated populations were run (each simulation was run for 100 generations and had 400 individuals on a lattice). The results are summarized in the following table:

<table>
<thead>
<tr>
<th></th>
<th>Martyr</th>
<th>Aristo</th>
<th>Emo</th>
<th>Kant</th>
<th>Norm</th>
<th>Psycho</th>
<th>BR</th>
<th>IM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve</td>
<td>27.1%</td>
<td>18.4%</td>
<td>12.4%</td>
<td>8.0%</td>
<td>10.0%</td>
<td>0.0%</td>
<td>8.3%</td>
<td>15.8%</td>
</tr>
<tr>
<td>Soph</td>
<td>23.7%</td>
<td>19.8%</td>
<td>10.5%</td>
<td>8.1%</td>
<td>12.2%</td>
<td>0.02%</td>
<td>8.3%</td>
<td>18.0%</td>
</tr>
<tr>
<td>Fully Soph</td>
<td>26.5%</td>
<td>17.7%</td>
<td>10.1%</td>
<td>8.6%</td>
<td>11.3%</td>
<td>0.0%</td>
<td>8.9%</td>
<td>16.0%</td>
</tr>
</tbody>
</table>

Table 3: The final results of the three simulations, each consisting of eight meta strategies.

What we find is that, in all cases (sophisticated and naïve), a convergence to stag hunting occurs while maintaining a fairly diverse set of second-order strategies. Once again, note how we treat the third-order strategies does not affect this outcome—both naïve and sophisticated populations maintain a plurality of second-order strategies while the majority hunts stag. In fact, the three outcomes have very similar distributions of second-order strategies.

Thus it appears that our original inclination was correct—it is possible for individuals in a society to all agree to certain moral facts, yet abide by them for a variety of reasons. What is interesting is which of the initial second-order strategies thrive and which die out. The first thing to notice is that the Psychopath population decreases to zero. In fact, just about all of the “mean” strategies in the population are not very successful—best response, which uses a cold hearted calculation to determine when it will cooperate, does badly overall. Not surprisingly, Kant, which is even worse at giving people the benefit of the doubt, is even less successful than best response.
The failure of mean strategies is accompanied by the predictable success of friendly strategies. Most notably is the success of the Martyr meta strategy, which acquires a large chunk of the population. This is followed closely by Aristotle, whose strategy is only slightly less nice than Martyr (Aristotle does hare hunt when all his neighbors are hare hunting). Rather neutral meta strategies such as emotional deliberator and Norm Follower tend to be fairly stable, neither losing nor gaining much share of the population. Imitation, however, which seems as though it should fall into this ‘neutral’ meta strategy category as well, does well, probably reflecting not so much the greatness of the strategy, but the fact that there are many nice strategies hunting stag.

Another important issue worth noting is how the structure of the interaction influences outcomes. For example, if we change our simulation to utilize a different social network, say, a circle, we get much different results. First off, the population does not converge toward stag hunting. In fact, in a simulation in which individuals were placed on a circle and second-order strategies were allowed to change, only 23% of the population settles on stag hunting. Interestingly, when we run these meta strategies on a circle but prevent individuals from imitating other meta strategies, cooperation is assisted—a little over 50% of the individuals end up hunting stag. These two findings seem to suggest that what is motivating the lack of stag hunting on the circle simulation is the increase of “mean” meta strategies in the population.

Figure 6: Simulation with meta dynamics, frequency of second-strat update, two generations.
When meta dynamics are permitted, the number of mean strategies in the population increases and the result is less stag hunting. Looking at the evidence from our simulation of the meta dynamics on a circle, we find that mean strategies do increase, over one-third of the individuals on the circle were either best responders or Psychopaths. This is in drastic contrast to our simulation on a lattice, which showed all of the Psychopaths had vanished in a matter of 20 generations.

The explanation for this is that mean strategies have an inherent advantage on less-connected social structures. This can be easily demonstrated. An individual playing Psycho on a lattice will consistently reap a hare (utility = 2) from each of his four neighbors, bringing his total fitness for that round to 8. However, his stag-hunting neighbor could potentially do much better than he could if he has either three or four stag-hunting neighbors. This is not the case in less connected social structure. On the circle, where individuals are connected to only two others, a Psycho hare hunter would never be enticed by his neighbor to switch second-order strategies. This is due to the fact that his neighbor, whether a stag or hare hunter, can never do better than the Psycho can. A neighboring hare hunter will obviously always do just as well as the Psycho and the stag hunter will always do worse, for one of two of his neighbors are Psychos! In other words, the more connected a social structure is, the more plausible it is that cooperators would thrive simply because there are more people to cooperate with! Hence the answer to our original question of moral convergence in the face of ethical plurality is contingent on the social structure and particular network we are dealing with.

5. Further Issues

Of course, regarding the main finding of Section 4, there is still much more to do and discuss. First of all, what is motivating this phenomena? Are particular meta strategies responsible for moral convergence in the face of diverse second-order strategies? This will be presented in detail in Section 5.1. Second, it appears we are not proving much if we are starting with second-order strategies, which we find in the world. If we start with intuitive second-order strategies, it seems as though it is no surprise that diverse second-order strategies can be maintained while stag hunting dominates a population. This is a form of reverse engineering, which we would want to avoid. To circumvent such flimsy
logic we must show how these second-order strategies come to dominate a population in the first place. This will be the task of Section 5.2.

5.1 The Motivation of the Phenomenon

To investigate what drives the majority of individuals to take up the socially cooperative act all the while sustaining a plurality of ethical theories within a population, we examine simulations in which only a select sub-set of the second-order strategies are present. By isolating certain meta strategies we can better study their effects on the population at large. Six different types of simulations are examined. Each simulation takes the following basic form. Three meta strategies are present: Martyr, Psycho, and an additional strategy from our remaining six meta strategies. Each simulation is designed to determine what the effect of a strategy, X, is with both second-order strategies Martyr and Psycho. In this way we can come to learn more about what is driving the phenomena we uncovered in Section 4.

The overwhelming trend in all six of our simulation is that Martyr took over a majority of the population. Among the six types of simulations ran, all had at least 50% of the population as Martyrs by the 50th generation. Additionally, almost all simulations had no Psychos present by the 50th generation. Simulations were split into two camps—some ended up where the Martyr and third meta strategy increased over time as the Psycho decreased, while others were characterized by an increase in the Martyr population at the expense of the other two meta strategies. All simulations shared the drastic decrease in the Psycho strategy. Below are two of the findings that illustrate the possible turns our simulations took:
It is worth noting that the strategy X usually did not decrease as rapidly or as much as the Psycho. This can be attributed to the fact that only the Psycho refuses to cooperate when surrounded by stag hunters. Friendly strategies, which cooperate more often, do significantly better than those that do not, although they typically do not do nearly as well as Martyrs do.

Another commonality in the simulations was the shifting of second-order strategies in the population within the first ten generations or so. These effects are seen in our main simulation in Section 4 but are more pronounced in these simpler simulations. What happened in four of

Figure 7: Psycho and Martyr with BR and Aristotelian, respectively.
the six types of simulations was an initial increase in the population of strategy X, accompanied by the decrease of Martyrs in the population. This trend was very temporary, lasting for four to five generations, at which point the Martyrs rebounded, increasing dramatically while the X strategy either slowly began to decline (as the case with the BR above) or continued to increase (albeit at a much slower rate than initially).

The explanation for this is quite intuitive. The X meta strategy in the simulation does well initially for it neither always hunts stag nor hare. Thus, it capitalizes on the initial heterogeneous nature of the population. Psychos will always do poorly (eight points) while Martyrs have a relatively small chance of doing better than Psychos. To get more than eight points they need at least three neighboring stag hunters. The probability of being situated on the first round next to three stag hunters is $\frac{5}{16}$, barely above a quarter. The expected utility a Martyr gets on his first round is 6 points, two below the expected utility of a Psycho. Thus, it is very intuitive why Martyrs drop off initially.

The following table documents this effect.

<table>
<thead>
<tr>
<th>Combo/Gen</th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>15</th>
<th>25</th>
<th>35</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Martyr</td>
<td>0.3266</td>
<td>0.2751</td>
<td>0.3186</td>
<td>0.369</td>
<td>0.5233</td>
<td>0.6507</td>
<td>0.7398</td>
<td>0.7964</td>
</tr>
<tr>
<td>Psycho</td>
<td>0.3314</td>
<td>0.3285</td>
<td>0.2886</td>
<td>0.2605</td>
<td>0.1681</td>
<td>0.0935</td>
<td>0.0435</td>
<td>0.0089</td>
</tr>
<tr>
<td>Best Response</td>
<td>0.342</td>
<td>0.3964</td>
<td>0.3928</td>
<td>0.3705</td>
<td>0.3086</td>
<td>0.2558</td>
<td>0.2167</td>
<td>0.1947</td>
</tr>
<tr>
<td>Martyr</td>
<td>0.333</td>
<td>0.315</td>
<td>0.3605</td>
<td>0.3885</td>
<td>0.481</td>
<td>0.492</td>
<td>0.49</td>
<td>0.493</td>
</tr>
<tr>
<td>Psycho</td>
<td>0.337</td>
<td>0.3235</td>
<td>0.2825</td>
<td>0.248</td>
<td>0.161</td>
<td>0.1405</td>
<td>0.1405</td>
<td>0.1405</td>
</tr>
<tr>
<td>Imitator</td>
<td>0.33</td>
<td>0.3615</td>
<td>0.357</td>
<td>0.3635</td>
<td>0.358</td>
<td>0.3675</td>
<td>0.3695</td>
<td>0.3665</td>
</tr>
<tr>
<td>Martyr</td>
<td>0.345</td>
<td>0.368</td>
<td>0.499</td>
<td>0.58</td>
<td>0.678</td>
<td>0.679</td>
<td>0.679</td>
<td>0.679</td>
</tr>
<tr>
<td>Psycho</td>
<td>0.308</td>
<td>0.255</td>
<td>0.169</td>
<td>0.101</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Norm Follower</td>
<td>0.347</td>
<td>0.359</td>
<td>0.332</td>
<td>0.319</td>
<td>0.321</td>
<td>0.321</td>
<td>0.321</td>
<td>0.321</td>
</tr>
<tr>
<td>Martyr</td>
<td>0.3375</td>
<td>0.2938</td>
<td>0.3406</td>
<td>0.3857</td>
<td>0.5303</td>
<td>0.6705</td>
<td>0.7239</td>
<td>0.725</td>
</tr>
<tr>
<td>Psycho</td>
<td>0.3324</td>
<td>0.3252</td>
<td>0.2844</td>
<td>0.2539</td>
<td>0.1525</td>
<td>0.0577</td>
<td>0.0222</td>
<td>0.0209</td>
</tr>
<tr>
<td>Kant</td>
<td>0.3301</td>
<td>0.381</td>
<td>0.375</td>
<td>0.3604</td>
<td>0.3172</td>
<td>0.2717</td>
<td>0.2539</td>
<td>0.2541</td>
</tr>
<tr>
<td>Martyr</td>
<td>0.3340</td>
<td>0.4124</td>
<td>0.4694</td>
<td>0.5012</td>
<td>0.5253</td>
<td>0.5254</td>
<td>0.5254</td>
<td>0.5254</td>
</tr>
<tr>
<td>Psycho</td>
<td>0.3346</td>
<td>0.2036</td>
<td>0.1028</td>
<td>0.0444</td>
<td>0.0003</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Aristotelian</td>
<td>0.3312</td>
<td>0.3839</td>
<td>0.4277</td>
<td>0.4543</td>
<td>0.4743</td>
<td>0.4745</td>
<td>0.4745</td>
<td>0.4745</td>
</tr>
<tr>
<td>Martyr</td>
<td>0.3334</td>
<td>0.2826</td>
<td>0.3325</td>
<td>0.3759</td>
<td>0.499</td>
<td>0.6023</td>
<td>0.6422</td>
<td>0.6514</td>
</tr>
<tr>
<td>Psycho</td>
<td>0.3256</td>
<td>0.3355</td>
<td>0.297</td>
<td>0.2642</td>
<td>0.1499</td>
<td>0.0661</td>
<td>0.0362</td>
<td>0.0294</td>
</tr>
<tr>
<td>Emotional Deliberator</td>
<td>0.341</td>
<td>0.3819</td>
<td>0.3705</td>
<td>0.3599</td>
<td>0.3511</td>
<td>0.3316</td>
<td>0.3215</td>
<td>0.3193</td>
</tr>
</tbody>
</table>

Table 4: Frequency of strategies.

5.2 On the Emergence of Second-Order Strategies

While we have shown that second-order strategies can peacefully coexist it deserves mention that we have assumed particular second-order strategies, and this assumption may tarnish our findings. The fact
that we only include intuitive second-order strategies we find in the world and use these alone to demonstrate the thesis may be sloppy. A stronger argument for our findings would be to assume an array of second-order strategies, a number of which are not currently found in the world and run them in a simulation. If we find that after the simulation the only ones left are our intuitive moral theories, then we have shown the stronger claim—that given a number of meta strategies, a plurality of them shall be retained and a number of strategies we deem to be intuitive will dominate the population.

To start this project we must first define a number of alternative strategies for our simulation. To do this let us re-examine our best response dynamic once again. It is the following: whenever I am neighboring three or more stag hunters, I hunt stag, for all else, I hunt hare.

\[
\begin{align*}
\text{BR (0)} &= H & \text{M(0)} &= S & \text{A(0)} &= S \\
\text{BR (1)} &= H & \text{M(1)} &= S & \text{A(1)} &= H \\
\text{BR (2)} &= H & \text{M(2)} &= S & \text{A(2)} &= S \\
\text{BR (3)} &= S & \text{M(3)} &= S & \text{A(3)} &= H \\
\text{BR (4)} &= S & \text{M(4)} &= S & \text{A(4)} &= S
\end{align*}
\]

Figure 8: Best response (BR), Martyr (M), and an artificial (A) second-order strategy.

Best response is a function that maps a domain to a given co-domain. In our example, the domain is the number of neighboring hunting stag: \{0,1,2,3,4\}. The co-domain is the set of possible actions, \{hunt stag (S), hunt hare (H)\}. There are thus \(2^5\), or 32, unique functions of this sort. Martyr is another such function (see figure above). There also exists a number of artificial functions which don’t correspond to an intuitive moral theory. A (for artificial) in Figure 8 is such a function. Function A alternates between S and H. This obviously does not correspond to any intuitive moral theory. It is, however, a completely viable second-order strategy. For convenience, let’s refer to function A as SHSHS, for it spits out S when surrounded by four hunt stag, H when around three hunt stag, etc.\(^{19}\) Our task now is to incorporate all such artificial meta strategies into a colossal simulation. Of our eight intuitive moral theories, four of them can be represented

---

\(^{19}\)We will refer to a function by its outputs. For example we will refer to best response as SSHHH for it demands the individual to hunt stag when surrounded by four stag hunters, stag when surrounded by three stag hunters, hare when around two stag hunters, etc. Likewise Martyr is referred to as SSSSS.
Harvest Moon

as one of the 32 functions (Martyr, Psycho, Aristotle, best response). Once again first-order and second-order strategies are randomly assigned to the individuals in the population. We run 50 simulations of 200 individuals for 200 generations and get the following data:

<table>
<thead>
<tr>
<th>Strategy</th>
<th>SSSS</th>
<th>SSSH</th>
<th>SSHH</th>
<th>SSSHH</th>
<th>SSSHHH</th>
<th>SSSHS</th>
<th>SSISH</th>
<th>SSSH</th>
</tr>
</thead>
<tbody>
<tr>
<td>% of Pop</td>
<td>18.60</td>
<td>17.23</td>
<td>14.32</td>
<td>4.70</td>
<td>0.45</td>
<td>15.96</td>
<td>7.32</td>
<td>5.25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Strategy</th>
<th>SSHSS</th>
<th>SSHSH</th>
<th>SSHHS</th>
<th>SSHHS</th>
<th>SHHHH</th>
<th>SHHSH</th>
<th>SHHSS</th>
<th>SHHSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>% of Pop</td>
<td>8.02</td>
<td>2.23</td>
<td>1.74</td>
<td>1.05</td>
<td>0.54</td>
<td>0.39</td>
<td>0.55</td>
<td>1.19</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Strategy</th>
<th>HHHHH</th>
<th>HHHHS</th>
<th>HHHSS</th>
<th>HHHHH</th>
<th>HHHHS</th>
<th>HHSSH</th>
<th>HHSsS</th>
<th>HHSSH</th>
</tr>
</thead>
<tbody>
<tr>
<td>% of Pop</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.44</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Strategy</th>
<th>HHHHH</th>
<th>HHHHS</th>
<th>HHHSS</th>
<th>HSSSS</th>
<th>HHHHH</th>
<th>HHHHS</th>
<th>HHSSH</th>
</tr>
</thead>
<tbody>
<tr>
<td>% of Pop</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 5: 32 strategies and their frequencies.

Martyr and Aristotle both have a large chunk of the population while best response is weak but alive and Psycho has been annihilated. Also, a number of other non-intuitive strategies such as SSSHS did quite well. This suggests that among a random pairing of second-order strategies we could expect Martyr and Aristotle to nonetheless thrive.

6. Conclusion

Can a population converge to correct behavior while nonetheless sustaining a plurality of normative theories of action? This paper employs a number of simulations to answer this question. We convert eight intuitive moral theories into second-order strategies and incorporate them into a simple evolutionary model. We find that given the correct social structure, individuals will opt for the socially cooperative act (hunt stag) while nonetheless maintaining an array of vastly differing moral theories. Specifically, we find that more highly connected social networks foster this phenomena, while less connected social structures lead to substantially less levels of cooperation. While these results are promising, further investigation is required to determine the robustness of these findings.

---

20Once again, by our notation SSSHS does hunt stag in all situations except when it is the neighbor of exactly one individual, in which case it hunts hare.
References


On the Key to Negative Judgments of Beauty in Kants Theory of Taste

“What I have in mind is that art may be bad, good or indifferent, but, whatever adjective is used, we must call it art, and bad art is still art in the same way that a bad emotion is still an emotion.”

Marcel Duchamp
Kant’s *Critique of Aesthetic Judgment* does not contain an *Analytic of the Ugly*. The *Analytic of the Beautiful*, as might be expected, does not systematically concern itself with negative judgments of beauty. In the entirety of the *Critique of Aesthetic Judgment*, where Kant’s texts on aesthetics are found, we find nothing on the ugly outside of §48, which explains only that artists can present things as beautiful which would otherwise be considered ugly, without giving a conceptual explanation of the ugly itself. Indeed, Kant seems to be almost entirely unconcerned with these negative judgments. There are, however, reasons to believe that Kant should allow for judgments of the ugly.

One such reason is systematic. Kant, throughout the entire *Critique of Judgment*, almost never talks of pleasure in general without connection to displeasure, bringing in the conjunction in section XI of the First Introduction, marking the feeling of pleasure and displeasure together as the mental power or faculty corresponding to the cognitive power of judgment itself.

And indeed, in the general talk of pleasure and displeasure in each of the four moments there is never a suggestion of disanalogy. This being the case, we might expect a certain amount of symmetry between pleasure and displeasure which permits of a displeasure in the ugly opposite to the pleasure in the beautiful. Indeed, there are certain places outside of the third *Critique* where Kant says of displeasure that it is a “negative pleasure,” and correspondingly, that “ugliness [is] a negative beauty.”

Another reason we should allow for judgments of the ugly lies outside of the text, and in the basic intuition that aesthetic appraisal often results, not in a pleasure, but in a displeasure, and that this requires us to exercise taste all the same. To do this intuition justice it seems we should spell it out in terms of a displeasure in the ugly opposite the

---

21 Kant wrote one pre-critical essay on aesthetics entitled *Observations on the Feeling of the Beautiful and Sublime*, which is not contained in the *Critique of Judgment*. However, it is largely if not unanimously agreed that what is presented in the *Critique of Judgment* represents his mature work on aesthetics.

22 The tables at 245 and 264 in the “First Introduction” (Pluhar, p. 434-435) progress into a final product:

<table>
<thead>
<tr>
<th>POWERS OF THE MIND</th>
<th>HIGHER COGNITIVE POWERS</th>
<th>A PRIORI PRINCIPLES</th>
<th>PRODUCTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive power</td>
<td>Understanding</td>
<td>Lawfulness</td>
<td>Nature</td>
</tr>
<tr>
<td>Feeling of pleasure and displeasure</td>
<td>Judgment</td>
<td>Purposiveness</td>
<td>Art</td>
</tr>
<tr>
<td>Power of desire</td>
<td>Reason</td>
<td>Purposiveness that is also law</td>
<td>Morals</td>
</tr>
</tbody>
</table>

pleasure in the beautiful.

Worries like these have even prompted one commentator, Henry All-

ison, to find that “an ability to account for the possibility of negative judg-
ments of beauty is criterial for the adequacy of an interpretation of Kant’s theory of taste.” In lieu of these considerations, this essay will concern itself with providing a reading of the Analytic of the Beautiful which allows for pure negative judgments of beauty, or judgments of the ugly. It will focus primarily on §9 of the Analytic, which Kant himself calls the “key to the critique of taste.” This essay will be divided into three sections: the first section will deal with the nature of judgments of the ugly, the second will give an analysis of a contemporary disagreement between commentators Henry Allison, Hannah Ginsborg, and Beatrice Longuenesse with regards to the possibility of the ugly, and the third section will provide an interpretation of the ugly which, one hopes, is phenomenologically plausible as well as consistent with Kant’s theory of taste as a whole.

I

Although Kant never officially concerns himself with judgments of the ugly, we are not left without resources for constructing or understanding a judgment which is opposite the beautiful.

In an essay of 1763 entitled An Attempt to Introduce the Concept of Negative Magnitudes into Philosophy, Kant distinguishes between two (jointly exhaustive) kinds of opposition: logical opposition and real opposition. Logical opposition is opposition through the law of non-contradiction, and a conjunction of logical opposites has the consequence of nothing at all: “A body which is in motion is something, a body which is not in motion is also something; but a body which is both in motion and, in the very same sense, not in motion, is nothing at all.” Real opposition, on the other hand, is opposition which is not through the law of non-contradiction, and the conjunction of real opposites has the consequence of something. “The motive force of a body in one direction and an equal tendency of the same body in the opposite direction do not contradict each other. As predicates, they are simultaneously possible in one body. The consequence of such an

24 Allison. 115.
25 Critique. §9, 216, 61.
opposition is rest, which is something.”\textsuperscript{26} When this distinction is applied to pleasure, we find its logical opposite to be not pleasure, and its real opposite displeasure. Kant himself says in the Negative Magnitudes essay that “Displeasure is accordingly not simply a lack of pleasure. It is a positive ground which [when combined with pleasure], wholly or partly, cancels the pleasure which arises from another ground.”\textsuperscript{27} And, since the conjunction of these opposites results in something, we should expect to find the conjunction to result in something of an indifference, in the same way that motion in opposite directions gives us rest.

Thus far we have only talked in general of pleasure and displeasure; however, Kant finds that there are in fact three species of pleasure: pleasure in the agreeable, pleasure in the good, and pleasure in the beautiful. Kant first defines the agreeable in §3 as “what the senses like in sensation”; later, in §7, Kant tells us that “about the agreeable the following principle holds: everyone has his own taste (of sense)”. The agreeable is described in §9 as having “merely private validity.”\textsuperscript{28} The beautiful and the good are both distinct from the agreeable in their claim to universal communicability, which we may for now take to mean roughly that we demand of others that they ought to judge as we do. The good is defined in §4 as “what, by means of reason, we like through its mere concept.” Something is either “good for” something else or it is “intrinsically good”—either way, “the good always contains the concept of a purpose,”\textsuperscript{29} and, since it is good, a satisfaction of this purpose. The agreeable and the beautiful are both distinct from the good in being that which is liked without concept. It will here suffice to define the beautiful in contrast to the agreeable and the good: beautiful is that which is liked universally, without concept.

Using the above analysis of the species of pleasure in conjunction with the real/logical opposition distinction, we can arrive at the following two tables:

<table>
<thead>
<tr>
<th></th>
<th>+</th>
<th>0</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Agreeable</td>
<td>Indifference</td>
<td></td>
<td>The Disagreeable</td>
</tr>
<tr>
<td>The Beautiful</td>
<td>Indifference</td>
<td></td>
<td>The Ugly</td>
</tr>
<tr>
<td>The Good</td>
<td>Indifference</td>
<td></td>
<td>The Bad</td>
</tr>
</tbody>
</table>

Table 1: Real Opposition

\textsuperscript{26} Negative Magnitudes. 211. AK 2:17.
\textsuperscript{27} Negative Magnitudes. 220. AK 2:181.
\textsuperscript{28} Critique. §3, 205, 47. §7, 212, 55. §9, 217, 61.
\textsuperscript{29} Critique. §4, 207, 48-49.
And it is clear from this picture of the symmetry of pleasure that the ugly properly concerns real opposition, and that it is the real opposite of beauty. We can thus see that the ugly is distinct from the disagreeable in being universally communicable, and distinct from the bad in being disliked without concept. *Ugly is that which is disliked universally, without concept.*

**II**

With this definition in place, we are now in a position to properly understand the contemporary disagreement with regards to the ugly. Much of this disagreement arises from an interpretive disagreement regarding the following problematic sentence from §9:

> “Hence it must be the universal communicability in the given presentation, which underlies the judgment of taste as its subjective condition, and the pleasure in the object must be its consequence.”[^30]  

If this sentence is taken literally, the ugly becomes problematic precisely because universal communicability itself becomes a source of pleasure, leaving no room for a universally communicable displeasure, as is requisite for a judgment of the ugly. The three commentators I will discuss here loosely exhaust the way in which one may take this sentence. That is, either (1) to reject or revise it in order to allow room for judgments of the ugly, (2) to accept it as it stands, and use it to endorse a view which says that Kant cannot in fact allow for judgments of the ugly, or (3) to accept it as it stands, but find that it is nevertheless unproblematic for judgments of the ugly.

Henry Allison represents the first way that one may take this sentence, proposing a revision which replaces “the universal communicability of the mental state” with “a universally communicable mental

[^30]: [Critique. §9, 217, 61.](#)
state,” allowing that which gives rise to the pleasure in the beautiful to be a state which is itself universally communicable, but not the feature of this state which is its universal communicability. According to Allison, “there is nothing inherently problematic in a universally communicable state of displeasure (as opposed to a displeasure in its very communicability).” In order to show how this revision is unproblematic, Allison has to now spell out the mental state corresponding to both judgments of beauty and judgments of the ugly, and show that both make a valid claim to universal communicability, while the former is pleasurable and the latter displeasurable.

We said earlier that aesthetic judgments are non-conceptual—that is, they possess all of the necessary components for judgment in general, but they do so without concepts. In order to show that the requisite work for judgment is being done here, Kant describes the cognitive faculties in judgments of beauty to be related in a harmony and free play, and talks of this relation as the subjective condition for judgment in general. Allison appeals to these relations of harmony and free play in order to explain his mental states.

And Allison finds that if we are to understand how the mental state in a judgment of the ugly differs from that in the judgments of beauty, “it is essential to distinguish between the free play and harmony of the faculties.” On his view, ‘free play’ is a term representing the ‘mere reflection’ operative in a judgment of taste, which is just to say that ‘free play’ refers to a non-conceptual and universally communicable activity of the cognitive faculties. This activity of free play or mere reflection, Allison tells us, can “either succeed or fail to produce a harmonious relation”; when it succeeds, the mental state is pleasurable, and consists in a harmonious free play—the mental state in a judgment of beauty—and when it fails the state is displeasurable, consisting in a disharmonious free play, or free play absent harmony, which is precisely how Allison wishes to understand the mental state in judgments of the ugly, this being the mental state which gives rise to a universally communicable state of displeasure.

---

31 Allison. 115.
32 Allison. 116.
Allison’s view of judgments of beauty, as we have understood it, can thus be understood through the following table:

<table>
<thead>
<tr>
<th>Aspect of Mental State (Relation of Cognitive Faculties Described in Mental State)</th>
<th>Aspect of Judgment (Which the Mental State Gives Rise To)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harmony</td>
<td>Pleasure</td>
</tr>
<tr>
<td>Free Play</td>
<td>Universal Communicability Without Concept</td>
</tr>
</tbody>
</table>

Table 3: Allison’s View

And if we wish to understand the ugly, we need merely to replace harmony with its real opposite, disharmony, and thus pleasure with its real opposite, displeasure.

Allison is correct to notice that the pleasure in the beautiful, if we are to account for a universally communicable displeasure in the ugly, must arise from a mental state which is universally communicable, and not from the feature of this state which is its universal communicability. He is also correct to find that understanding the mental state in a judgment of ugliness requires of us that we distinguish between the harmony of the faculties and their free play. However, I will argue that the way in which Allison attempts to spell out the mental state in a judgment of the ugly does not allow of this state that it be universally communicable, and that it thus fails to properly represent the mental state corresponding to judgments of the ugly. In order to understand Allison’s error, it is essential to better understand what Kant means when he talks of the cognitive faculties in a judgment of beauty as related in a harmony and free play.

Kant first brings in these relations later in §9 in showing how judgments of beauty pertain to cognition, for, according to Kant, “nothing can be communicated universally except cognition, as well as presentation insofar as it pertains to cognition.” True cognitive judgments are universally valid objectively because they are conceptual, and thus “valid for every object contained under [a] given concept”; since judgments of beauty are non-conceptual, they do not have this objective reference point. Thus, if they are thought to have universal validity, it must be a subjective universal validity. Kant explains that “this subjective universal communicability can be nothing but the mental state

---

33 Critique. §9, 117-118, 61.
34 Critique. §8, 215, 58.
in which we are when imagination and understanding are in free play (insofar as they harmonize with each other as requisite for cognition in general).”\textsuperscript{35} Similar to the parenthetical proposition in this passage, Kant also explains that “for cognition, the determination of the object with which the given presentations are to harmonize (in any subject whatever) is the only way of presenting that holds for everyone.”\textsuperscript{36} That is, the harmony of the cognitive faculties is to be understood as the subjective condition for cognition in general, and this in turn is to be seen as how the mental state in the judgment of taste pertains to cognition, legitimating its claim to universal validity. We find in §35 of the Deduction that “the imagination’s freedom consists precisely in its schematizing without concepts,” such that we can understand the relation of the faculties which is responsible for the validity of the claim to universal communicability to be the harmony, and the relation of the faculties referring to the non-conceptual nature in the judgment of beauty to be the free play. Given these passages, we should understand something like this to be the picture, rather than (Table 3):

<table>
<thead>
<tr>
<th>Aspect of Mental State (Relation of Cognitive Faculties Described in Mental State)</th>
<th>Aspect of Judgment (Which the Mental State Gives Rise To)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harmony</td>
<td>Universal Communicability</td>
</tr>
<tr>
<td>Free Play</td>
<td>Without Concept</td>
</tr>
</tbody>
</table>

Table 4: Possible Response to Allison

However, if, as Allison maintains, the mental state in a judgment of the ugly consists in a disharmony or absence of harmony in the free relation between the representational powers, then there seems to be nothing in the mental state which pertains to cognition by subjectively grounding it, thus no ground for claiming the universal communicability of the mental state. While Allison is correct to suggest that “there is nothing inherently problematic in a universally communicable state of displeasure,” there is something inherently problematic in a universally communicable state of disharmony.

Hannah Ginsborg, in disagreement with Allison, proposes that Kant in fact leaves no room for judgments of the ugly in his theory of taste.

\textsuperscript{35} \textit{Critique}. §9, 217-218, 62.
\textsuperscript{36} \textit{Critique}. §9, 217, 62.
This is because she wishes to accept the problematic sentence of §9 as it stands, and use it to endorse a view of the judgment of beauty as a self-referential act of judgment, in which “I take my mental state in perceiving an object to be universally communicable, where my mental state is none other than the mental state of performing that very act of judgment.” According to Ginsborg, the universality of a judgment of taste consists precisely in taking it that everyone ought to judge as I do. She explains this as inherently pleasurable by appeal to a handful of passages in which Kant explains pleasure in terms of an “inner causality.” A good representative sample of these passages comes in at §10:

“Consciousness of a presentation’s causality directed at the subject’s state so as to keep him in that state, may here designate generally what we call pleasure; whereas displeasure is that presentation which contains the basis that determines [the subject to change] the state [consisting] of [certain] presentations into their own opposite (i.e., to keep them away or remove them).”

As well as in §60 of the *Anthropology*:

“What immediately (through the senses) drives me to leave my state (to go out of it) is unpleasant to me—it pains me; what similarly drives me to maintain my state (to remain in it) is agreeable to me, it satisfies me.”

And in these passages we get a picture of pleasure as that which causes me to maintain or persist in my current state, and displeasure as that which causes me to exit or leave my current state. According to Ginsborg, universal communicability consists in the judgment that everyone, and thus that I myself, ought to be in the mental state which I am in fact in. This can be seen to have a self-maintaining structure precisely in that the judgment, as well as the corresponding mental state, is appropriate; it, as Ginsborg puts it, “approves of itself”—the mental state encourages itself to go on, to go about its business.

However—and Ginsborg readily acknowledges this point—in virtue of its taking a literal reading of the problematic sentence of §9, this

---

37 *Key*. 299.
38 *Critique*. §10, 220, 55.
39 *Anthropology*. §60, 231.
40 *Key*. 302.
reading makes it incoherent to talk of a universally communicable displeasure, precisely because, on her view, universal communicability is inherently pleasurable—thus, wherever there is universal communicability in a judgment, according to Ginsborg, there will be pleasure, and this leaves no room for a universally communicable displeasure, as is needed for a judgment of the ugly.

It is interesting though, that Ginsborg wants to agree with Allison in maintaining that “it is a reasonable criterion of any interpretation of Kant’s theory of taste that it be able to account for the possibility of negative judgments of taste,” and she wishes to satisfy this criterion by showing that all one wishes to call ugly in fact collapses into beauty’s logical negation: the not beautiful.

She begins in §48, where Kant gives the only examples we have of ugly things in the entire third Critique, these examples being “the furies, disease, [and] devastations of war.” And Ginsborg rightly does away with these as candidates for judgments of ugliness, for they are all either “clearly connected with interest” or “seem to counterpart pleasure in the agreeable rather than pleasure in the beautiful.” Either their status as disagreeable or their status as interested remove these as candidates for ugly things, and places them simply in the category of not beautiful things.

Ginsborg, though, does not find these examples to exhaust the kinds of judgments where we take something to be ugly: “We might be struck by the ugliness of, say, a piece of gravel worn as an earring, or the institutional-style folding chairs a friend has chosen to place in her newly decorated living room, or of a sudden cacophony in a piece of otherwise tonal music.” However, Ginsborg would like to understand these as cases where we have an expectation of beauty, or find beauty to be appropriate, and it is only because our expectations are disappointed that we find these things ugly. “We are likely to be indifferent to the same thing in contexts where beauty is not expected or required: for example when the piece of gravel is part of a gravel path, or the metal chairs are in a school auditorium, or the cacophony is the sound of the orchestra tuning up.” And we do not tend to find these things

---

41 Critique. §48, 312, 180.
42 Aesthetic Judging. 177.
43 The first moment of the Analytic of the Beautiful concludes in saying “taste is the ability to judge an object, or way of presenting it, by means of a liking or disliking devoid of all interest. The object of such a liking is called beautiful.” (211) We should then expect this to carry over to the ugly, making it a disinterested displeasure.
44 Aesthetic Judging. 178.
ugly, according to Ginsborg, precisely because we do not here have an expectation for beauty. Having thus traced the problem of the ugly to the problem of the not beautiful, Ginsborg sees her view as adequately accounting for (the impossibility of) judgments of ugliness. Beatrice Longuenesse, like Ginsborg, offers a reading of §9 which takes the problematic sentence literally, and understands universal communicability to be inherently pleasurable. However, she does not see this move as devastating for judgments of the ugly; instead, she merely uses the sentence to expose what she finds to be the complex nature of both judgments of beauty and judgments of the ugly.

Longuenesse finds that the feeling in the beautiful consists in both “a first order pleasure in the [harmony and] free play of our own mental capacities” as well as a second order pleasure which is a pleasure “in the universal sharability of the first order pleasure.” This second order pleasure is necessary because we can, according to Longuenesse, have a pleasure in the free and harmonious play of the faculties without being aware that our pleasure is universally communicable, and “this sense of a possible universal sharability of a pleasure is the source of the second order pleasure” the second order pleasure being the judgment of beauty proper.

Longuenesse also believes that this two order reading of judgments of beauty allows the possibility of universally communicable judgments of the ugly. In order to understand how she accounts for such judgments, we should first recall that “in the case of judgments on the beautiful, the pleasure is [taken] in the universal communicability of a pleasure” and that “both pleasures have to be in play to make the judgment a judgment of the beautiful.” In the case of the ugly, the first order feeling is one of displeasure. Since Longuenesse accepts the problematic sentence as is, she does find that universal communicability

\[\begin{array}{|c|c|c|}
\hline
\text{The Agreeable} & \text{Indifference} & \text{The Disagreeable} \\
\hline
\text{The Beautiful} & \text{Indifference} & X \\
\hline
\text{The Good} & \text{Indifference} & \text{The Bad} \\
\hline
\end{array}\]

\[\text{Longuenesse. 153.}\]
\[\text{Ibid.}\]
\[\text{Longuenesse. 154.}\]
\[\text{Longuenesse understands the first order pleasure in the beautiful as taken in the free play and harmony of the faculties. As regards the ugly, Longuenesse says here only that “presumably, imagination and understanding are arrested in their free play... whatever this may mean in any particular case”(154)—This is obviously vague, however it does seem to be committed to something like Allison’s view of the mental state.}\]
is a source of pleasure, making the second order feeling a pleasure which is taken in the universal communicability of the first order displeasure. However, this still allows the judgment to communicate a displeasure universally, allowing for the judgment here to be both non-conceptual and universally communicable: the judgment of the ugly proper.

Longuenesse does admit that “this suggestion is speculative,” however, she finds the tension between pleasure and displeasure here to be “phenomenologically plausible,” and finds that her account, in any case, is “at least compatible with what Kant actually says.” However, I would like to argue that her account of the ugly, while it may be phenomenologically plausible, is not consistent with Kant’s overall theory. This is because it was the distinction between real and logical opposition which allowed us to distinguish between judgments of ugliness and judgments of the not beautiful, the ugly properly understood as the real opposite the beautiful. In combining real opposites, though, “one thing cancels that which is posited by the other”; only “the consequence is something.” In the motion example, motions in opposite directions combine in the same body to form rest. With regards to feeling, the conjunction of real opposites (pleasure and displeasure) concludes in indifference. Thus, we should expect that the combination of a first order displeasure and a second order pleasure leads to an indifference, and not to a displeasurable state with a further element of tension.

III

By way of its making universal communicability inherently pleasurable, the problematic sentence of §9 forces us either, as with Ginsborg, to reject the possibility of a universally communicable displeasure, or, as with Longuenesse, to understand the feeling involved in a judgment of ugliness to consist in a problematic admixture of pleasure and displeasure. For reasons like these, I would like to accept Allison’s revision to the problematic sentence, and understand the pleasure in the beautiful as due to a mental state which is in fact universally communicable, but not the aspect of this state which is its universal communicability. Rather, I would like to understand the pleasure as due to a different aspect or relation of the cognitive faculties described in the mental state.

\[50\]Longuenesse. 155.
But what, then, shall we say is responsible for the pleasure? As we have seen in our analysis of Allison, we cannot understand this mental state as disharmonious, for this would negate the legitimacy of the claim to universal communicability. We also cannot understand it as somehow unfree, for this would make judgments of the ugly conceptual, collapsing them into ordinary cognitive judgments rather than aesthetic ones. It thus seems that there is not enough room for variation in the mental state to allow for a state which represents a universally communicable, non-conceptual displeasure.

However, this conclusion may be a bit premature; allow me to suggest the following picture of the mental state in a judgment of the ugly. The mental state must be both harmonious and free, for the reasons we have just stated; however, I wish to argue that the displeasurable mental state in a judgment of the ugly differs from the pleasurable state in a judgment of beauty precisely in its not being in a state of play. And indeed, that the relation of play described in the mental state is precisely what is responsible for the pleasure.

The relations of harmony and freedom are never themselves talked of as inherently pleasurable without being connected to play, play itself being the only aspect of the mental state defined independently as inherently pleasurable: in §43 Kant describes play as “an occupation that is agreeable on its own account.”51

Kant contrasts play with task in §51 and the General Comment after §29; task being a “law governed occupation,”52 play being activity or occupation which is not so governed. This makes freedom a necessary condition for play; however play is not a necessary condition for freedom. This is most apparent in Kant’s talk of the sublime in §23, describing it as a “seriousness, rather than play, in the imagination’s activity”; this in turn used to explain the feeling of the sublime as a “negative pleasure.”53 The sublime, like the beautiful, involves a universal communicability which is free of concept; but instead of a harmonious free play, the sublime consists in a harmonious and free seriousness.

We do not want to understand the mental state in a judgment of the ugly to also consist in a seriousness rather than play, for then we would be unable to distinguish between the mental state in the sublime and that in the ugly. We cannot understand the mental state to consist simply in not play either, if we understand the ‘not’ here to represent

---

51 *Critique*, §43, 304, 171.
52 *Critique*, General Comment, 268.
53 *Critique*, §23, 245, 98.
a logical negation, for if the play is responsible for the pleasure, then not play would simply get us not pleasure, and what we need is a displeasure. We must understand the cognitive faculties to be engaged in something, then, that is the real negative of play. Kant does not give us a term which occupies this post\textsuperscript{54}, and, from what I have seen, neither does any of the secondary literature. I propose, then, that the cognitive faculties are engaged in something of a mischief\textsuperscript{55}; rather than play, in a judgment of the ugly; this mischief correctly understood as the real negative of play.

Let us look at this a bit more systematically. If what has been said so far is correct, then we can come up with the following picture of the correspondence between the different kinds of relations of the cognitive faculties, and the features of the judgments which arise from them:

\textsuperscript{54}The only terms which we are given to contrast with play are task—which Kant in some places also calls “business”—and seriousness. Task is play’s logical negative, for something which is both a law governed occupation (task), and, in the very same sense, an occupation which is not governed by law (play) is nothing at all. Seriousness also is not the real negative play, for it manifests itself in the feeling of the sublime, which is not the real opposite the beautiful, but rather “a mere appendix to our aesthetic judging of the purposiveness of nature.” (Critique. §23, 246, 100.)

\textsuperscript{55}Credit for the term goes to my editor, Laura Davis. The word ‘mischief’ was chosen because it is has a destructive connotation, and seems to be more in the realm of play than task or seriousness, granting us the ability to draw its relation to play as one of opposition, and its structure as self-defeating (displeasurable).
Notice that what is under the column “Positive” is the judgment of beauty, and that judgments of the ugly and judgments of the beautiful, as we saw in our analysis in Section I, share many of the same aspects. Indeed, the only aspect of a judgment of the ugly which is different from the judgment of beauty is that with beauty there is a pleasure, whereas with the ugly there is a displeasure. The aspect of the mental state or relation of the faculties which gives rise to displeasure, as derived from the table above, is mischief. Thus, if what we have said is correct, something like this should be the picture of judgments of the ugly:

<table>
<thead>
<tr>
<th>Aspect of Mental State</th>
<th>Aspect of Judgment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harmony</td>
<td>Universal Communicability</td>
</tr>
<tr>
<td>Freedom</td>
<td>Without Concept</td>
</tr>
<tr>
<td>Mischief</td>
<td>Displeasure</td>
</tr>
</tbody>
</table>

How, then, are we to understand this relation of mischief? As we have said, it is that aspect of the mental state which is responsible for the displeasure in the ugly. The question we should ask, then, is this: how are we to understand the operations of displeasure in the mental state?
We can, similar to Ginsborg, understand the inner causality of pleasure and displeasure to be normative. However, Ginsborg believes that we should understand the normative causality which maintains the mental state in a judgment of beauty to consists in the feeling that everyone, and thus that I myself, ought to be in the mental state that I am in, and since I am in fact in it, this judgment can be seen as having a tendency to maintain itself, for I feel I ought to be in it. We cannot understand this to be the way in which the normative causality functions if we are to account for judgments of ugliness, because here also I feel that everyone and thus that I should be in the mental state I am in fact in, only here displeasure must be understood as having a self-defeating and not a self-maintaining structure.

As Longuenesse correctly emphasizes, the judgment of beauty proclaims a pleasure or liking to be universally communicable; thus, as Allison suggests, there should be nothing problematic in a judgment proclaiming a displeasure or disliking to be universally communicable, so long as we can account for the universally communicable state which gives rise to it. When we make such a judgment, we proclaim that everyone, and thus I, ought to be in this mental state of displeasure, which is just to say that everyone/I ought to drive themselves/myself away from this state or presentation. This is best understood if we consider the distinction between logical and real opposition when it is applied to the universal normative demand itself.

And when we carry out this application we find that ought is opposed both to not ought—its logical negative, informing us that this is not a proper application of the normative modality—and ought not, where not here signifies the real negative of that which it takes as its object. The judgment of beauty declares that ‘everyone ought to share my pleasure’, or that ‘everyone ought to maintain themselves in this mental state’; the judgment of the ugly, that ‘everyone ought to share my displeasure’, or ‘ought not to maintain themselves in this mental state’, which is just to say that everyone ought to share my displeasure.\(^{56}\) or ‘ought not to maintain themselves in this mental state’\(^{57}\), which is just to say that everyone ought to share my displeasure. Notice the only normative difference here between judgments of beauty and judgments of the ugly is that, with the beautiful, one feels everyone ought to maintain themselves in this mental state, whereas

\(^{56}\)It may be considered artifactual of our language that ‘not’ has only a common usage which refers to negation and not real opposition, however the not here is intended to indicate such a real opposition.

\(^{57}\)Here we are simply replacing pleasure and displeasure with their definitions in terms of inner causality—pleasure being defined as self-maintaining, displeasure self-defeating.
with the ugly, one feels everyone ought not to maintain themselves in this mental state: it is a negative judgment of beauty.

My suggestion for an interpretation of the ugly which is both consistent with Kant’s overall theory as well as plausible phenomenologically is this: in a judgment of the ugly, we feel everyone ought to feel our displeasurable mental state, consisting in a free harmony of the faculties which is engaged in mischief rather than play, and precisely because it engaged in this real negative, does not possess the self-maintaining structure of pleasure, but rather the self-defeating structure of displeasure, and this is precisely what is communicated universally.

Allow me to end in an elucidation by example. Hannah Ginsborg offered us the example of a “cacophony in a piece of otherwise tonal music,” describing it as something we often think of as ugly. She then attempted to reduce it to an example of something which is not beautiful, feeling that we would likely be indifferent to the same cacophony if it were the sound of “the orchestra tuning up,” or in a case when we do not have an expectation of beauty. According to Ginsborg, we only find something to be ugly in this way when we expect it to be beautiful and it fails to strike as us such. I would like to challenge this understanding.

Consider a beginning pianist practicing a newly acquired piece. Knowing her skill level, we expect nothing beautiful to come from her performance. Nevertheless, the false notes strike us as ugly, and we may find ourselves eagerly anticipating her improvement so that this process may one day be pleasurable for us. The sound of an orchestra tuning up is not often accompanied with an expectation for beauty; however, this does not revoke its ability to make ugly noises. If the entire process is atonal, or consists entirely in cacophonous noises, these things may still be ugly, even though we do not expect them to be beautiful.

What Ginsborg has in mind here may not be that we must understand the cacophony as not beautiful, so much as understand its context dependence to make it a case of an adherent and thus an impure judgment of the ugly. However, here we should note that it need not be heard as a cacophony in order to be heard as ugly: one may simply say of the racket ‘this is ugly’, without applying concept to it so as to make

---

58 A judgment of adherent beauty is defined in §16 as a judgment of beauty which has a partly conceptual basis; for example the judgment ‘this sunset is beautiful’ as opposed to ‘this is beautiful’ full stop. This mingling with concept makes it impure.
it impure. A cacophony or false note, then, seems to be a paradigm example of something ugly.

References

Allison, Henry E.. *Kant’s Theory of Taste*.


Ginsborg, Hannah. *On the Key*.

Kant, Immanuel. *Anthropology from a Pragmatic Point of View*.

Kant, Immanuel. *An Attempt to Introduce the Concept of a Negative Magnitudes into Philosophy*. Trans. David Walford.


Longuenesse, Beatrice. *Kant’s Theory of Judgment*.
The Utility of Maximal Properties and the Possibility of Mereological Gunk Both as Sufficient Reasons for Preferring Mereological Universalism over Nihilism
The ontological debate over the existence of composite objects has long been criticized as nothing more than a substanceless metaphysical amusement. This allegedly pointless debate—which we shall call the compositional debate—centers around claims like ‘Are their tables?’ and ‘Do composite objects exist?’ to which mereological compositionalists answer ‘yes,’ and mereological nihilists answer ‘no.’ Clearly, however, the philosophers who engage in this debate believe that their answers to such questions—which objectively differ at least in semantics—also express opposing and metaphysically significant facts about the world. Nonetheless, other philosophers have argued that the compositional debate is ultimately insubstantial and futile. Karen Bennett is one such philosopher; in her paper ‘Composition, Colocation and Metaontology,’ she argues that though the compositional debate avoids some types of dismissivisms (i.e., metametaphysical arguments that claim that a metaphysical argument is futile), it does fall victim to at least one dismissivism—which Bennet calls epistemicism: that is, that though the ontological dispute over ‘There are Fs’ may actually have a defined truth-value, there is insufficient reason for believing which truth-value that is, thereby rendering the debate futile. Bennett substantiates her argument by first showing that there are no methodological reasons to prefer either side of the compositional debate. She then continues to evaluate that certain philosophical arguments attacking mereological compositionalists apply equally to and therefore do not favor mereological nihilism. I will argue that Bennett’s objection that the compositional debate is insubstantial is false by first showing that one philosophical argument—namely, the Problem of the Many—favors the compositionalist. Furthermore, I will show that the possibility of mereological gunk (i.e., the theory that every part is a composite\textsuperscript{59}) serves as a further sufficient reason for favoring the mereological compositionalist’s ontology over that of the mereological nihilist.

Before we can evaluate Bennett’s metametaphysical arguments, we must first understand the metaphysical debate at hand. The compositional debate asks, “Are there tables?” Or, more generally, “Do composite objects exist?” When the mereological compositionalist answers ‘yes’ to these questions, he affirms that some objects have proper parts: parts that are contained in but not equivalent to the whole composite object. In order to explicate the concept proper part, the mereological compositionalist must hold that at least some classical mereological definitions are intelligible and that some mereological theses truly cor-

\textsuperscript{59}These definitions will be explicated later.
respond to the way the world is composed. Mereology is defined as taking a parthood relation as primitive. In other words, sentences like ‘x is part of y’ (symbolized as ‘x ≤ y’), or equivalently ‘y contains x’ cannot be explained by sentences containing more elementary concepts, as the relations ‘is a part of’ and ‘contains’ are taken as primitive, and do not supervene on other concepts. The parthood relation, then, has the following properties:

(r): x ≤ x, (reflexivity: all objects contain themselves as parts)

(t): If x ≤ y and y ≤ z, then x ≤ z, (transitivity: if z contains y and y contains x, z contains x)

(a): If x ≤ y and y ≤ x, then x = y, (antisymmetry: if y contains x and x contains y, then x is equivalent to y).

A compositionalist accepts these axioms, as they resemble our natural intuitions about the composition of objects. After all, it follows naturally from an intuitive understanding of parts that “The table contains itself,” and “Since the leg is part of the table, a part of the leg is also part of the table.” With these definitions, we can define a proper part.

(i): x is a proper part of y (symbolized by x < y) iff x ≤ y and x ≠ y.

A leg of a table is, thus, a proper part of the table given that it is a part of the table but not identical to the (whole) table. It follows that when a composite has a proper part a, then there is some further proper part b of that composite that does not overlap with the a. Overlap is defined as:

(ii): x and y overlap iff there is an object z such that z ≤ x and z ≤ y. If there is no such z, then the objects do not overlap and are disjoint.

So, (iii) if x < y, then y has a part z that is disjoint from x. This disjoint part z is such that z < y, i.e., z is a proper part of y.

The whole of the table is defined as the fusion of all of its proper
parts. Fusion is defined as:

(iv) A region $f$ is a fusion (= mereological sum) of a set of $y$s $S$ iff: for an object $x$, $x$ overlaps $f$ iff $x$ overlaps some $y$ contained in the set $S$.

Sometimes the fusion of just some proper parts will sufficiently express the table—for example, the fusion of all four of its legs and the chunk of wood attached on top of its legs; the fusion of these parts is the table. The whole table may be expressed as the fusion of other proper parts as well since its proper parts may overlap and need not correspond to regular English words; for example, the table may be defined as the fusion of the three following proper parts: the whole table except for its south-west-most molecule, its south-west-most molecule and some strange-shaped-region contained within one of its legs. Since both the above fusions express the same total region, they thus express the same table. So, when the mereological compositionalist answers ‘yes’ to the question “Are there tables” he commits himself to there being some unique composite object (‘table’) that is equivalent to the fusion, or mereological sum, of its proper parts (Russell 4-5; Van Cleve 321-323; Markosian 341-343).

The mereological debate is a debate about ontology—i.e., what objects exist. The compositionalist must then make a choice about how many objects there are total—that is to say, he must explain when two objects have a fusion (which itself counts as a further object). Since the mereological compositionalist wants to say that all things that we call ‘tables’ are whole tables—and not merely a conglomerate of non-table parts that together look like tables—the compositionalist might say that:

(v) Necessarily, all sets of objects have fusions.

---

60 For example, say we grant three non-overlapping objects in a hypothetical ontology: $A$, $B$, $C$. The compositionalist thinks that at least some fusions exist, so he holds that there will be more than three objects in our ontology when completed. If all fusions exist, this will include seven objects: $A$, $B$, $C$, $A+B$, $B+C$, $A+C$, $A+B+C$. 
This thesis is called *unrestricted composition* and is held by the *mereological universalist*; unrestricted composition, then, guarantees that whenever we see a table, we have before us one unique composite object (‘table’) and not merely a set of parts that look like a table.\(^{61}\) In such a way, it is this thesis (\(v\)) that the mereological universalist uses to account for the truth of “There are tables.” But, unrestricted composition does not just add whole tables to our ontology. Now, all fusions of any set of objects must count as an object in our ontology, including gerrymandered, disjoint objects like ‘the mereological sum of my pen, half the watermelon in my kitchen and a pebble on the moon.’ What’s more, if, as Ned Markosian suggests, unrestricted composition leads to a commitment to Four-dimensionalism, then objects might be composed of any set of objects at any time, so that the mereological universalist must commit to even stranger objects like ‘the mereological sum of a dinosaur’s tail 20,000,000 BC and a column of air in my room today’ (Markosian 345-346). Unrestricted composition guarantees that all the composite objects we want to say exist do exist, but at the cost of committing to a horde of unfamiliar and unintuitive objects.\(^{62}\)

---

\(^{61}\)What is a conglomerate of table parts that looks like a table but is not a table? This is admittedly a strange way to speak. The idea is that if not all fusions exist then even when we look at some ‘object’ and see that there is stuff there we have no way of knowing whether the ‘whole’ that we are considering is a genuine object or whether the stuff we see is a bunch of non-fused parts that only seem ‘whole’ to us but do not count together as a whole or fused object. Skepticism of this sort is a denial of *mereological universalism* and the very fact that the idea underlying this denial resists common sense expression seems telling against it. Nonetheless, we should try to clarify this denial, and for this task I provide an analogy:

Imagine that four table-legs are standing up-right on the ground and we place a wooden board on top of them. In some sense what we have created is—or, at least, is like—a table. But perhaps we would be hesitant to call it a genuine table because the parts have not been fastened together with nails, etcetera, so that really it seems just to be some table legs with a wooden board on top and not a completed whole. What the mereological skeptic—who resists thesis (\(v\)) that all objects necessarily have fusions that count as objects as well—says is that even if the parts if the table-like thing that we created were to be fastened together, we should still hold a similar hesitance to call it a genuine table as opposed to *just* a conglomerate of table-looking parts (e.g., the legs, wooden board) that itself does not count as a genuine object.

\(^{62}\)Alternatively, if the compositionalist wants to avoid these bizarre objects, he must find *some* way to allow for fusions of some objects to exist but not *all*. I will not consider these more ‘moderate’ compositionalist views in the body-text of this
Harvest Moon

paper, but something might be said about why I treat mereological universalism as the paradigm compositionalist theory. One such reason is that many moderate compositionalist theories are doomed to failure: for example, committing to some principle that restricts fusion so that it occurs only with certain sets of objects will suffer some flaws mentioned below. Such a moderate compositionalist accepts some claim in the following form:

\[(v') \text{Necessarily, for any set of objects, there is a fusion of those objects iff the objects bare some relation } R, \text{ (where } R \text{ might be defined by contact, or fastenation, etcetera).}\]

It is the compositionalist’s burden to explicate what defines the relation \(R\). Intuitive candidates will have some counterexamples; for example, if we define contact of regions (overlapping parts) as what accounts for their fusion, then specific tokens of the letter ‘i’ will not be single objects as the dot and the base regions of the ‘i’ do not overlap. But, more substantially, since, in any possible \((vi)\) principle, fusion will be relative to a relation \(R\) whose holding is a matter of degree, it follows, then, that whether fusion occurs will either be objectionably arbitrary or give rise to an objectionable vagueness that conflicts with Leibniz’s law of excluded middle (i.e., that a single object cannot both have and not have property \(A\)). It will be objectionably arbitrary as to when fusion occurs if some near duplicate objects differ with respect to whether or not they count as in relation \(R\) only by an infinitesimally minute difference. However, if the cutoff point for counting as in relation \(R\) is not so strongly and arbitrarily defined, then it is vague as to when some set of objects has a fusion. This is a problem with vague predicates in general. Consider the predicate ‘is bald.’ We can intuitively determine whether a man is bald or not, but there are cases where it is difficult to decide either way. But, certainly no one is both ‘bald’ and ‘not bald’ as that is contradictory. So we might try to define baldness as having a less-than-15% area of hair covering the top of one’s head. But this seems hugely arbitrary—a man with a 15% area of hair will not be bald, and a man with a 14.9999998% area of hair will be despite looking essentially identical. If the cutoff is sharp, one hair might make the difference. If it is not, then some people will either be both ‘bald’ and ‘not bald’ or ‘indeterminate concerning baldness,’ both of which are logically invalid. The problem with basing a compositionalist theory on a relationship which determines whether there will be a fusion by a matter of degree is that it will be based on a vague or objectionably arbitrary predicate, thereby creating skepticism about what counts as genuinely in our ontology or not. This consequence implies, there are some sets of objects for which it is neither clearly true nor false whether they have a fusion; the fusions of such sets of regions can neither be said to be in our ontology nor not, so there is no truth to the matter about how many objects exists, which contradicts the law of excluded middle. Since the compositionalist wants to maintain that reality is completely determinate yet reject any objectionably arbitrary answer to whether fusions occur, he must reject that composition occurs relative to a relation (Van Cleve 326-331; Sider 2007 257-262).

There is one further way a moderate compositionalist may attempt to claim that composition occurs only sometimes. This view is compositional brutalism. Compositional brutalism holds that:
The mereological nihilist, on the other hand, answers ‘no’ to the question “Are there tables?” To the nihilist, there are no composite objects. As such, all everyday objects are merely fictional representations that exist only insofar as they are useful logical constructions. Though the nihilist’s motivation for this denial might be to rid of the now-unnecessary parthood-relation (which the compositionalist takes as primitive), the basis of mereological nihilism is expressible in mereological terms. That is:

(vi) Necessarily, no objects have proper parts.

If this is the case, then no objects may contain other objects. Instead, every object is thus a logical atom or mereological simple that only contains itself as a part. In this picture, talk about fusions becomes unnecessary, as the only fusions are fusions of single objects, and thus are still merely simples. Believing only in simples need not contradict our intuitions about everyday objects, like tables, forks and see-saws. Simples, which are the only ‘real’ objects, then, serve as a basis for all our everyday, fictional objects—nonetheless, these fictional composite objects do not appear in the ontology of our universe. When we re-

(v”) some sets of objects have fusion and others do not, but there is no true, non-trivial and finitely long way to express whether or not fusion occurs.

There may be some true infinitely long expression that expresses when composition necessarily occurs by listing every set of objects that has as a fusion, but this proposition is far and beyond human comprehension and epistemically out of reach. That composition occurs sometimes is a brute fact that cannot be justified with any form of substantiation, yet it must be related to our perceptual intuitions about objects. Brutalism is particularly attractive as it can be consistent with our commonsense intuitions as to when objects have a fusion, therefore the brutalist need not affirm the existence of gerrymandered composite objects of the mereological universalist. Additionally, brutalism entails no vagueness about what objects exist, which is a shortcoming of other moderate positions (Markosian, 352-354). I think that compositional brutalism has a great deal of merit, but it is difficult to argue for and it does not fall within the scope of this paper to discuss it any further, as it will not be necessary to consider concerning Bennett’s arguments about the compositional debate. Any subsequent reference to the compositionalist theories will thus concern those of the mereological universalist, who hold that unrestricted composition (v) is true. To reiterate: Necessarily, for any set of objects, it has a fusion ( = unrestricted composition).
fer to ‘a table,’ we actually refer to many simples arranged in some specific way (e.g., ‘table-wise’) or instantiating some properties and relations so that they fulfill the ‘table-role.’ So, even though the ‘table’ does not exist, propositions about that ‘table’ may have a defined truth-value given that those propositions reduce to propositions about simples arranged-tablewise. Since the nihilist holds that all propositions about material objects express claims about simples, the nihilist can recapture all our intuitions about how (what seem like) composite objects work without committing to the existence of composite objects. There are no tables, but whenever we see ‘a table,’ there are simples arranged-tablewise over there.

Both the mereological compositionalist and nihilist theories are adequate for expressing claims about material objects. And, both are consistent with the way we perceive the material world. So, the question remains: which ontological theory should we favor? Bennett argues that we have insufficient justification to choose either and that because of this lack of justification the compositional debate is futile. Let us trace through her argument.

Bennett holds that there is no available, adequate justification to settle which side of the compositional debate has the stronger theory. She thinks that “Are there composite objects?” cannot be settled as either true or false by virtue of the compositional debates being too speculative to be solved, i.e., lacking justification in favor of either party. To Bennett, whether a debate suffers this futility boils down to whether the debate lacks justification in the following three ways: When judging which party has the correct theory, (1) there is no methodological reason to prefer either party’s theory (i.e., no answer to the question: “which party presents a simpler model for the metaphysics in question?”); (2) there is no empirically derived reason to prefer either party (i.e., no way of scientifically testing either party’s theory); or, (3) there are no convincing philosophical argument to act as evidence for selecting one party’s theory as correct. Bennett holds that all three of these ways to lack justification in favor of one side apply to the compositional debate, and therefore it is futile. Bennett’s argument that (1) applies is sophisticated and convincing, and (2)’s application is hardly difficult enough to warrant discussion; I find, however, that Bennett’s argument for (3)’s application—which attempts to show that some of the major arguments attacking the compositionalist’s theories present similar problems for the nihilist and therefore do not count as evidence in favor of nihilism—is incomplete and therefore
false. Particularly, accepting Theodore Sider’s definition of maximal properties may actually twist the Problem of the Many to be in favor of the compositionalist—despite the fact that it is usually leveraged against mereological universalism. Finally, I want to raise a new argument in favor of the compositionalist overlooked by Bennett; this argument is that accepting the possibility of mereological gunk commits the nihilist to an empty ontology, which I take to be absurd.

That (2) applies to the compositional debate can be argued for easily. No laboratory study will be able to make the distinction between whether an object is a genuine fusion of its parts or whether it is merely a bunch of simples. Parthood is not a concept that we can see, feel or taste in things—and though it may be related to our linguistic conceptions about our senses, no empirical study will serve to favor either party in the debate.

On to methodology—Bennett has a powerful argument to show that (1) applies to the compositional debate, i.e., that neither side presents a simpler methodology with which to express claims about objects. Before getting into the specifics of this argument, it is important to note that, as Bennett observes, both parties of the compositional debate exhibit a tendency to difference-minimize, that is, both parties wish to show that they can adequately capture the intuitions that the other party claims are advantages of its theories.

The compositionalist wishes to downplay his higher-ontology by explaining that its commitment to the existence of more objects really requires nothing more than committing to the nihilist’s ontology, i.e., commitment to the existence of fusions has an ‘ontologically innocent’ character. “It’s a mistake to think we are making an extra step,” the compositionalist says. Bennett explains that, in order to resist objections from the nihilists by way of difference-minimalizing, the compositionalist might make a claim along the lines of, “When I say tables exist, it means nothing more than there are some simples arranged-tablewise over there.” That is to say, though the compositionalist has a larger ontology, he wants to show that his theory is adequate to represent what the nihilist has with his simples-talk, namely, that (alleged) composite objects are in some way equal to and not more than their constituents (e.g., simples). For reasons that will be important later, I think that this paraphrase of the compositionalist’s attempts towards difference-minimizing might better be phrased by: ‘When I say tables exist, it means nothing more than there are some table-like parts over there, where those parts might be simples arranged-tablewise or other
things whatever.’ This rephrased version will allow the compositional-
ist to capture the difference-minimizing he wants, as it still coheres with simples-talk, but it does so without committing the compositionalist to the necessary existence of simples (Bennett 47-50).

Returning to issues of methodology: Prima facie, the nihilist’s methodology for representing objects seems to enjoy a greater simplic-
ity than that of the mereological universalist, as the nihilist only has to commit to a minimal ontology of simples, avoiding the infinites of gerrymandered fusions that the universalist must claim also exist. But, as we shall see, the situation is not so clear-cut because, in order to capture normal intuitions about (alleged) composite objects, the nihilist must up-play his ontology’s expressive power, which has the consequence of complicating his methodology. We have already discussed how the nihilist holds that claims about (alleged) composite objects may reduce to claims about simples. This move requires that the nihilist use plural quantifiers and non-distributive predicates.\(^{63}\) That is, since the nihilist holds that ‘there are simples arranged-tablewise over there’ does not entail the existence of the composite object ‘table,’ then the predicate ‘arranged-tablewise’ necessarily quantifies over some plural simples and not a set of simples.

This is good and well until the nihilist is asked to make sense of claims whose truth already relies on plural quantification when applying to composite objects, like \((q)\): ‘the tables are in a row.’ The plural quantifier is already, so-to-speak, ‘used up’ in paraphrasing reference to a table (which is already defined in terms of a plural quantifier placing a non-distributive predicate ‘arranged-tablewise’ over some simples) and thus must appeal to a further plural quantification over plural quantification to handle the further non-distributive predicate ‘in a row’ in order to capture \((q)\). The situation only gets more complicated if the rows of tables are aligned in columns. In fact, all statements that explain how bigger things are composed of smaller things—for example, scientific claims like, “cells are composed of organelles which are

\(^{63}\)An archetypical example of a non-distributive predicate is ‘scattered about the yard’: after a November rain, I set foot outside and observe that some mushrooms have grown and are ‘scattered about the yard.’ This predicate is quantified plurally: it can only make sense if applied to a plurality and does not distribute to each individual mushroom, which by itself cannot be ‘scattered about the yard.’
composed of molecules which are composed of atoms..." etcetera—too will rely on plural quantification upon plural quantification upon plural quantification, ad infinitum. So, if the nihilist wants to preserve notions of compositional structure—which are necessary for the sciences—he must commit himself to an arsenal of unreasonably complicated predicates. Bennett argues effectively that committing to such difficult predicates convolutes the nihilist’s methodology (Bennett 57–61). There is no reason, then, to think then that the nihilist’s having a smaller ontology of objects will count as having a simpler methodology than that of the universalist, whose ontology contains many gerrymandered things but whose predicates are far simpler. It is a trade-off between simpler ontology of objects or simpler ontology of predicates: So, since neither side has a methodological advantage over the other, (1) applies to the compositional debate and simplicity of method cannot be a determining factor in favoring either the compositionalist or the nihilist.

So far so good—but, perhaps there may be some philosophical argument that can count as evidence in favor of one side of the debate. Bennett holds that (3)—“no philosophical argument acts as evidence sufficient for selecting one party’s theory over the other’s”—applies because she sees there to be no adequate argument to count as such evidence. To substantiate this claim, Bennett shows that three prominent arguments against the mereological compositionalist can be equally leveraged against the nihilist and thus still do not favor either side. This paper will only focus on one: “The Problem of the Many,” conceived by the mereological nihilist Peter Unger.

Unger’s argument expresses a concern that when the mereological universalist admits to one composite object of a kind $F$, he must then also admit to a finite but unreasonably large number of slightly different objects of the same kind $F$ that nearly spatiotemporally overlap. Unger holds that because assuming unrestricted composition leads to the conclusion that any one composite object must always give rise to many, then composition must not be the principle by which some (alleged) composite object counts as ‘one object.’ To Unger, some other linguistic intuition provides for the fact that (alleged) composite objects may be counted (i.e., numbered); as such, the compositionalist’s views are unneeded and even incoherent. The proof goes something along these lines: Assuming unrestricted composition, consider a rock,
which we shall call ‘Rocky.’ Rocky is made up of molecular parts. But, surely if Rocky were to lose a molecule \( m \), we should still have a rock before us, though this new rock is not Rocky. Let us call this new rock ‘Rocky-minus-\( m \).’ Surely, a duplicate Rocky-minus-\( m \) (i.e., some other object that shares all the same perfectly natural intrinsic properties\(^{65}\) as Rocky-minus-\( m \)) in other contexts will count as being a rock. So, we must say now that we have at least two rocks before us, Rocky and Rocky-minus-\( m \); continue this procedure with any of Rocky’s other parts and suddenly we have a very large number of rocks indeed. This is a problem for the compositionalist, as his theory seems to contradict the linguistic intuitions we rely on for counting objects. Unger find this an adequate reason to favor mereological nihilism (Sider [2001] 358; Bennett 66-68).

Bennett is not impressed. It is not enough for the nihilist to fold his arms and claim that the compositionalist loses this battle; if the nihilist is to declare a genuine victory here, he must provide some sort of explanation for the aforementioned linguistic intuition that gives us the ability to count (alleged) composite objects. Bennett then proceeds to evaluate how the nihilist might attempt to do so. It turns out that when the nihilist attempts to capture this linguistic intuition, parallel logical reasoning to that of the Problem of the Many argument reapplies: Let us once again consider Rocky, who now might be defined by relation and properties of some simples. If one were to ask, “How

---

\(^{64}\)Why choose ‘a rock’? A rock is an example of an object for which our linguistic intuitions hold that its being a rock seems to involve nothing more than its instantiating some material/spatiotemporal properties. This is opposed to things like ‘a chair’ whose being a chair might be argued to involve specific essential properties involving origins, causal relations or intentions like made by beings for the purpose of being sat upon, etcetera (Sider [2001] 361).

\(^{65}\)These terms ‘duplicate,’ ‘perfectly natural’ and ‘intrinsic’ are borrowed from David Lewis. Perfectly natural properties are those most fundamental properties on which all other properties depend and from which all other properties may be defined. The sum of all perfectly natural properties and relations, then, act as a sufficient minimal supervenience basis for the whole of reality. Naturalness is, then, a primitive undefined concept. Intrinsic properties, which are commonly defined as those properties which may be instantiated by one individual even in the case that no other individuals exist (i.e., properties whose instantiation relies only on the existence of the individual it is instantiated within, now may then be defined in Lewis’s terms; a property is intrinsic iff: necessarily when \( x \) instantiates it, then any duplicate of \( x \) instantiates it (Lewis 343-351, 355-361; Sider [2001] 359).
many rocks (as alleged composites) do we have before us?” the nihilist still wants to be able to answer, “One.” ‘One,’ then, refers to some plural simples instantiating some non-distributive property being arranged-rockwise. Since only claims about simples are true, then being arranged rockwise must supervene on the properties and relations of the simples that (allegedly) compose the (alleged) composite: ‘a rock.’ But, if Rocky’s simples and their instantiated properties and relations are not to be the only way that some simples can count as ‘an (alleged) rock,’ then some certain small changes in the properties and relations of the simples that (allegedly) compose Rocky cannot be the deciding factor as to whether the plurally quantified non-distributed property being arranged-rockwise is instantiated. It follows that if instantiations of arranged-Rockywise and arranged-Rocky-minus-m-wise are both to count as instantiations of arranged-rockwise, then the nihilist must count at least two (alleged) composite objects. This problem occurs whenever the nihilist uses any instantiation of arranged-F-wise to count as an (alleged) composite object of the kind F. So, the nihilist is confronted with the same problem as the compositionalist—that is, just as the compositionalist must admit the existence of many nearly overlapping rocks, the nihilist is similarly committed to the existence of many nearly identical instantiations of arranged-rockwise (Bennett 66-68).

It seems that Bennett has argued that the Problem of the Many favors neither side of the compositionalist debate. This is because Bennett seems to have shown that both sides fall into the same sort of inconsistencies with our linguistic intuition involved in counting composite objects (whether those composite objects be alleged or actual). But, I think that there is a way that the compositionalist might be able to avoid these inconsistencies by defining the property of being some composite object \( \phi \) as an extrinsic property equivalent to a conjunction of maximal properties and intrinsic properties. Furthermore, since I believe that the nihilist will run into trouble making a parallel move using simples-talk, we have a reason to favor the compositionalist’s theory, and consequently a reason to believe that the compositionalist debate is thus one worth having.

Theodore Sider, in his paper ‘Maximality and Intrinsic Properties,’ identifies the instantiation of maximal properties as a way in which compositionalist theories might represent our linguistic intuition for counting composite objects. If Rocky is to count as only one rock, and not many, there must be a way to discredit Rocky-minus-\( m \) and all
other rock-ish proper parts of Rocky from also counting as rocks. The problem is: a duplicate of any the proper part of Rocky, when isolated in a world only inhabited by it, must still count as being a rock. So, how are we to define being a rock so that it may be false of Rocky-minus-m when he is part of Rocky, yet true of a lonely Rocky-minus-m duplicate in an otherwise empty world? Sider’s answer to this problem is to express being a rock as a conjunctive maximal property. Sider states, “A property, F, is maximal property iff, roughly, large [proper] parts of an F are not themselves Fs” (Sider [2001] 357). Thus, instantiating a maximal property involves not being a proper part of some greater thing of the same kind (or substance). Rocky instantiates being a rock because it is not contained in any larger rock-ish object; likewise Rocky-minus-m does not instantiate being a rock because it is contained in the larger rock-ish Rocky. In such a way, because instantiating being a rock involves things outside of the object in question, being a rock is an extrinsic property; but, being a rock still involves an intrinsic property as well, as the object must be rock-ish in order to be a rock (this prevents objects like ‘water not touching rock’ from counting as a rock). Thus, most of the nearly overlapping objects that Unger claimed that the compositionalist must count as rocks merely instantiate being rock-ish, but only Rocky, who is not the proper part of a larger rock-ish composite, instantiates being a rock. It is instantiation of these sorts of properties—being a P iff it is (intrinsic) P-ish and (maximal, extrinsic) not a part of any larger P-ish object—that the mereological universalist uses to capture the linguistic intuitions used in counting objects (Sider [2001] 357-8). The compositionalist can count the number of rocks he sees by virtue of counting the number of things that instantiate the maximal property of ‘being rock-ish and not contained in a larger rock-ish object.’

Sider has shown an advantage to committing to a thesis of compositionalism—that is, by using compositionalist ideas in the definition of properties like being a rock, we can explain what these properties consist in and how we can count the number of objects that instantiate such properties. The compositionalist thesis has told us a great deal about the nature of certain properties, as well as how we can count—and these seem to be good reasons for accepting this thesis. Can the nihilist mimic this move? I think not, as the definition of a maximal property requires certain mereological assumptions that the
nihilist denies. If I am correct, then it seems the compositionalist has a philosophical advantage over the nihilist. Nihilism is based entirely on the denial of objects with proper parts. This is what the nihilist means when he denies that there are fusions or composite objects. But, maximality relies entirely on the proper part definition, or more specifically, the transitivity and antisymmetry axioms of mereology (see (t) and (a) on page 2). But, might there be a way to do the work of these mereological axioms using simples-talk? Might relations between instantiations of non-distributive properties over plural simples be used to capture the intuitions that give rise to maximal properties? To me, it seems impossible. Relating instantiations over plural simples to other instantiations over plural simples might seem like a promising route, but in the end the nihilist will have to eventually accept some sort of mereological thesis in order for these complicated relations over plurals to capture what maximality does. Perhaps, there is some hope for the nihilist, but it seems to require some new logical constructions and quantifiers, and figuring out how to do that is the nihilist’s burden. Until then, it seems that the compositionalist has the upper hand in the compositional debate, and if that is so, then there is justification for believing the debate is genuinely substantial.

In addition to this, I wish to present one more argument that (yet more surely) favors the compositionalist over the nihilist. This argument rests on the assumption that it is possible that the world be such that objects are composed of atomless gunk. Gunk theory, first explicated by David Lewis, is the claim that all objects have (and are fusions of) their proper parts. So, if something is made of gunk then it is infinitely divisible—nowhere are mereological atoms to be found. If gunk is possible, then it is not necessarily the case that simples exist. This is entirely consistent with compositionalist theories, as objects are defined as the fusion of their proper parts. The nuance, though, is that gunky objects necessarily have infinite proper parts. On the other hand, nihilism, which relies on the existence of simples, runs into trouble. Just

---

I do not argue that there necessarily cannot be a way to define properties like being a $P$ and for grounding our faculty of counting that does not rely on maximal properties and may be available to the nihilist. But until the nihilist explains these, or explains why explaining these is of no philosophical interest, the compositionalist seems to have the advantage.
in case our world is gunky, the nihilist, who denies all composite objects, is left with an empty ontology; this is absurd. Moreover, even if our world is not such that objects are gunky, this seems to be a contingent fact; since the nihilist claims that necessarily no parts have proper parts, the nihilist’s language is incapable of expressing possible worlds in which objects are gunky. But, if we accept that it is possible that objects are gunky, then there are possible worlds in which objects are gunky, and the nihilist’s claim that necessarily no parts have proper parts causes a contradiction. It follows then that mereological nihilism is a deficient theory (Russel, 1-3; Sider [1993] 285-289).

Bennett mentions in footnote 8 of her paper that she will not address the possibility of gunky objects in her theory. I imagine that her motivation for leaving gunky objects out of the picture originates in her wishing to address the difference-minimizing spotlight of the compositional debate. There is a sense in which when the compositionalist downplays his high-ontology he might want say that composite objects (fusions) are nothing more or above the simples instantiating some plurally quantified property. But, as mentioned earlier in this paper, I think that it is a mistake to think that the compositionalist must admit to simples in order to downplay his ontology. That is—the compositionalist can difference-minimize while still maintaining neutrality as to whether simples exist or whether objects are gunky. He might express his downplaying as: “If simples exist, then the composite object ‘table’ is nothing more than simples arranged-table wise; or, if simples dont exist, then the composite table is nothing more than gunky stuff over there instantiating being a table (or having a table-ish fusion and not being a part of a larger table-ish part, etcetera).” This disjunction captures all the compositionalist needs to insofar as he attempts to portray what intuitive advantages the nihilist’s theory enjoys, while simultaneously being open to the possibility of gunky objects. So, since even the difference-minimizing compositionalist can allow for the possibility of gunky objects, then Bennett must acknowledge arguments based on this possibility as relevant to the compositionalist debate.

In summary, I have shown two arguments that act as evidence in favor of the compositionalist and are consistent with Bennett’s scope of difference-minimizing. Since these arguments do act as evidence in favor of one side of the compositionalist debate, Bennett’s claim that (3) applies to the compositionalist debate (i.e., no philosophical arguments favor either side) is false. It follows, then, that the compositional debate is a fruitful metaphysical labor after all. We might take a lesson from
this enquiry that mereological axioms—in that they give rise to the intelligibility of maximal properties—may be essential to the foundation of certain linguistic intuitions, like that of counting, or recognizing objects as having number, unitariness, etcetera. Moreover, accepting these axioms opens a framework through which we can explore the possibility of gunk. If we can make sense of mereological gunk, perhaps there are adventures to be had in assessing gunk in other spatial and temporal dimensions, like topology, measure and duration.
References

Bennett, Karen. “Composition, Colocation, and Metaontology.” (In Chalmers: 38-76)


There has been much debate among decision theorists about which solution concept is normatively superior or captures a better insight into rationality. This debate presupposes two key pieces of machinery: subjective probabilities and preference orderings. These two pieces of machinery, in turn, depend on a third formalism: outcomes. We take outcomes for granted because there are many cases where it seems clear what types of considerations are relevant in a decision problem. For example, it seems clear that the relevant considerations for the question of whether I should bring an umbrella are [it will rain] or [it will not rain]. Then I can figure out how many different classes of rainfall I can distinguish between, perhaps in increments of inches, and assign each a probability based on how likely I think each state is to obtain. From this, it is a matter of using the expected utility (EU) theorem and calculating which option has the highest “state+act” score. The problem with this picture is that it is not what people do in practice. Because of the incongruity between what people do and what decision theory prescribes, there is a tension wherein we must either label people irrational or seriously reconsider decision theory as a viable normative theory as well as an explanatory one. In this paper, I will do the following:

1. Present a challenge to decision theory from Jose Bermudez based on people’s shifting preferences in response to different descriptions of the same problem.

2. Consider an alternate framework from Frederic Schick that adds descriptive realism and then explain two shortcomings of Schick’s proposal stemming from the imprecision in its algorithm for solving EU calculations, which leaves it without a normative basis.

3. Attempt to strengthen and extend Schick’s account by appealing to the values we use to describe counterfactual situations and how these values can provide a normative basis for decision theory, at least for cases of materially identical propositions.

Background

In the 70s, Tversky and Kahneman conducted a series of experiments exploring whether humans act as rational agents. According to Kahneman and Tversky, a rational agent has to obey four axioms under standard decision theory: cancellation, transitivity, dominance, and
invariance. The axiom which Bermudez focuses, with the aim of undermining decision theory, is the Invariance Principle (IP) axiom, which states:

Invariance Principle (IP): “Different representations of the same choice problem should yield the same preference.” (Kahneman and Tversky 1986)

One study that Bermudez uses to defend his objection that decision cannot be both predictive and normative involves framing effects on choice of treatment for cancer patients. The study was administered to patients, business students, and doctors, all of which chose similarly after being given the following prompt:

**Survival frame**

Surgery: Of 100 people having surgery 90 live through the post-operative period, 68 are alive at the end of the first year and 34 are alive at the end of five years.

Radiation Therapy: Of 100 people having radiation therapy all live through the treatment, 77 are alive at the end of one year and 22 are alive at the end of five years.

**Mortality frame**

Surgery: Of 100 people having surgery 10 die during surgery or the post-operative period, 32 die by the end of the first year and 66 die by the end of five years.

Radiation Therapy: Of 100 people having radiation therapy, none die during treatment, 23 die by the end of one year and 78 die by the end of five years.

Among the results was that only 18% of the subjects chose radiation in the survival frame while 44% of people with similar backgrounds, age, ethnicity, etc. choose so differently is hard to dismiss. The reason this result is troubling is because we would expect that whatever a doctor believes about a treatment, her opinion would not vary across different phrasings with equivalent referents. It is in light of these kinds of inconsistency findings that Bermudez criticizes decision theory’s ability to be predictive and normative.
Bermudez’s Objection

If descriptions make a significant difference to an agent’s subjective probabilities and utilities, decision theory is not going to be a viable theory that straddles the roles in the way we want it to. To formalize the notion of IP, we will make use of the concept of co-reporting, which can be broken down into two types: material and logical.

Co-reportive_{material}: “Two propositions are co-reportive just if they are materially equivalent (i.e. they coincide in truth value) and their material equivalence is a function of independent physical self-identities.” (Bermudez 83)

Co-reportive_{logical}: Two propositions are co-reportive just if they are logically equivalent where their logical equivalence is a function of their identical truth conditions. (Bermudez 84)

The difference between the two will become apparent when we consider two examples by Frederic Schick in which an agent has different utilities for the same event and compare those to the surgery/radiation example. The first example Schick uses is an account by George Orwell of his own experience as a soldier in WWI. He describes himself on the lookout for fascists to shoot when suddenly a man with his pants down comes stumbling across the battlefield. Orwell realizes this man is a fascist, but is torn because seeing the fascist with his pants down reminded him that the man is also a fellow creature. Orwell has the following preference orderings:

shoot fascist < don’t shoot fascist

don’t shoot the fellow creature > shoot the fellow creature

Orwell knows that the proposition “shoot the fascist” refers to the same outcome as “shoot the fellow creature,” but he still assigns different utilities to the two propositions. He decides to refrain from shooting, but does not have a solid justification for his action in terms of his preferences. The second example Schick mentions is the case of the doctor who is trying to save a patient when he learns that the patient is actually a vicious convicted murderer found guilty of his crimes. The doctor has a low utility from losing the patient in surgery, but also a high utility from losing the murderer in surgery. Again, the agent
knows the propositions “I save the patient” and “I don’t save the murderer” have the same extension in the world. This is to say that the same physical thing in the world determines the truth-value of both propositions completely, yet the doctor has these preference orderings.

From the point of view of decision theory, Orwell, the doctor, and the participants from the study are all irrational because they violate the invariance principle and, along with it, the axiomatic basis of the expected utility theorem. However, they violate the IP in different ways depending on the version of co-reporting we apply. The participants from the study are guilty of violating both the material and logical versions of co-reporting because the treatments are materially the same as well as have the same truth-value as a matter of necessity. Both versions are violated because saving 90% of the patients is the same as losing 10% of the patients yet there is a difference in response based on which number is presented to the subjects. On the other hand, Orwell and the doctor do not violate the logical version of the IP because it is not a matter of necessity that the fascist is the same as the fellow-creature or the patient is the same as the murderer. I will address this difference in violations later in the paper when I present my attempt at reconciling the rationality of Orwell and the Doctor with their seemingly irrational choices. Before I do that however, I will present Schick’s model for preserving the rational status of Orwell and the Doctor.

**Schick’s Proposal**

Schick believes that we need to give up the normative aspect of decision theory in favor of preserving its predictive/explanatory role. He violates the axioms of decision theory by introducing new machinery into expected utility calculations called understandings, which are roughly what Kahneman and Tversky call frames. In the Orwell example, Schick would say that there are at least two possible understandings, that of shooting a fellow creature and that of shooting a fascist. After identifying the possible understandings, the agent must then determine which understanding is “compelling.”

Compelling: A proposition is compelling just in case it presents a fact in a way that the agent understands it.

After an agent finds an understanding that is compelling, she can continue as usual with subjective probabilities and utilities, which de-
termine expected utilities. This new machinery is very powerful in accommodating a wide range of environmental effects on agents, which could influence what utilities are assigned. In fact, the evidence that framing affects reasoning as well as mental heuristics is too strong, in my opinion, to continue without considering framing effects. However, to do it at the cost of the normative force of the theory would be premature and Schick’s account has some gaps that are too large to overlook. In the next section, I will distinguish between a theory’s normative role and its predictive role, address why normativity in decision theory is desirable, consider the gaps in Schick’s account, and finally help Schick bridge some of his gaps by considering counterfactual situations and values.

**Predictive Theories versus Normative Theories**

Theories typically have at least one of two roles, or reasons why we use them. One role is to predict and organize a set of phenomena. For example, the theory of evolution is used to organize the varied physical evidence that organisms that adapt to a given environment tend to survive and have offspring while those that are not successful at adapting to their environment tend to die off, thereby explaining the phenomenon of speciation and extinction. We can use the theory of evolution to predict how future species will diverge in lineage or become extinct. However, we do not grant evolution a normative role. We do not endorse evolution as a way things should be in the world. In fact, we make various efforts to preserve endangered species because we, as a society, value diversity in our environment among other reasons. Other theories are used normatively. For example, Marxism is a descriptive theory of social classes and economic systems, but it also prescribes that we should act to bring about a certain state of the world. Marxism combines a model of the world with advice on how we should behave or think. Expected utility theory is a similarly structured theory in that it models how we actually make decisions and advises us how we should make decisions. For example, it claims that we know whether we prefer vanilla ice cream to chocolate ice cream and that we will order vanilla ice cream rather than chocolate if we prefer vanilla to chocolate, all things being equal. It will also claim that if we like vanilla twice as much as chocolate, then we would be indifferent between a 50/50 gamble of getting vanilla and getting chocolate with certainty. Expected utility
theory would tell us to watch for this result. In addition, EU would
tell us that if we like vanilla ice cream twice as much as chocolate, then
we should be indifferent between the gamble option and the certainty
option. If we paid someone a premium to get chocolate with certainty
over the 50/50 gamble for vanilla, we would be acting irrationally. We
would be acting contrary to what we want.

The real usefulness of a normative theory is best brought out by a
more urgent and real world example. Suppose an agent for the Center
for Disease Control had to choose whether to quarantine a town of
10,000 to stop the spread of a moderately harmful flu or send trucks
of retroviral shots to stop the infection. Quarantining the town
would result in millions of dollars of foregone profits and tax dollars to fund
the quarantine, but the action would stop the outbreak with certainty
while the retroviral shots would be much cheaper and most likely treat
enough folks that the threat of outbreak would be very small. The
decision, under standard decision theory, is entirely dependent on our
preferences across money, lives, and how much we are willing to pay
to prevent deaths. The intuition is that we typically make decisions
like this every day. We live in high-risk neighborhoods to save on rent
or purchase unhealthy foods that shorten our lives because we do not
want to spend time or money to provide ourselves with healthier food.
In each choice, we reveal a little bit about our preferences across our
safety and some other thing. Expected utility theory tells us how we
must act when we reveal these sets of preferences. In the example, the
CDC would estimate how much they would be willing to risk human
lives to save millions of dollars. If the risk of outbreak under the second
option is so small as to be insignificant compared to the risks we happily
take every day like driving a car, then the CDC agent would be better
off, according to our own societal preferences, using those tax dollars
and foregone profits elsewhere and sending the retroviral shots instead
of instating a quarantine. If he instated the quarantine, then we, as a
society, could hold him responsible for choosing the wrong action on
our behalf. We hold this person accountable because he did not choose
the outcome with the most beneficial effect. The reason we can hold
him accountable is because he should have chosen the shots, but we
need to ensure that decision theory gives a normative basis before we
allow ourselves the ability to blame someone on rational grounds.
Why Care About The Normative Basis Of Decision Theory?

The objection to decision theory that I want to focus on is that we systematically choose different options if the options are presented differently, thereby violating the invariance principle. The invariance principle is derived from the expected utility theorem. This theorem underlies the utility calculations of the type I noted in the introduction, such as whether to bring an umbrella. A breach of the axiomatic basis of decision theory as well as any principle that is derived from it, such as the invariance principle (IP), extinguishes the decision theory’s normative force. One might wonder why it is important to preserve the normative role as well as the predictive/explanatory role. I can give two related motivations.

First, everyone has been in a situation where the group they are a part of is deliberating what course of action to take, and the group ends up taking the non-optimal course of action even though you, as an agent, had the EU calculations to prove a different course should have been taken. In fact, you could have accused the group of acting irrationally in a substantive way. What you would mean by irrational would not be relative to some nominal difference. A stronger claim would be available to you to accuse the majority of acting against its own interests. Without the axiomatic basis upon which decision theory rests, that substantive accusation would not be available to you and instead would necessarily be relativized, but it seems that there are times when clearly there is an optimum choice for a group that is overlooked because of variation in description. In the example with the CDC agent, if the agent succumbed to the framing effects that afflicted the research participants I discussed in the background section, he would be acting directly in opposition to the people he is supposed to serve. On the other hand, if the public heard of the plan on the news and suddenly wanted to be quarantined because of a framing effect, then they blame the agent even though they were acting against their own preferences.

Second, an analogy can be drawn with a common feature of agents that endorse ethical theories. Moral agents seem to be able to hold an ethical theory as right or true and yet violate the tenants that they hold to be normative. However, they do not scrap the basis of their ethical theory and produce a different theory that attempts to account for actual behavior so that it is more realistic. Instead, they face up to
their shortcomings and strive to do better.

The alternative is to have a model of moral relativism, one that works really well at describing actions like killing, stealing, and torturing, but is useless at being able to tell us not to kill, steal from, and torture others. These two points give us a sense of what is at stake. A descriptive theory is nice. It can tell us what people will do which can have practical applications, but it does not do the kind of heavy lifting we sorely need—telling us what we should do.

**Pitfalls of Schick’s Proposal**

**Imprecise Procedure**

The first issue Schick must confront is he takes for granted that we know the understandings in an agent’s mind. There is no way for us to evaluate whether a certain understanding is the right one (i.e. most compelling) because understandings are not “shareable,” “transitive,” “communicable,” or “truth-evaluable” (Bermudez 87) as Bermudez interprets Schick. It would be hard for an outside agent evaluating Orwell to determine what mental states were available to him and how he saw the situation. This is the first kind of imprecision Schick must face. The second kind of imprecision that Schick must address is the procedure for determining when to find an understanding compelling. An example Bermudez cites from Schick is one involving a corrupt judge who sees two understandings available to him about his own actions. Understanding A is that he is receiving a bribe and understanding B is that he is performing a service for a fee. The judge knows that these two understandings are materially co-reportive yet he assigns different utilities to the two understandings. To make Bermudez’s objection clear, I will overlook the breach of the invariance principle for the moment and focus on the fact that the judge chooses to find B compelling just because it will take away any moral censure. It is not the case that he just sees the facts under a frame, but that he chooses a certain frame through which to see it. How can it be that the judge can do this? It seems very strange that an agent can determine what frame to find compelling based on what is most self-serving. The uncomfortable feeling we get when the judge chooses to look at his action as taking a payment for services rendered rather than taking a bribe cannot be explained by appealing to the invariance principle because we removed it from the theory when we decided to add understandings to our de-
cision theory machinery. Instead, Schick believes we can preserve the legitimacy of understandings by looking to morality as a guide for when to find an understanding compelling. To handle the judge case, Schick checks to see if the judge is being “sincere” or making some kind of good faith effort to choose the understanding he actually finds compelling. Clearly the judge both knows that the understanding of payment for services has a higher utility and chooses it for that reason, but puts on airs as if he actually finds the understanding compelling. This game of pretend is done just to avoid moral censure and the only reason we are able to detect it is because this is an idealization, so we stipulate the judge’s mental state. In the real world, Schick’s method of calling the judge out for being insincere would fall flat because we just could not tell, based solely on his decision, what is going on in his mind, which again is a result of understandings not being communicable. However, moral norms will serve as a basis for my proposed enhancement to Schick’s theory.

Breach of Normativity

The deeper issue with Schick’s account is that it removes the normative basis of decision theory in order to make strides for its predictive/explanatory dimension. I have mentioned above that the invariance principle is derived from the expected utility theorem and adding understandings violates the IP which means that the EU theorem is violated (Bermudez 91). We might consider trying to rewrite the axioms of decision theory to accommodate understandings, but I do not think this would be a fruitful project. Normativity is principally about what people should do. An “ought” or “should” claim must be objective if it is to be a substantive claim. I do not mean there must be some reality or physical object which makes an ought claim true. What I do mean is that one can present an ought claim to another person with the assurance that they will see what you mean. The claim will not seem too strange and hopefully will be as salient for them as it is for you. The problem with understandings is that they are mental states that, by definition, are not communicable or shareable. Schick postulates that the machinery he uses allows for violations of the IP, which violates the EU. We are able to analyze the judge case for sincerity because we pretend we are in his head and idealize his cognitions; real applications would be difficult to handle. Because of this difficulty, which stems from Schick’s own definition of understandings, it is doubtful that a reworked axiomatic basis for decision involving understandings could
be produced. This is not to say that Schick is completely off the mark. As it stands, decision theory assumes a straightforward Sidgewickean utilitarian conception of morality and I think there is room for other ethical theories to make their presence felt. The next section presents one way a value or virtue based account of ethics can be inserted into decision theory, which may help capture a part of ethics that, up until now, has been overlooked in decision theory.

A Different Way of Saving Orwell and the Doctor

I have presented two ways of modeling what is going on in the head of Orwell and the Doctor. The first, pre-theoretic pass is that they have two competing descriptions of the same physical object in their heads. Orwell asks himself whether the thing in front of him is a fascist or a fellow creature while the Doctor asks himself whether the thing in front of him is a patient or a murderer, but that these are just an error in the proper symbolic manipulations. The second is that there are two competing frames or ways of parsing reality that move beyond a difference in which signs are used to refer to an object. Both agents go back and forth on which details to attend to because their active interpretation finds different sets of facts salient. I want to look at another way to explain why Orwell and the Doctor might be having such a hard time figuring out what to do in their respective decision problems. This new way draws on the second model, but gives it a normative framework. My proposal for examining their situations entails a consideration of how moral values are tied to descriptions of possible worlds. First, I want to spell out a few background assumptions about the possible worlds framework I want to use. When I say possible worlds, I do not mean parallel worlds with a physical existence that we cannot reach, as David Lewis would suggest \(^1\). Instead, I want to think of possible worlds as ways the world might have been. These ways I take to be primitive and sets of these primitives are what make up propositions.

---

\(^1\)It is not clear that assuming a Lewisian framework would affect the outcome of my argument.
For example, the proposition (1) I shoot the fellow creature is made up of a set of possible worlds wherein there is this fellow creature in front of me and I shoot him. The proposition (2) I shoot the fascist is made up of a subset of the possible worlds which make up (1). In the doctor case, the situation is a bit different. The propositions (3) I save the murderer and (4) I save the patient have at least one overlapping world, but one is not a subset of the other because this same man may not have come to this particular doctor (possibly not patient) for treatment and he also may not have committed the crime in question (possibly not murderer). We do not know which possible world we are in, but, based on the clues around us, we know which worlds we are not in. Orwell is in the same epistemic boat as us, he does not know exactly which possible world he is in because he does not have all the facts that would identify a unique world. He does know that he can exclude a set of worlds from the pool of candidate worlds he might be in, namely the worlds where the fellow creature is not a fascist. The Doctor can also exclude some possible worlds from the list of places he might be, the worlds where the patient and the murderer are different people. We might wonder why this should matter. We know we are not in those worlds where the two terms come apart, but I want to argue that Orwell and the Doctor might take into account these possible worlds, even though they have no credence (read: belief) that they are in them. I do not have a formal argument for why we should take into account possible worlds which we know we can exclude from the possible places we are located in logical space. Instead, I have examples that seem to point to a behavior we constantly engage in, thereby either making a large swath of human behavior very strange or indicating that we need to modify decision theory to accommodate this behavior.

**Fragility**

Fragility is a property based entirely on whether or not the bearer of this property shatters easily. It is a property based on an answer to the modal question “what if X amount of force were applied to this object?” The possible worlds where it shatters or not determines the property of fragility in the actual world even though we know we are not in the world where the object shatters. In this case, possible worlds (i.e., ways the world might be) determine a property in the actual world.
Near-death Experiences

Many people who come to the brink of death make life-altering decisions such as to quit smoking, pursue a career they have dreamed of, reveal secrets, etc. The reason that people make such choices is not because they almost died in this world, but because they realize how close they were to the possible world where they would have died. They take into account what could have happened and use that thought experiment data to inform their values in the actual world. After all, almost getting hit by a train does not physical change a person, apart from the temporary fight-or-flight response they would have. If you are willing to grant that decision theory should take into account non-actual phenomena, then I only need to argue for one further claim: the language we use is heavily influenced by our values. In our everyday discussions, the words we choose to refer to a topic may imply our political views, our hobbies, and origins. This is uncontroversial. The interesting discovery that comes from cognitive science and linguistics is that our choice of words actually indicates which values we hold toward a topic. For example, cognitive linguist George Lakoff offers an analysis of the metaphors that are used when talking about tax cuts from the right and left of the political spectrum. The right frames the issue by calling it tax relief, which evokes the metaphor of taxes as an affliction or disease. The value they hold is that if you make money, you should be entitled to keep it. The left frames taxes as paying one’s fair share for the opportunity to make profit by using the various public goods in this country; tax cuts are thought of as corporate welfare (which incidentally is a poor frame to use for several reasons). Very different value systems are evoked by these different frames and the words we use to describe a bill with tax cuts will be determined by which value system we endorse. One recent frame contest was “Obamacare” versus “Patient Protection and Affordable Care Act.” The right is focused on government involvement in healthcare with an emphasis on strict self-reliance and the left is focused on making sure sick people can see a doctor. Different values: self-reliance versus safeguarding life. The Schick cases make it clear that we are not passive participants in value-based language. Agents, at times without knowing it, describe themselves and their choices to indicate which values they hold as most important. If the doctor frames himself as saving a patient rather than taking revenge on a murderer, then we know he values compassion and life over retribution and anger-satiation. We can analyze this value determination by looking at the possible worlds
the frames evoke. First, we identify the two statements that are materially (but not logically) co-reporteive for which the agent has two distinct utility values. In this case, it is (3) and (4). Next, we identify the sets of possible worlds of which each proposition consists. For each set of possible worlds, we see what the preference ordering for the action suggested by the proposition is in relation to the act of doing nothing. In the possible worlds where the man in front of the Doctor is a murderer, but is not his patient, does the Doctor prefer saving him to doing nothing? The answer to that question will determine part of his value system. After that, we look to the possible worlds where the person in front of the Doctor is his patient but not a murderer and see if the Doctor prefers saving the patient to doing nothing. Now we have two clear preferences, determined by what the Doctor will do in each set of possible worlds. In the first case, it seems that if the Doctor would kill the murderer or not save the person to whom he has no legal obligation to save, he values punishment or retribution. In the second case, if the Doctor saves the patient, he, at the very least, shows that he values life and perhaps duty. The doctor must establish what each action would mean in terms of endorsing a value. Preference ordering might not be a plausible interpretation of this thought experiment, but rather it might be described better as a weighting system. You might have a utility function that is not merely a continuous line where all of your preferences are arranged on one axis, but rather several lines that order actions with respect to certain values and then combine them based on which value you think is most important. Thus, the Doctor would have to figure out which value he thinks is more important to endorse: punishment or preservation-of-life/duty. The answer to the question of what to do in the actual world will depend on the answer of which possible worlds to endorse. What one should do now is partly determined by the ways things might have turned out, just like in the case of fragility and near-death experiences. The possible worlds we focus our attention on are shaped by our choice of language, but this choice is heavily influenced by our values. My goal has been to explain how we might retain normativity, but still accommodate framing effects into decision theory. Orwell neither has inconsistent values nor is irrational. Instead, I would argue that when he calmly considered killing the fascist, he discovered what his values were. He discovered these values by considering the different ways to frame the situation (fascist or fellow-creature), the possible worlds that correspond to these frames, and ultimately decided on the language to describe the situation based
on which possible worlds conveyed the values he agreed with more. If in the future, he starts shooting fascists, but claimed to share the same values as before, then I believe we would have grounds to call into question either his rationality or the well-orderedness of his preferences. One possible objection might be that we have no idea why the Doctor might save the patient when the patient is not a murderer. That no unique value is picked out. The strength of this objection comes from the doubt we may have that we know what value drove the Doctor to save the patient. Surely saving the patient could have just been the most expedient thing to do because the Doctor would have gone to jail if he had not saved the patient. This is a valid concern, so let me give a plausible account of how to resolve this problem. Suppose we are unsure why the Doctor reports that he would save the patient in the set of possible worlds where the man in front of him is not a murderer. There are several explanations we can come up with that all seem to be able to explain his actions. What we could do is repeat the steps I provided above to test for each co-reportive pair of alternate explanations. For example, we could present the Doctor with the scenario that a man is in front of him and needs saving. Then we could give him two further understandings to consider. The further understandings do not have to be exhaustive, but must represent the case that would cast the most doubt on which value is endorsed by the Doctor. The following serve as examples of these kinds of cases:

(5) The man is a legal citizen of my tribe/nation, and thus I would be morally and legally censured if I did not save him.

(6) The man is not a legal citizen of my tribe/nation, and thus I would not be legally censured if I did not save him.

Then we could see what he would do in each set of possible worlds. If he saves the man regardless of the legal/societal obligations placed on him by the man’s creed, then we can rule out obligation as an alternate explanation for why the Doctor saves the patient. This procedure could be repeated until several pairwise comparisons of sets of possible worlds come up heterogeneous in reply. At this point, what the Doctor reports as his choice in the main decision problem will determine a set of values that he holds such as saving a man regardless of his creed which sounds very much like the concept of compassion or basic consideration of another’s humanity.
All Is Not Lost For Schick

This proposal helps Schick out of a tight spot because now the move to consider morality as a guide to compellingness is clarified. The machinery is explicit, communicable, and has clear guidelines for how ethics can help decide which frame to find compelling. Take, for example, the corrupt judge. The two relevant propositions are (7) I take a bribe and (8) I provide a service for a fee rendered. The two propositions come apart because there are possible worlds where a judge provides a service for a fee rendered which is not a bribe. One can then see which the judge prefers more: receiving a bribe or receiving a fee for a service rendered. The judge can lie to us about his preferences, but this lie is now explicit. It is no longer confounded with wanting a fee for a service rendered because he might have preferred to perform a service to the bribe from our point of view. Finding a reliable lie-detecting system is a practical hurdle, but under my analysis he at least commits to a set of worlds clearly distinguished from the fee-for-service worlds. A good analogy from the field of ethical theories is the trolley problem and the doctrine of double effect. The setup of the trolley problem is that there are five people tied to a track and a trolley is racing towards them. You have the power to flip a switch to divert the trolley to another track, which has one person tied to it. The situation can be understood in at least two ways, either you are saving five lives or taking one life. If you flip the switch, you can pick the following understanding: you are saving the five and killing one is an unintended side effect that happens to obtain from saving five lives. Now pretend I add a loop to the track. If you do nothing, the five will die and then the one will die. If you pull the switch, only the one will die because he is so heavy set that he will stop the trolley. Under this scenario, you can no longer say that the killing of the one is a side effect of saving the five because you are counting on the one’s death to save the five. It reveals just how far you would go to save the five. Before, we were not sure if the agent pulling the lever was using the one person on the track as a means for saving the five or if the one person’s death was an unintended side effect. What I have effectively done is add a loop to the track and put pressure on the agent to save the five by killing the one. The analogy works like this: the judge can choose to receive the bribe/payment-for-service in the actual world and we would legitimately not be able to accuse him of taking a bribe if he did not see it that way because he might have seen it as service for payment and the illegal nature of the transaction was just an unhappy side effect. However, if we asked him
to consider the possible world where he performs honest work and the possible world where he is just receiving a bribe for not doing his judicious duties, then we finally see what his preferences are. If he prefers being paid for honest work to doing nothing and nothing to taking a bribe, then we can accuse him of not obeying his own values because in the actual world he decided to take the bribe. If he prefers a bribe to doing nothing and doing nothing to honest work, then we can expose his lie that he sees the bribe as fees for services rendered. Remember, the judge has two sets of possible worlds and has to weigh which value is more important, as represented by each set of possible worlds. If he chooses to take the bribe/payment-for-services, then we know that the judge thinks financial gain is more important than say doing one’s duty. Furthermore, looking at possible worlds has eliminated the need for understandings to be mental states that are not communicable or shareable because they are a logical construction. Modal propositions can be represented with modal logic, which acts as a conduit for information between two agents that allows one agent to evaluate whether the other is correctly finding a particular way of looking at the world compelling. In addition, we could empirically test this new machinery by asking subjects to tell us how they weigh their values and then give them scenarios that would record how they respond to a series of moral decision problems involving those values. If this theory predicts well, it can also help decision theory improve its predictive/explanatory power. Of course, this would have to be an empirical finding rather than an a priori thought experiment.

Conclusion

I have outlined a new piece of machinery that can be added to decision theory, which utilizes possible worlds that an agent knows he is not in to make decisions in the actual world. The three criticisms leveled against Schick earlier have been addressed. First, communicability in the content of mental states is less of an issue as sets of worlds can be conveyed with modal logic. The issue has not been put to rest because communicating modal logic is not a natural process, so this part of the theory may be objected to on the grounds that we might not engage in this practice. As I have mentioned above, this is an empirical matter. Second, I have explained the procedure by which one can find a set of possible worlds compelling. Third, because the first two objections have been answered, the third objection about preserving normativity
is also answered. The decision of what to find compelling is no longer ours to make but rather a function of what we prefer which is based on our values and how we describe the world. Decision theory allows us to hold any preferences we like as long as they are well ordered. If values are the substratum of our preferences, then we can also hold whatever values we like as long as they are well-ordered. The debate about whether there is a specific set of correct values to endorse and which those would be I leave for another to address, however I do assume that there are values or mutually compatible sets of values that are correct to hold. If there are no normatively right values for us to hold, then my argument would fail. Lastly, I want to reiterate that this does not explain violations of both versions of the IP, but rather only apparent violations of the materially co-reportive version. The logically co-reportive version of the IP cannot be explained using the schema I have laid out because logically co-reportive statements are equivalent across all possible worlds, making it impossible to break apart the statements into the sets of discrete possible worlds that make them up since the very same worlds would make up both statements. On the other hand, materially co-reportive statements can be broken up by what values each statement espouses. A method for deciding between them now exists, which holds people responsible for aligning their actions with their values or risk suffering moral censure.

References


Notes on Contributors

Justin Bruner graduated in the spring of 2010 from UC Berkeley with a degree in Economics (high honors) and a minor in Philosophy. His academic interests include decision theory, the philosophy of economics, and evolutionary game theory. He is currently pursuing a Ph.D. in philosophy at UC Irvine in the Department of Logic and Philosophy of Science.

Matthew Clark graduated from Berkeley with a B.A. in Philosophy in the Spring of 2010; he will begin an M.A. program at the University of Chicago in the fall of 2011. His interests in philosophy are still quite general, but when pressed, states his interest in non-conceptual or non-rule-based cognition and action, with a serious (and not wholly unrelated) side interest in feelings and moods. These interests began with Kant’s third Critique, and later brought him to many others in the history of philosophy, including (but not limited to) Aristotle, Heidegger, and Wittgenstein. He would like to thank Laura Davis, Chuck Goldhaber, Adam Blazej, Eugene Chislenko, Janum Sethi, and Professor Hannah Ginsborg for their thoughts, contributions, and encouragements in writing about the ugly in Kant.

Charles Goldhaber will graduate UC Berkeley with a double major in Philosophy and Japanese in summer 2011. His main interests in philosophy are in philosophy of action, ethics and phenomenology, but he admits to having a soft-spot for mereology. He is currently concerned with the problem of how to characterize practical intelligence in light of Aristotle’s account of phronēsis in the Nicomachean Ethics. He believes that philosophy is foremost a practical art, used for overcoming those obstacles that we lay ahead of ourselves when we forget to think. Philosophy, he thinks, is a defense against the covering-up of the meaningful by the mundane.

Eduard Meleshinsky graduated with a double major in philosophy and psychology. His interests include foundational issues in decision theory, behavioral economics, and ethical theory; usually all at once. He would like to thank the professors and peers that have
challenged him these past four years to think with clarity, approach problems creatively, and question assumptions incessantly.