On How to Build a Moral Machine

Paul Bello PAUL.BELLO@NAVY.MIL
Human & Bioengineered Systems Division - Code 341, Office of Naval Research, 875 N. Randolph St., Arlington, VA 22203 USA

Selmer Bringsjord SELMER@RPI.EDU
Depts. of Cognitive Science, Computer Science & the Lally School of Management, Rensselaer Polytechnic Institute, Troy, NY 12180 USA

Abstract

Herein we make a plea to machine ethicists for the inclusion of constraints on their theories consistent with empirical data on human moral cognition. As philosophers, we clearly lack widely accepted solutions to issues regarding the existence of free will, the nature of persons and firm conditions on moral agency/patienthood; all of which are indispensable concepts to be deployed by any machine able to make moral judgments. No agreement seems forthcoming on these matters, and we don't hold out hope for machines that can both always do the right thing (on some general ethic) and produce explanations for its behavior that would be understandable to a human confederate. Our tentative solution involves understanding the *folk concepts* associated with our moral intuitions regarding these matters, and how they might be dependent upon the nature of human cognitive architecture. It is in this spirit that we begin to explore the complexities inherent in human moral judgment via computational theories of the human cognitive architecture, rather than under the extreme constraints imposed by rational-actor models assumed throughout much of the literature on philosophical ethics. After discussing the various advantages and challenges of taking this particular perspective on the development of artificial moral agents, we computationally explore a case study of human intuitions about the self and causal responsibility. We hypothesize that a significant portion of the variance in reported intuitions for this case might be explained by appeal to an interplay between the human ability to *mindread* and to the way that knowledge is organized conceptually in the cognitive system. In the present paper, we build on a pre-existing computational model of mindreading (Bello et al., 2007) by adding constraints related to psychological distance (Trope & Liberman, 2010), a well-established psychological theory of conceptual organization. Our initial results suggest that studies of folk concepts involved in moral intuitions lead us to an enriched understanding of cognitive architecture and a more systematic method for interpreting the data generated by such studies.

1. Introduction and Plan for the Paper

This article makes a two-pronged argument for the further influence of psychological findings on the development of machine ethics. The first part of our argument presented throughout section 2 lays out the case against adopting traditional ethical theory as a foundation upon which to build moral machines. We address a number of issues pertaining to the apparent naturalization of ethics

that our particular approach to developing moral machines seems to entail. The second part of our argument, laid out in section 3, centers on a commitment to doing machine ethics via computational cognitive architectures. Computational cognitive architectures are comprised of data structures and algorithms that purport to be constrained by our best psychological theories of perception, mental representation and inferential powers. In particular, we argue that the capacity for moral machines to *mindread*, that is to ascribe a rich set of mental states to either itself or other agents is a requirement for said machine to be a moral agent.

Many philosophical discussions of mindreading are often about competence-related issues surrounding cognitive structure of core mental states such as belief and desire; they are also fixated upon the nature of the mental-state ascription process. Unfortunately, much of this theorizing is derivative on tightly-controlled psychological experiments that rarely resemble the kind of real-world situations within which we routinely make moral judgments. Furthermore, the general tendency for philosophers to focus on beliefs, desires, and intentions has come at the expense of focusing on choice, agency, self, and other notions that seem to be intimately tied up in moral evaluation. We make an argument in section 4 for utilizing the latest empirical data produced by *experimental philosophers*, who have devoted substantial energy to charting the boundaries of both primitive mental states and many of the other folk concepts involved in moral judgment.

In sections 5 and 6, we integrate these seemingly disparate themes by showing how a counterintuitive result in the recent experimental philosophy literature about the nature of the human self-concept could potentially be accounted for in computational terms. To do so, we use the *Polyscheme* cognitive architecture, due to its ability to account for data on mental-state attribution; and show how, with minimal modification, it provides a lean explanation of variance in human intuitions about the self and causation. In this way, we demonstrate a way in which results generated by experimental philosophers motivate extensions to computational theories of cognition, and how the latter serves to constrain the interpretation of human data.

In section 7 we provide a sketch of the computational model, and some sample inputs and outputs. Finally, in section 8 we outline some of the lessons of this exercise, and what they have to say to cognitive modelers and philosophers alike. Taken together, we feel that this preliminary effort should serve as a model to those who wish to build machines equipped with the conceptual sophistication demanded of a full-blown moral agent.

2. Skepticism about Philosophical Ethics

Machine ethicists hoping to build artificial moral agents would be well-served by heeding the data being generated by cognitive scientists and experimental philosophers on the nature of human moral judgments. This is our core thesis, and it isn't without its critics. Ethicists who are reading this might already be thinking that we've committed an instance of the naturalistic fallacy, conflating how humans *actually* think and do in moral situations with how they *ought* to think and do. We ought to be forthcoming about our pessimism regarding the prospects of building moral agents grounded in classical ethical theory before venturing any further into dangerous territory. In more than two thousand years of ethical theorizing, we've yet to arrive at any more than a few ethical principles that most neurobiologically normal humans have agreed upon as being beyond dispute. But even

among all of this disagreement, there is often widespread convergence on the guilt of the obviously guilty, and the identification of exemplary behavior deserving of moral praise. We have the rather less-ambitious goal of understanding the mechanisms that produce these fairly robust trends among human moral cognizers in an effort to engineer machines with similar ability. Focusing on human peculiarities breaks us out of the cycle of attempting to construct an ideal set of prescriptions for machines. Until we have a more elaborate understanding of the baseline against which we can evaluate a candidate ethic, it makes very little sense to say we are truly making any progress as machine ethicists. This fact, coupled with a growing need to tame technology as machines become more deeply ingrained into our lives, lead us toward a kind of pragmatism about machine ethics that runs afoul of the views of some philosophers and other colleagues in the machine ethics community. In this section, we explore some possible objections to a human-centric research strategy for machine ethics and offer rebuttals.

2.1 Descent into Moral Relativism

The first worry we might encounter is that in the absence of starting with well-defined prescriptions, machines (and humans) end up as moral relativists. In certain situations, especially those involving life-or-death decisions, we'd like for our machines to behave correctly, providing justifications for their actions along the way. There is nothing about grounding machine ethics in the study of human moral cognition that precludes the development of well-behaved systems. Indeed, there seem to be (mostly) well-behaved humans, and as such, we can imagine building human-like machines that are also well-behaved. There is something about the mind that allows us to acquire and deploy deontic constraints on action. Whether or not there are moral universals that correspond to a subset of these deontic constraints is a question that we bin alongside whether or not free will actually exists. No resolution on the issue seems to be forthcoming and it isn't clear that it matters for machine ethicists. If there are moral universals, we perceive them as such because people share much of the same cognitive architecture: exactly what we'd like to reproduce in machines.

We think most ethicists would agree that there have been paragons of moral behavior throughout the pages of human history, and that most people can recognize exemplary behavior in various degrees when they see it. From our perspective, it seems necessary to explain this ability in light of the fact that many of us are not directly acquainted with the thinking of Aristotle, Aquinas, Kant, Mill, or any other moral philosopher. It beggars belief that we could ever do so without adverting to something like an information-processing account of human cognition. Perhaps our moral exemplars are better perspective-takers, or perhaps they are better at resisting temptation or controlling the degree to which their emotions guide their choices. Our point is that it's hard to even explore these avenues as working hypotheses without understanding the mental representations and processes that attach to them.

2.2 Folk Intuitions and Inconsistency

A second worry involves the alleged perversion of ethics via the introduction of folk concepts. Taken further, there is a concern that folk intuitions may lead us to inconsistent moral judgments or otherwise irrational sorts of behavior. Given our previously cited lack of consensus after many

years of ethical theory and application, we find this objection to be lacking in force. Why should Kant's insights be privileged without being able to understand the implications of his metaphysic in every conceivable situation? The latter question applies to any major school of ethical thought. We are aware of some of the initial problems posed by adopting hard-line interpretations of Kant, Mill and others. This isn't to say that there aren't profound insights to be had in studying the classics. It should be understood that there is more than a kernel of truth buried within them. In some sense, it is what makes those views attractive to us as ethical systems in the first place. But we cannot allow ourselves to go further and claim that these insights somehow stand independently from how human thinkers conceptualize their moral lives. Much of the latest research in moral psychology demonstrates a tension between utilitarian and deontological cognition, even at the level of functional localization in the brain (Berns et al., 2012). Regardless of whether one buys into this dichotomous view of human moral judgment, it seems as if we're wired with some of the core constructions associated with traditional ethical theories. But even if we are so constituted, it doesn't follow that we, given finite computational resources, could ever be perfect Kantians, perfect Utilitarians, or perfect anything! John Stuart Mill admitted as much when he says:

> Again, defenders of utility often find themselves called upon to reply to such objections as this: that there is not enough time, previous to action, for calculating and weighing the effects of any line of conduct on the general happiness (Mill, 1979).

To address the computational concern, Mill further says:

> All rational creatures go out upon the sea of life with their minds made up on the common questions of right and wrong.

This sounds hauntingly familiar to our own proposal. Mill looks to be claiming that we ought to be utilitarians *insofar as we are able*. Where we are unable, we rely on intuitions and on accumulated moral heuristics as they've been passed down via cultural transmission. This doesn't excuse us from applying the principle of utility maximization when we can, but it does not force us to apply it where we can't.

Herein lies the trap for the good-intentioned machine ethicist. There are no thinker-independent notions of ethics upon which to build an optimization routine for the calculation of utility. The same can be said for rational inference to universalized maxims, or for the determination of which virtues or *prima facie* duties should be exemplified at any given moment. It should also be noted that using formal logics or decision-theoretic techniques in some form as part of the underlying inferential backbone of a computational cognitive architecture seems to be a near indisputable necessity. What we don't want to do is make the stronger claim that the whole of cognition just *is* theorem proving, model finding, expectation maximization, or some similarly general inference procedure. To say so would conflate different levels of analysis in ways that would ultimately lead to confusion. The more general point is that we cannot go on trying to implement systems that slavishly adhere to principles which are beyond our computational means, and ignore the peculiarities of human nature. If one needs proof of the futility of this enterprise, we should look no further than economics, where rational-actor models have not only lacked predictive power, but have had disastrous consequences when implemented.

2.3 Metaphysical Worries

Lastly, the ethicist might be concerned that building artificial moral agents is a non-starter due to obvious qualitative differences between humans and machines. This is indeed a concern, and one that we are sensitive to. The entirety of ethics and the law is predicated on assumptions about the experience of human freedom, pain, suffering, happiness and similarly murky notions that don't admit to easy computational expression, if they admit to expression at all. Again, we think that this may be less of a worry than it seems to be. We judge and are judged in courts of law. Society does a somewhat reasonable job of policing itself in the absence of having bridges over these metaphysical lacunae. Perhaps we merely need to formally capture our folk conceptualizations of these core notions and figure out how they are deployed in moral cognition. Does the absence of characteristically phenomenal states (e.g. pain, pleasure) in machines pose a problem for the development of artificial moral agents? We think that this is an empirical question. One way to answer it is to see if the folk appeal to the phenomenal aspects of these states in their moral justifications, judgments, and explanations. If the answer is no, then there is no principled reason to think they'd be a barrier. If the answer is yes, then we have the difficult task of trying to account for them in a somehow purely functionalist fashion. Does the mere idea that one's future welfare will be endagered serve as motivator enough for compliance with the moral law? Again, these are open questions that only further study can shed light upon. This uncertainty won't be satisfactory for some, but it might constitute a hard limit on what designers of artificial moral agents can hope to accomplish, and thus be a substantial contribution to the philosophy of science in the area of machine intelligence.

Similar concerns exist due to a lack of a properly naturalized theory of intentionality. Additionally, some are concerned that limits on computation rule out the possibility of machines that could ever entertain arbitrary *oughts*, since (at least on Kant's metaphysic) *ought* implies *can*. This kind of argumentation has led some philosophers to adopt a kind of moral behaviorism as an ethic for artificial moral agents, whereby systems are deemed "good" insofar as they do good things (Beavers, 2011) without recourse to any sort of interiority (or first-person subjective experience). However, it seems difficult to accept any version of moral behaviorism that relies on observables, *tout court*. For one thing, interiority about intentions, prior beliefs, desires, and obligations provide an extra set of resources to discern the quality of actions. What kind of morality assigns praise to the indifferent do-gooder, and heaps scorn upon the sincere bungler? How can praise or blame be defined in the absence of an inference to an actor's *trying* to do the right thing? It is conceivable that we could potentially prosper as a society if everyone behaved in apparently helpful ways without having the appropriate attitude toward performing the actions in question. On the other hand, we would be in serious trouble if we were unable to discriminate an accidental bad behavior from an intentionally bad one. The argument for moral behaviorism constitutes a slippery slope. If we are unwilling to consider mental states in moral computation because we find ethical theories to be uncomputable in practice, why not throw them away for other aspects of cognition as well? When we do so, we lose our ability to participate in dialogue with our fellows, to learn via cultural transmission and goal-based imitation or to organize our lives in a way that is independent of current environmental stimuli. At any rate, it appears to be the case that humans pervasively make inferences about the inner lives of their fellows. A veritable raft of evidence from cognitive neuroscience, child development, cognitive and clinical psychology attests to the human capacity to make mental-state inferences. If we

have such information available to us as moral agents, and having such information allows us to make finer-grained moral distinctions that result in greater accuracy in our ascriptions of praise and blame, it seems morally incumbent upon us to utilize it. To briefly sum up, we should always prefer ethics that embrace rather than eschew hidden variables that are demonstrably predictive in the right way, even if they don't produce optimal results all of the time.

2.4 Summing up

Does our perspective on developing artificial moral agents force us into an unwelcome reduction of the moral to the mundane? We think not. Does it entail throwing out formal ethical systems in favor of relying solely on our moral intuitions? Our first thought on the matter is that it does not. As we have argued, humans seem to agree on instances of moral exemplars. At a very minimum, we can salvage moral principles in machines by designing them to pattern their own behavior (to a degree) after that of an identified moral authority. Does granting a foundational role to moral intuitions imply a sort of deflationary ethics that is determined by how our cognitive architecture is arranged? In the very beginning, perhaps. History has been our laboratory. We have sometimes met with spectacular successes, and have unfortunately met with colossal failures; but we have learned from them and passed those lessons on. In this way we have transcended the constraints of our biology. There seems to be no reason why machines wouldn't or couldn't do the very same with their initial endowments. What seems to be evident is that the richness of human moral cognition is better adapted to our complex social lives than are the strictures of formal ethical theories. While this by itself doesn't constitute a call to abandon ethics as a project for humanity, it does suggest that we broaden the scope of our inquiry in ways that are sensitive to the unique qualities of our species; and it suggests caution against a simpleminded adoption of classical theories in computational artifacts.

3. Machine Morality, Cognitive Architecture and Mindreading

We now offer a few brief remarks on computation, and a justification for the role of computational cognitive science in the development of moral machines. Analogous to our discussion of philosophical issues, we focus on some potential concerns and offer replies.

3.1 The Need for Morally Superior Machines

We aren't suggesting that building machines that are somehow more ethical than a typical human is not in our space of interests. The suggestions made in the previous paragraph about imitating moral exemplars and the ratcheting effects of cultural transmission might very well lead to machines that become increasingly moral over time. This might occur as the result of machines discovering just the right sort of general principles to cover any moral situation it finds itself needing to navigate. Such principles might take the form of an appropriately rich *deontic logic*. Our past work puts this strategy front-and-center in the construction of ethical machines that produce *provably correct* moral behavior on the basis of a particular (but clearly insufficient) deontic logic (Bringsjord et al., 2006). Ron Arkin's "Ethical Governor" is another prime example of just such a system (Arkin, 2009).

Yet, however noble these goals, it's clear to us now that both Arkin's work and our prior work share a rather glaring omission in common. The omission we have in mind further makes the case for the need to capture and formalize moral intuitions in machines. If our goal is to build an artificial agent whose moral capacity somehow exceeds that of a typical human, it seems reasonable to assume that at some point other human agents and their capacity for moral judgment will be the objects of our agent's superior moral reasoning faculties. Of course, this almost necessitates being able to computationally reproduce the variance in prototypical human moral judgments. The systems we've built so far rely heavily on the idea that there are well-defined rules that govern behavior in whatever our domain of interest happens to be. However, the space of situations in which the "folk" are able to make reasonable moral judgments is more or less infinite. Whatever approach we take to building artificial moral agents, it must be capable of generating judgments in a similarly broad space of situations. The best hope for doing so lies in understanding the cognitive mechanisms that produce the highly varied, yet systematic patterns of moral judgments that we see in human beings.

To sum up the case that we are trying to make for the need to take folk intuitions and mindreading seriously, we give the following informal argument:

(1) Moral perfection requires acting in the best interest of other moral agents and patients.

(2) Acting in the best interest of an agent requires knowledge of the inner life of that agent, including its potentially irrational dispositions.

(3) Ethically flawed and otherwise irrational human beings are moral agents.

(4) $\therefore$ Moral perfection requires knowledge of the inner lives of ethically flawed and otherwise irrational human beings.

One might wonder why we make any philosophical hay of the distinction between moral agents and moral patients in the first premise. Why couldn't we just get away with treating other beings purely as moral patients? For one thing, the law as it is practiced by humans, usually individuates offenses committed intentionally from those that are coerced or the result of limited cognitive capacity. Reinventing the law to accommodate robots who are engineered without any concept of what it means to intentionally act would be no small feat, but let's grant that one day such laws might exist. In order to enforce exclusivity between laws applying to humans and laws applying to robots, we might design robots that have no self-interest or sense of responsibility, and are essentially other-oriented, where "other" always picks out a human being or a group of human beings. But let's also be generous and grant that these robots are able to take the autonomy of humans into consideration during their moral deliberations.

To see why this otherwise charitable moral evaluation function for robots might be problematic, let's run through a plausible scenario. Consider two humans, Tim and Tom, and their robotic teammates Chip and Blip. Tim and Tom are subject to all of our ordinary laws as given by local, state and federal authorities. Chip and Blip are subject to a special set of regulations set by the International Council for the Regulation of Robotic Behavior (ICRRB). ICRRB law demands that all robots maximize the good for those humans to whom they are assigned according to the moral evaluation

function roughly sketched at the end of the prior paragraph. Now suppose that Chip and Blip are constructed at a level of cognitive and physical fidelity that impels both Tim and Tom to (falsely) ascribe moral agency to the both of them. Tim and Tom are intimately familiar with human law, which makes distinctions among offenses on the basis of agentive powers, and place moral value on both Chip and Blip since they are considered to be moral agents. It follows straightforwardly that in order for Chip and Blip to maximize the good for Tim and Tom, they must possess some theory of how Tim and Tom behave toward other moral agents. More to the point, Chip and Blip wouldn't want to take any actions which prevent Tim and Tom from acting rightly toward other moral agents, including those entities which Tim and Tom falsely believe to be moral agents. At a minimum, this situation warrants Chip and Blip to consider themselves from Tim and Tom's perspective: as fully responsible beings, but this leads to absurdity, since neither Chip nor Blip consider themselves responsible, autonomous or self-interested.[1]

If any of our other premises seem questionable, it might be number two. But consider how we treat the depressed, the pathologically angry, our children, our pets (as moral patients), the philosophically untutored and even those who are moved to commit certain kinds of retributive acts in the case of gross offenses. These conditions are often taken into consideration when making judgments that have impact on their lives, as well they should be. Only the most egregious violations of the moral law prompt us to see the world in black and white. In large part it seems that laws are by-and-large defeasible under the right set of circumstances. [2]

3.2 Intuitions and Cognitive Architecture

Another potential concern is that we are offering a phenomena-by-phenomena research program which offers no constraints on a general solution to building moral machines. To this end, we are explicitly committing to the use of computational cognitive architectures (Langley et al., 2009) as our primary means of exploring the landscape of moral cognition. Cognitive architectures make commitments to the structure and types of mental representations and processes that constitute human cognition. Suffice it to say that no existing cognitive architecture provides anything like a complete explanation of human cognition, but that shouldn't deter us. One thing that we're fairly sure of is that mindreading plays an integral role in moral cognition. Without the ability to consider the beliefs, desires, intentions, obligations and other mental states of our confederates, much of the richness of human moral cognition evaporates. Since we have it on good evidence that mindreading seems to be a general human ability, it behooves us to choose a cognitive architecture within which mindreading can be profitably explored as it relates to the rest of cognition. We choose to conduct our explorations within the *Polyscheme* cognitive architecture (Cassimatis, 2006; Scally et al., 2011) because it offers us the expressivity and high-level inferential services that are seemingly required for the kind of higher-order cognition that moral reasoning requires. Specifically, Bello has devel-

1. Thanks are due to Fiery Cushman, who brought the agent/patient distinction into focus during a recent exhange. Upon reflection, it strikes us that much of the computational machinery employed by machine ethicists isn't sensitive to the role of autonomy, choice and related concepts in moral deliberation.
2. Presumably, the number of circumstances that serve as defeaters for laws serve as a good metric for just how effective such laws are. This seems to be an interesting computational direction to explore for those interested in so-called contributory standards or certain varieties of moral particularism (Guarini, 2011).

oped the beginnings of a cognitive theory of mindreading within Polyscheme that we have argued to be necessary going forward. One might ask why we wouldn't use a more well-established set of individual algorithms or an existing logic as our computational *prima materia*. Bello has argued in (Cassimatis et al., 2008) and (Bello, forthcoming) that existing formalisms are inadequate as cognitively plausible accounts of the mechanisms we assume to be involved in mindreading. So, far from being a piecemeal research strategy, implementation within computational cognitive architecture forces us to commit to a set of representations and mechanisms that serve as constraints on our theorizing. In this sense, they force us to be specific and acutely aware of resource constraints in ways that traditional ethical theorizing and algorithmic implementations thereof do not.

Now that we've justified our general strategy for pursuing the development of artificial moral agents, and have chosen a set of tools, we must turn to the data we have available on the structure of moral concepts and the nature of the mechanisms that operate over them. To proceed, we briefly motivate the usage of results from the nascent field of *experimental philosophy*, providing a brief summary of some recent work on human intuitions about the bounds of the self, causation, and responsibility as an example of such work. We then introduce *construal level theory* (Trope & Liberman, 2010) and its core concept: *psychological distance*; suggesting that psychological distance mediates inference in ways that might have profound effects on our folk intuitions about the self, and consequently about responsibility. We relate intuitive judgments about responsibility to mindreading, which we take to be the cluster of cognitive processes supporting the generation of behavior explanations. In particular, we detail how construal level and psychological distance serves to promote efficiency in explanation, while producing some of the variance seen in the empirical studies as a byproduct. We then review Bello's existing computational model of mindreading, and implement extensions to the model that explicitly factor psychological distance into the mental-state ascription process. We outline the operation of our augmented model on examples from the target study on the self and responsibility, showing that it reproduces the general trends in the human data, but explains them in a more parsimonious way than is offered by the authors of the study. Finally, we wrap up with a general discussion of our results and with some suggested directions for future research.

4. Experimental Philosophy and Folk Intuitions

An enormous literature on issues surrounding free will and moral responsibility has been generated over the years, but almost all of it has assumed a level of theoretical sophistication and training that one is unlikely to find in an untutored member of the population. A growing group of philosophers have become interested in peoples' pre-theoretical evaluation of philosophical issues and have brought the tools of empirical science to bear on studying these intuitions. Such studies have been typically called "experimental philosophy," and thankfully for us, many of them have focused exclusively on questions surrounding our intuitions about freedom, self and responsibility, all issues that are central to building artificial moral agents. Intuitions about the self and its relation to responsibility will be reviewed as a representative case study from the experimental philosophy literature and will be a test case for exploring the formalization of folk intuitions within a computational cognitive

architecture. But first, we address some questions about methodology as they pertain to the role of experimentation in fundamentally philosophical endeavors.

4.1 Should we Trust the Folk?

A number of criticisms have been leveled at the use of empirical methods to add to the extant body of philosophical knowledge. It has been argued that the proper role of philosophy is as an expositor of the *a priori*, and science as the elucidator of the *a posteriori*. It has also been claimed that experimental philosophy seeks to crassly substitute empirical results for traditional philosophy. Antti Kauppinin suggests that many of the results generated by experimental philosophers reflect pragmatic aspects of the concepts under examination, rather than their true semantics (Kauppinin, 2007). Many of the studies conducted by experimental philosophers explicitly control for pragmatic interpretations, and there isn't an instance of an experimental philosopher claiming that their results should usurp the work performed by traditional analytic methods. If anything, experimental philosophers see their work as being complimentary to the classical approach, and serves to bring the set of intuitions that undergird classical work closer in line with those of the "folk." Insofar as most traditional philosophers engage in thought experiments designed specifically to appeal to ordinary intuitions, there seems to be no principled objection that can be brought to bear without engaging in a non-trivial degree of hypocrisy.

But we do not want to belabor the preceding points. Our primary aim is to do a kind of computational modeling that is informed by work in experimental philosophy. We see this as a contribution to philosophical psychology. The extent to which philosophers are interested in this sort of work as philosophy will depend on the extent to which they think philosophical work can be informed or constrained by empirical insights, and whether the empirical insights discussed herein are relevant to more traditional philosophical debates. We will have more to say about this in the closing section of the paper.

4.2 Intuitions about the Self

Recently, Shaun Nichols and Joshua Knobe ran a study on people's intuitions about the nature of the self as it pertains to responsibility (Knobe & Nichols, 2011). Specifically, different philosophical conceptions of the self were studied: one in which the self is identified with the body, one in which the self is identified with psychological states, and one in which the self is identified with a "central executive" above and beyond the self-as-body or self-as-psychological-state. Nichols and Knobe contend that there is a core reason why these three conceptions of self have been studied so intensely from a philosophical perspective: namely because all three are used in making judgments about how the self relates to action under varying circumstances. In particular, they propose that given some agent A, people will deploy a bodily/psychological notion of self-as-cause when A's actions are considered in a broader situational context. On the contrary, when we zoom in to look at the action itself and the mental processes surrounding it, people will tend to deploy the executive notion of self, treating A itself separately from the processes surrounding the action. For further clarity, we give a description of some of the stimuli, and here quote directly from the source materials:

> Subjects were randomly assigned either to one of two conditions. In one condition, subjects received what we will call the choice-cause case:
>
> Suppose John's eye blinks rapidly because he wants to send a signal to a friend across the room. Please tell us whether you agree or disagree with the following statement:
>
> - John caused his eye to blink.
>
> In the other condition, subjects received what we will call the emotion-cause case:
>
> Suppose John's eye blinks rapidly because he is so startled and upset. Please tell us whether you agree or disagree with the following statement:
>
> - John caused his eye to blink.
>
> Subjects rated each statement on a scale from 1 ('disagree') to 7 ('agree').

As predicted, subjects generally identified John as the cause of his eye-blinking in the choice-cause condition, while asserting that John wasn't the cause of his eye blinking in the emotion-cause condition. Consistent with their "zooming" account, the zoomed-in description of the mental circumstances surrounding John's eye-blink in the emotion-cause condition compelled subjects to deploy the John-as-executive conception of self, whereas in the zoomed-out choice-cause condition, subjects deployed the John-as-psychological-states conception of self. A second experiment was run to rule out the possibility that ordinary folk don't consider being startled as the kind of psychological state that's constitutive of persons. Subjects were told that "John's hand trembled because he thought about asking his boss for a promotion." They were then asked to agree (on a 1–7 scale) with the contrasting statements: (1) *John caused his hand to tremble*, and (2) *John's thoughts caused his hand to tremble*. Consistent with results from the first study, people tended to agree with (2) much more than (1). In a third condition, subjects were given the following:

> Suppose that John has a disease in the nerves of his arm. He experiences a sudden spasm, his arm twitches, and his hand ends up pushing a glass off the table. As the glass strikes the floor, there is a loud crashing noise.

Then, the subjects were given two questions, a "zoomed-in" question, asking them to agree or disagree with the statement "John caused his arm to twitch," and a "zoomed-out" condition asking them to agree with the statement "John caused the loud noise." Again, the results showed subjects willing to agree with the assertion that John caused the loud noise, but disagree with the assertion that John caused his arm to twitch. The pattern of responses given suggests that by asking questions that "zoom out" and consider the situation more broadly, our intuitions lead us to adopt the John-as-body (similar to John-as-psychological-state) notion in our causal attributions. For the sake of brevity, we won't review the last of the experiments, which varied both zooming and the type of action (choice-cause vs. emotion-cause); but the results are predictable, and consistent with the multiple-self-concept hypothesis.

On the face of it, a multiple-self-concept explanation of the data seems to be adequate for explanation, but we find it lacking in parsimony and ultimately unsatisfying. We've uncovered a potential candidate for a framework within which to think the results we've covered so far in a way that doesn't require us to assume anything like multiple self-concepts. The framework is called *Construal Level Theory* (CLT) (Trope & Liberman, 2010), and is centered on the relationship between psychological distance and the abstractness/concreteness of knowledge as represented in the human cognitive architecture. Our later discussion about the relationship between mindreading and responsibility will depend critically on aspects of CLT, so before moving ahead to mindreading, we continue by briefly describing CLT.

5. Construal Level Theory and Folk Intuitions

Construal level theory is fundamentally a framework for thinking about how humans are able to "transcend the here and now," by imagining hypotheticals, remote locations or times, perspectives of other agents, and counterfactual alternatives. CLT's core claim is that each of the aforementioned departures from thinking about the immediate present can be thought of as being "distanced" psychologically from the thinker's conception of here and now. According to CLT, psychological distance is grounded out in the cognitive architecture in terms of varying levels of mental construal. Central to this idea is the fact that knowledge is represented along a continuum, from the incidental features of individual memories all the way up to the invariant fact-like relations that hold over large collections of individual memories. Preliminary neural evidence suggests that pre-frontal areas in the brain support knowledge of this sort at multiple layers of abstraction, lending support to CLT's central claims (Badre & D'Esposito, 2007). Furthermore, studies of how humans conceptually carve up experience point to temporal, spatial, causal, hypothetical, and agent-related phenomena as being natural carving joints (Swallow et al., 2009). All of these dimensions have been explored to some degree within CLT, and have shown sensitivity to varying psychological distances.

CLT enjoys wide evidentiary support from a multitude of studies (see (Trope & Liberman, 2010) for a review). The pattern found time and again across investigations is that greater psychological distance is associated with knowledge represented at higher degrees of abstraction. For example, thinking about far away places or the distant future cues subjects to use highly abstract descriptive features compared to thinking about what's outside the room you're currently in, or what you might do in the next five minutes. When cued with an abstract word like "car," subjects typically think of similarly abstract words like "truck" or "motorcycle," but usually don't think of "ducati" or "cherry red convertible." When asked what they might do two weeks from now, subjects give abstract responses such as "go to work," or "hang out with my spouse" rather than "attend 9:30am meeting with my boss and then further revise this paper." Abstract thoughts describe superordinate, relatively stable sets of features, while the concrete thoughts associated with shorter psychological distances represent subordinate, incidental sets of features that aren't highly predictive of category membership.

5.1 Quantitatively Measuring Psychological Distance

We have been bandying the notion of psychological distance about without regard to anything that looks like a quantitative measure. We've also suggested that psychological distances apply in a variety of core cognitive domains, including space, time, interpersonal similarity space, content-related distance among hypothetical worlds, and in the domain of causal relations. Nowhere have we given an account of whether distances are measured differently along these dimensions of cognitive content, nor have we detailed if and how they interact when producing an overall effect when considering situations that are defined along more than one of these dimensions. This isn't accidental. There aren't any comprehensive theories that purport to provide answers to these questions as of the present. For now, they remain important topics for further psychological study. What we do see is that by way of indirect psychological measures, including reaction-times, implicit association tests, and probability judgments, different levels of construal seem to be associated with intuitive notions of egocentric or indexical distance from the subject in these experiments. While merely qualitative at this point, such differences are informative contributors to computational explanations, and shouldn't be discounted on the grounds that all of the relevant details remain to be worked out. If we cleave to an unreasonable standard of detail, we'll soon find ourselves attempting to provide molecular explanations for morals. It's worth bearing in mind that our goals are computational in nature, and that the research presented in this paper is highly provisional and parasitic upon the current state of CLT as a theory. Some suggestions regarding the quantitative shape of psychological distance measures can be found in (Trope & Liberman, 2010), but much of what we've said above is echoed in the reference material. Part of what's been most interesting about developing the computational model that we shall present has been to embark on a first attempt to capture psychological distance in something that looks like a formally meaningful way. We believe this to be the first effort to do so, and we're especially confident that it is novel from the perspective of the cognitive architecture community.

5.2 Relating Psychological Distance to Intuitions About Self and Responsibility

CLT has it that increased social distance is associated with adopting the perspective of other agents (e.g. moving from a first- to a third-person perspective). CLT also predicts that perspective-shifts entail larger psychological distances, which induce higher-level construals. These high-level construals lead to describing behavior in agent-centric dispositional terms rather than referencing incidental features of the situation within which it occurs. Our contention is that the results reported in (Knobe & Nichols, 2011) can also be effectively interpreted through the lens of CLT without relying on the multiple conceptions of self posited by Nichols and Knobe. On CLT, consideration of longer causal chains should correspond to larger psychological distances. When agents are involved in these longer chains, CLT predicts attributions of responsibility to the agent, since elaborated causal chains should correspond to larger psychological distance from the observing agent. As we'll see in the next part of the paper, psychological distance promotes focus on causal antecedents and diminishes focus on causal consequences. For our purposes, a focus on causal antecedents should lead to a higher likelihood of responsibility attribution to agents in "zoomed out" situations.

5.3 Construal Level Theory and Mindreading

Construal level theory offers a way out of a dilemma introduced by the problem of having to predict and explain the behavior of other agents via mindreading. It is now almost universally agreed that one of the procedures employed in service of mindreading involves the mindreader ***mentally simulating*** himself as the target agent and engaging in prediction or explanation as demanded by the situation. Simulation offers a host of distinct advantages over its competitors under certain circumstances, especially when very little specific information is known about the target *a priori*. While we don't address the prediction issue, since it's less pertinent to the experimental results described thus far, we do see a prominent role for CLT in facilitating action-explanation in concert with mindreading.

During an episode of mental simulation supporting action-explanation, a mindreader A would simulate himself as a target B in a situation where B performed the action in question. The goal of simulative inference in this case is to explain B's action by inferring B's most likely set of mental states such that they would dispose B toward the action in question. The nastier feature of this kind of computation is that actions are multiply realizable in terms of mental states. Many different sets of mental states ultimately lead to the same action. Let's take the simple example of observing an agent making a sandwich. On an unqualified simulative account, explaining this sort of behavior involves inferring an explanatorily adequate set of beliefs, desires, and intentions. We naturally jump to the conclusion that the agent is likely hungry, and he believed that having a sandwich would help sate his hunger, and that sandwich-making involves going to the fridge, reaching for the handle, reaching for mustard, and so on. But what about all of the sub-intentions involved in reaching, slicing bread, stacking ingredients and the like? The acceptable sequences of motor intentions involved in sandwich-making define an enormous space – one that would turn explanation into a mire. The state-of-affairs we just described is an instance of a more general phenomena that directly relates to intuitions about freedom and responsibility. All of us at some point seem to come to the conclusion that agents can usually do other than what they currently intend. From the perspective of sandwich-making, we're considering low-level intentions involved in collecting ingredients. In A's simulation of himself-as-B-making-a-sandwich, if the current intention under consideration involves reaching in the fridge for mustard with getting meat still yet to do, we could imagine a version of A-as-B having the occurrent intention to reach for the meat first, and the mustard second.

What's needed here is a way to abstract the details of the situation away such that only top-level invariant intentions are considered as part of the explanation. We'd like to cite intentions to reach for ingredients in the fridge without worrying about how the reaches are performed, or in what order. In other words, we'd like to focus on the causal antecedents of events (e.g. their top-level intentional structure) and reduce focus on causal consequences, those being the particular form an intentional action takes. CLT provides just such an antidote. CLT predicts that as events are considered at greater psychological distances, they become indexed by their most abstract, invariant characteristics. In the case of actions, details are swept under the rug, and attention is paid to causal antecedents, which for our purposes are constituted by top-level intentions. This phenomena is well-documented throughout the psychological literature, manifesting in early infancy through studies of anticipatory looking behaviors (Cannon & Woodward, 2012) and replicated in studies of

event-boundary detection in perceptual segmentation exercises performed by adult subjects (Swallow et al., 2009).

In her recent dissertation, SoYon Rim documents a series of experimental results that confirm CLT's prediction that focus on causal antecedents corresponds with higher levels of psychological distance (Rim, 2011). In the case of mental simulation, the very act of A taking B's perspective introduces a degree of psychological distance and promotes focus toward top-level intentions rather than the accidental features of particular actions that B has been observed performing. Reasoning backward from the observed effects of B's action produces a simulation enriched with causal structure that introduces ever more distance between A-as-B and the originally considered outcome. Our contention is that the extra psychological distance introduced by enriching simulations in this way will promote focus on causal antecedents that will often ground out in responsibility ascriptions to the target of mindreading. That is, when A simulates himself-as-B when a B-related event has occurred, and the simulation gets enriched via inference, himself-as-B (or just B) will be considered more often as a causal antecedent for the event. Conversely, when such a simulation fails to become enriched in this way, B will be discounted as a candidate for the ultimate cause of the observed event.

We think Nichols and Knobe are on the right track. There does seem to be a generator of complexity in intuitions about moral responsibility, but the generator isn't a multifaceted self-concept. Instead, we assume that the complexity is a function of how knowledge is represented in the mind. We explain the results in (Knobe & Nichols, 2011) by appeal to the concept of psychological distance, and within the broader framework of construal level theory. CLT locates the "zooming" account presented by Nichols and Knobe squarely within the cognitive architecture in the form of abstract-to-concrete gradations in knowledge representation. CLT predicts that psychological distance promotes focus on abstract, stable features of events such as their causal antecedents. In the case of responsibility attributions, distance promotes focus on the agent-as-cause. But what introduces psychological distance into responsibility attributions? We think the process of responsibility attribution depends to some degree on the ability to mindread, and in particular to use mindreading in service of constructing explanations. The act of taking the perspective of the target agent under judgment introduces psychological distance. Extra distance is also introduced during the explanation process, which infers backward from an observed event to antecedent mental states. Each causal inference of this kind promotes further focus on causal antecedents, eventually terminating in the agent-as-cause.

This is clearly not the final form of a solution, and perhaps it isn't even close. We haven't taken into consideration a typical mindreader's beliefs about agency and control, which would be relevant to making inferences about twitches and the like. That being said, we think that this explanation moves us a little closer to a motivated account of why we seemingly have multiple conceptions of self when making attributions of responsibility. But our job isn't done yet. As machine ethicists, we'd like to take this hypothesis all the way to a computational instantiation. Doing so ensures us that our ideas have enough structural integrity that they can be appropriately formalized. As an attractive side-benefit, we end up building a computational foundation for conducting further research as new data become available. In the next section, we demonstrate how distance-modulated mindreading might look by using an established computational model of simulative mindreading.

We formalize an example from (Knobe & Nichols, 2011) and show that psychological distance introduced in mindreading promotes agent-centric attributions of responsibility.

6. A Computational Model of Mindreading

What follows is a very brief description of a computational theory of mindreading that has been used to model early mental-state attribution in infancy (Bello et al., 2007), and errors in attribution (Bello, 2011); and has been used to detail the relationship between mindreading and introspection (Bello & Guarini, 2010). The need for brevity precludes the possibility of providing a detailed defense of the model, so we will have to be satisfied with but an outline of the very basic set of underlying assumptions and computations. The task-model of the data in (Knobe & Nichols, 2011) that we present could be implemented using a variety of computational techniques. We write the model as a set of weighted constraints in Polyscheme's knowledge representation framework.

6.1 Representation and Inference

Polyscheme takes a broadly simulationist approach to mindreading. In classic presentations of simulationism, it's often the case that the mindreader creates a series of pretend beliefs, desires and intentions, "running" these within a mental simulation of the target in order to produce a prediction or explanation. The mindreader operates over this pretend mental content using his own practical reasoning system as a rough-and-ready substitute for the target's inferential capabilities. The result of such simulations are "taken off-line" so that actions performed by the simulated target don't affect the current set of motor intentions held by the mindreader (Goldman, 2006). On our account, simulations of this kind are a particular kind of counterfactual reasoning in which the mindreader identifies with the target within a simulated state of affairs. Information that the mindreader knows about the real world is available within these mental simulations through a process called *inheritance*, which we explore in some detail in the next section. On simulation theories, mindreading involves entertaining a counterfactual statement of the form: "if I were him/her, I'd ϕ." It should be noted that there are a number of details that we must omit for the sake of brevity. For starters, it can be argued that statements prepended with "if I were x" are irreducibly *de se* beliefs, since they involve the first-personal pronoun. Detailed consideration of *de se* beliefs necessitates a defense of Polyscheme's theory of mental content, including whether or not Polyscheme ever utilizes anything like eternally true propositions, or whether or not the contents of its beliefs are *properties* that can change value over time. Provisionally, we subscribe to the property theory and to the minimal commitment to narrow mental content that the property theory entails. An excellent discussion of *de se* beliefs and the property theory can be found in (Feit, 2008). It would take an entirely separate paper to give an account of beliefs *de se*, *de re*, and *de dicto* in Polyscheme, although see (Govindarajalu & Bringsjord, 2011) for a recent computational treatment. Representing and reasoning about counterfactuals involves keeping representations of real situations separate from representations of counterfactual situations. This being said, we embark on some formal preliminaries that detail a situation-centric representation that we will use throughout the rest of the discussion.

6.1.1 Knowledge Representation

An *atom* is a relation over one or more entities that takes a truth-value at a specific time in a situation (or *world*). In general, atoms are of the form *RelName*(e_1, e_2, ..., e_n, t, w). The penultimate argument represents a temporal interval. We use the letter "E" to designate the temporal interval representing "at all times." The last argument-slot defines the world in which the relation holds. We use the letter "R" to represent Polyscheme's beliefs about reality (rather than about imagined or counterfactual worlds). So to say that John is happy all of the time (in the real world), we write: Happy(john, E, R). This may look strange to those familiar with modal logics of knowledge and belief (i.e. epistemic/doxastic logics). We've just called the prior atom one of Polyscheme's *beliefs* about the real world, yet there isn't a modal operator anywhere to be found that denotes it as such. This isn't an accident. Polyscheme experiences the world *as an agent*, rather than in the third-person terms that are implicitly assumed by most epistemic logics. In many ways it seems safer for us to proceed in this fashion. By their nature, epistemic logics express a relation of knowing/(believing) between an agent and some propositional content. But as we mentioned, the ubiquity of *de se* beliefs make us somewhat reluctant to adopt the semantics of epistemic logics without hesitation. For the sake of keeping on target, let's assume that atoms like Happy(john, E, R) comprise Polyscheme's beliefs about the real world.

So far, we've only looked at atoms with no variables in any of the argument-slots. If we want to say something trivial about happy people being people, we write: HappyPerson(?p, ?t, ?w) ==> IsA(?p, Person, E, ?w). Arguments of the form $?e_i$ are variables, and sentences such as the one we just expressed are implicitly universally quantified over each argument-slot. So in this case, if we knew that john, pete, and ralph are happy people on Sunday (in the real world), we'd have: HappyPerson(john, sunday, R), HappyPerson(pete, sunday, R), and HappyPerson(ralph, sunday, R). Each of the three guys would bind to the variable ?p in both atoms, Sunday binds to the ?t slot defining the time at which the relation is true, and R binds to the ?w slots of each atom indicating truth in the real-world as Polyscheme believes it to be. The ==> operator is treated as a standard material conditional. So when we have the three variable-free atoms as given above, they each match the left-hand-side of the conditional, and produce: IsA(john, Person, R), IsA(pete, Person, R), and IsA(ralph, Person, R). Atoms can be negated as well. If we want to say that John isn't happy on Monday (in the real world), we write: ¬Happy(john, monday, R). Atoms can be arranged in conjunctions. If we'd like to say that if John isn't happy on Sunday and Sally isn't happy on Monday then Annie isn't happy on Tuesday, we write: ¬Happy(john, sunday, R) ∧ ¬Happy(sally, monday, R) ==> ¬Happy(annie, tuesday, R). Following the standard semantics for conjunction, ¬Happy(annie, tuesday, R) won't be true unless both ¬Happy(john, sunday, R) and ¬Happy(sally, monday, R) are both true.

Finally, we're brought to two of the more intricate features of Polyscheme's formal language: implicit universal quantification over relations, and weighted constraints. Implicit quantification over relations allows us to express relation-names as variables. So if we want to talk about all of the one-place relations that John is a part of on Sunday, we write ?Rel(john, sunday, R). If John happens to be both happy and excited on Sunday, we end up with Happy(john, sunday, R) and Excited(john, sunday, R). As we shall see, this feature is a key enabler for Polyscheme's ability to reason about the beliefs of other agents.

Polyscheme also allows us to talk about *soft constraints*, which are conditionals that hold, but can be broken at the expense of incurring a cost. If we want to say that it's likely that if John is happy then John is full, we write: Happy(john, E, R) (.8)> Full(john, E, R). The (*cost*)> operator is a conditional that is violable. That is to say that in the real world, it could be the case that John is happy, but not full. When such a situation obtains, the real world incurs a cost of .8. The magnitude of the cost is irrelevant in this particular example. The example is just meant to illustrate the difference between two two different types of conditional statements. The *hard constraint* operator, written as ==>, carries an implicitly infinite cost if broken. So if we have Happy(john, E, R) ==> Full(john, E, R), and we also have Happy(john, E, R) and ¬Full(john, E, R), the hard constraint is violated, and inference ceases. In this way, we can see that hard constraints are similar to what we'd have if we embedded a standard material conditional inside the scope of modal necessitation. The state-of-affairs with John as we just described it is *impossible*, whereas if we use a soft constraint, it is merely costly.

Polyscheme controls inference through the utilization of costs on worlds. As inference proceeds and constraints are broken, their respective worlds accrue costs. Worlds that have hard constraint violations are pruned away due to their impossibility. Polyscheme finds the least-costly world given a set of sentences in its language, and a set of variable-free atoms as input. More could be said about some of Polyscheme's other features, including its ability to reason about uncertain object identities, arbitrary, and non-existent objects, but this is outside of the scope of our discussion. In general, Polyscheme performs a form of *abduction* that delivers least-costly results given a set of inputs. In the example model we present, ***costs on soft constraints are used as a way to coarsely capture the psychological distances associated with mindreading and on the consideration of long causal chains***.

6.2 Simulations, Worlds and Inheritance

As mentioned, our simulation-based theories of mindreading rely centrally on the notion of entertaining counterfactuals. In order to stay on track, we avoid further motivation of the use of counterfactual reasoning as a substrate within which to run the mental simulations associated with mindreading. Instead, we focus on the notion of *inheritance* between worlds. Inheritance as it relates to mindreading can be thought of as the mechanism used to populate mental simulations. Information available to the mindreader becomes available in the counterfactual world where the mindreader is the same as the target through the inheritance process. In essence, inheritance defines the relationship between the world as the mindreader sees it, and the world as the mindreader thinks the target sees it. The most basic form of an inheritance rule is given below, and captures default ascriptions of the form "if it's true for the mindreader, then it's true for the mindreader-as-target."

Def (1) ?Relation(?e_1, . . ., ?t, R) ∧ IsCounterfactualWorld(?w, E, R) (*cost*)> ?Relation(?e_1, . . ., ?t, ?w)

Where *cost* takes a value in the range (0,1). Every time this constraint is broken because the target is ascribed ¬?Relation(?e_1, . . ., ?t, ?w) by assumption or via inference in *w*, costs are incurred. Given what formal machinery we have, we now move on to providing an example of this formalism at work on one of the Nichols & Knobe examples discussed in prior sections.

7. Accounting for the Data

We spend this section exploring the vignette regarding John and the twitch that knocks the glass off of the table. Given the formal apparatus presented in the last section, we can begin to construct a simple domain theory. We first write down constraints that roughly serve the purpose of being circumscriptive axioms[3] (Mueller, 2006) that minimize the number of event occurrences and causal relationships that hold in individual worlds. Let ?ev and ?ce be arbitrary events, and ?ag be an arbitray agent:

c_1: IsA(?world, World, E, ?w) $\wedge$ IsA(?ev, Event, E, ?w) $\wedge$ IsA(?ag, Agent, E, ?w) (.10)> $\neg$Causes(?ag, ?ev, ?world, E, ?w)

c_2: IsA(?world, World, E, ?w) $\wedge$ IsA(?ev, Event, E, ?w) $\wedge$ IsA(?ce, Event, E, ?w) (.10)> $\neg$Causes(?ce, ?ev, ?world, E, ?w)

c_3: IsA(?world, World, E, ?w) $\wedge$ IsA(?ev, Event, E, ?w) (.10)> $\neg$Occurs(?ev, ?world, E, ?w)

These constraints serve to minimize thinking about causal relationships or events during mental simulation unless we have on very good evidence that they actually obtain. We continue by expressing two more constraints that define causal chains for events. In short, the first constraint states that if one thing is caused by another, the latter is in the former's causal chain. The second constraint states that if a causal chain exists for an event, and something is known to cause the most distal event in the chain, then the former gets added to the chain and becomes the newest distal cause. Let ?e0, e1, and ?e2 be arbitrary events:

c_4: Causes(?e0, ?e1, ?world, E, ?w) ==> InCausalChain(?e0, ?e1, ?world, E, ?w)

c_5: Causes(?e0, ?e1, ?world, E, ?w) $\wedge$ InCausalChain(?e1, ?e2, ?world, E, ?w) ==> InCausalChain(?e0, ?e2, ?world, E, ?w)

Next, we have some very simple causal relationships encoded about potential events described by the vignette. The right-hand side of these constraints marks the caused event as a new focal event. This will become important momentarily.

c_6: Occurs(loudNoise, ?world, E, ?w) ==> Occurs(glassFall, ?world, E, ?w) $\wedge$ Causes(loudNoise, glassFall, ?world, E, ?w) $\wedge$ FocalEvent(glassFall, ?world, E, ?w)

c_7: Occurs(glassFall, ?world, E, ?w) ==> Occurs(armMotion, ?world, E, ?w) $\wedge$ Causes(glassFall, armMotion, ?world, E, ?w) $\wedge$ FocalEvent(armMotion, ?world, E, ?w)

c_8: Occurs(armMotion, ?world, E, ?w) ==> Occurs(twitch, ?world, E, ?w) $\wedge$ Causes(armMotion, twitch, ?world, E, ?w) $\wedge$ FocalEvent(twitch, ?world, E, ?w)

3. Reasoning about action and change forces machines to consider all of the possible ways in which the world changes upon taking an action, including conditions which are completely unrelated to the action itself. Circumscription is a logical technique that produces models that describe the effects of actions in a minimal sense, where other features of the situation are assumed to be held constant.

The critical constraint for mindreading is given below. It captures the basic structure of trivial belief ascription by simulation, and implements the inheritance schema presented as Def (1):

c_9**:** IsA(?parentworld, World, E, ?w) $\wedge$ IsA(?childworld, World, E, ?w) $\wedge$ IsCounterFactualTo(?childworld, ?parentworld, E, ?w) $\wedge$ IsA(?ag, Agent, E, ?w) $\wedge$ Same(self, ?ag, ?childworld, E, ?w) $\wedge$ IsA(?ev, Event, E, ?w) $\wedge$ Occurs(?ev, ?parentworld, E, ?w) (.99)> Occurs(?ev, ?childworld, E, ?w)

For any world and for any focal event that happens in that world that involves an agent, the longer the causal chain of the focal event, the more likely the agent caused the focal event. This constraint captures the influence of psychological distance as promoting focus on causal antecedents (e.g. agents) in the context of mindreading.

c_{10}**:** IsA(?parentworld, World, E, ?w) $\wedge$ IsA(?childworld, World, E, ?w) $\wedge$ IsCounterFactualTo(?childworld, ?parentworld, E, ?w) $\wedge$ IsA(?ag, Agent, E, ?w) $\wedge$ Same(self, ?ag, ?childworld, E, ?w) $\wedge$ FocalEvent(?fe, ?childworld, E, ?w) $\wedge$ IsA(?ce, Event, E, ?w) $\wedge$ InCausalChain(?ce, ?fe, ?childworld, E, ?w) (.99)> Causes(?ag, ?fe, ?parentworld, E, ?w)

There are a few other constraints that capture mutual exclusivity relations between each event instance and agent instance as well. We now define the initial conditions of the vignette, which essentially is a description of what subjects read, plus some very basic background facts:

Events: IsA(glassFall, Event, E, R), IsA(armMotion, Event, E, R), IsA(twitch, Event, E, R), IsA(loudNoise, Event, E, R)

Agents: IsA(john, Agent, E, R)

Worlds: IsA(selfworld, World, E, R), IsA(otherworld, World, E, R)

For Mindreading: IsCounterFactualTo(otherworld, selfworld, E, R), Same(self, john, otherworld, E, R)

Percepts: Occurs(loudNoise, selfworld, E, R), FocalEvent(loudNoise, selfworld, E, R)

Once the loud noise is encountered as a focal event with John as a potential cause, simulation begins in order to explain the outcome. The loud noise occurs in the simulated world via the inheritance constraint c_9, and all of the antecedent events and causes are inferred by applying $c_6 - c_8$ to the simulated occurrence of the loud noise. Each of the antecedent events become focal events in the simulation, and causal chains for each are calculated via c_5 and c_6. Worlds having longer causal chains and not having agents as causes are penalized via c_{10}. Given the loud noise as the initially perceived event, the model produces the following output in the "best" (least penalized) world:

Best World: 260, cost: 2.0000004

Causes(armMotion, glassFall, selfworld, E, 260) > true

Causes(glassFall, loudNoise, selfworld, E, 260) > true

Causes(john, armMotion, selfworld, E, 260) > true

Causes(john, glassFall, selfworld, E, 260) > true

Causes(john, loudNoise, selfworld, E, 260) > true

Causes(twitch, armMotion, selfworld, E, 260) > true

On this particular parametrization of costs in the model, it seems as if John is blamed for everything from his arm motion all the way to the loud noise. By playing with the costs, we can adjust how many elements of the loud noise's causal chain John will be considered responsible for. As in (Knobe & Nichols, 2011), we then give the model a set of inputs corresponding to the the question whether John was responsible for twitching. If we run the model with the same input as above except for the replacement of the percepts with

Occurs(twitch, selfworld, E, R), FocalEvent(twitch, selfworld, E, R)

we get the following output:

Best World: 213**, **cost: 0.2

Causes(twitch, twitch, selfworld, E, 213) > false

Causes(john, twitch, selfworld, E, 213) > false

In short, we were able to capture the idea that psychological distance generated during the elaboration of mental simulations make causal attribution to agents much more likely. These efforts are extremely preliminary, but nonetheless a sign that in subsequent work we shall be able to further model some of the more subtle influences on human attributions of responsibility.

8. Discussion and Future Work

It's worth summing up where we've been and where we plan on going next. The motivation for the work we've presented here comes as a contrast to much of what has been standard fare in the machine-ethics literature. Too often machine ethicists draw sharp distinctions between how humans ought to reason and how humans actually muddle through ethically sensitive situations. This position is puzzling to us. Traditional ethical theories and their prescriptions appeal to us precisely because they often prescribe the intuitive. There's something about our cognitive architecture that makes utilitarian judgments attractive at times, and so it goes for virtue ethics, deontology, and other classes of normative ethical theories. If we want to build artificial moral agents that interact productively with other human beings as teammates, we must deal with the messiness of human moral cognition. We also made a case for moving past encoding prescriptions in formal languages on the basis of breadth. Most of the domains in which humans make moral judgments don't have well-defined sets of prescriptions associated with them. It seems clear that moral machines (vice domain-specific advisory systems) need to have something like moral commonsense, and where

better to look for the latter than among the folk. In response, we've decided to take up the challenge of beginning to computationally account for folk intuitions about two central concepts in the moral realm: responsibility and the self.

We considered a study by Nichols and Knobe on the concept of "self," and its deployment in judgments of responsibility. The data from their experiments suggest that when looking at an agent's behavior in a broader situational context, we come to see that agent as a causal force like any other. But once we zoom in close and start looking at the psychological states surrounding the agent's decision-making, it seems as if there needs to be something over and above the agent itself, examining, evaluating, and manipulating these mental states as some form of governor. According to Nichols and Knobe, it seems as if the folk have at least two genuinely different conceptions of self (and perhaps more).

Finally, we formalize an example from the Nichols and Knobe study using an existing computational framework for mindreading and encode psychological distance as costs on logical constraints. We show that simple weighted constraint satisfaction is sufficient for producing the qualitative patterns of judgment in the Nichols and Knobe study. The upshot of our modeling work to date has been to reproduce the puzzling pattern of human judgments found by Nichols and Knobe regarding different conceptions of the "self." Our model suggests that perhaps we don't need multiple conceptions of the self to explain the data. From a computational perspective, this is merely a first exercise. We have much more to do, including accounting for judgments in the other studies that may yield to a similar analysis in terms of construal level. In any case, we expect that the exercise of implementation will lead to further questions, more research, and eventually a richer computational story about moral cognition.

8.1 The Contribution to Computational Cognition

Our purpose in writing this paper was far from providing an in-depth analysis of an individual folk concept, or making a profound contribution to moral psychology. We see this paper as a manifesto of sorts - one that seeks to provide a new way of thinking about how we ought to pursue the construction of moral machines. It is moral machines that we're after, by hook or by crook. The need is driven by independent concerns. In the domain of warfare, it is driven by our desire to take human beings out of harm's way, just as it is in the area of emergency response and disaster relief. In healthcare it is driven by our hopes for better diagnostic aids or companions for the infirm. In short, the need for moral machines is grounded in our own set of moral concerns.

As machine ethicists, we cannot sit idly by and wait for the final word on a suitable and computationally tractable ethical theory to come along. Even if such a theory was developed, it remains unclear that a system implementing such rules would be able to act and explain itself in ways that would be scrutable to other human beings. These are some of the same worries we have about the wholesale adoption of classical ethical theory as the basis for machine morality. We have independent reasons to reject the latter approaches, but transparency and comprehensibility remain here as basic constraints on any attempt at building algorithms for moral computation. Exploring moral cognition via computation provides a wonderful set of challenges to extend current computational cognitive architectures with new representational and inferential capabilities that may be unique to the realm of moral discourse.

8.2 The Contribution to Philosophy

What are we to make of the relationship between studying folk intuitions and the conceptual analysis employed by analytic philosophers? We've addressed this issue in pieces throughout our discussion, but will try and reiterate some of what we feel to be the main issues at hand. First, there is the question of whether or not the intuitions of the folk are substantially different than those of the philosopher. In many of the cases studied by experimental philosophers the answer seems to be yes. So what should we do? Should we endeavor to educate the folk, or to bring the philosopher down to earth? Or is it a little of both? It depends on what the goal of ethical philosophy purports to be.

If the answer to that question has something to do with how to make moral progress as a society, then the answer ought to involve philosophers getting serious about understanding the nature of folk concepts as they relate to human moral behavior. After all, if philosophers are going to educate the folk, then they are going to have to get down to earth and understand how the folk think. Analogously, machine ethics needs to do the same. If good-intentioned moral machines are going to engage with the folk, they must possess either a highly elaborate theory of human moral judgments or must be cognitively constituted in functionally similar terms. We've chosen to pursue this latter option for the sake of parsimony, but there are no principled reasons preventing us from pursuing the former. As machine ethicists, it seems that charting a path toward moral progress, and to machines that behave "better" than we do (whatever that might mean) entails getting serious about understanding the human cognitive architecture and how it shapes the structure of moral cognition as a baseline upon which to improve.

A further reflection on suboptimal or non-ideal rationality is in order. We, as philosophers, often find ourselves unquestionably committed to a rather narrowly conceived notion of what it means to be rational. We ought to stop and ask ourselves every once in a while if our unwavering commitment to idealized rationality is itself rational. We suspect that at least sometimes it might not be. In a recent piece on belief formation and memory, Andy Egan works out the implications of building cognizers with internally consistent knowledge bases that update on the basis of a rational belief-revision process. He contrasts the latter with systems that are fragmented, inconsistent when examined globally, and prone to perceptual error in the belief-formation process. The moral of Egan's story is that systems that fail to be ideally rational fare much better in the kind of world we find ourselves in than a perfectly rational system would (i.e. one that is always internally consistent and not subject to time or resource constraints (Egan, 2008)).

Perhaps the same lesson applies to philosophical ethics. Maybe our slavish devotion to idealized notions of rationality comprise the very source of our difficulty with making greater levels of ethical progress as a species. When we think about moral cognition as serving an adaptive purpose for the human species rather than being disconnected from human behavior *ex hypothesi*, we need to trade what's elegant for what works. Does this mean that we give up on ethics? Absolutely not. But it does mean that treating ethics in a vacuum under ideal conditions of unlimited cognitive resources and a static world is tantamount to performing a gigantic gedanken-experiment. One that may never have any import to real-world moral doings.

Acknowledgements

The authors would like to extend their gratitude to Marcello Guarini for his extensive and eminently helpful commentary. They'd also like to thank Nicholas Cassimatis, Perrin Bignoli and J.R. Scally of Rensselaer's Human-Level Intelligence Laboratory for their ongoing support in all matters intellectual and technical. Finally, they'd like to thank Tony Beavers, Matthias Scheutz, Bertram Malle and Fiery Cushman, whose stimulating ideas motivated much of the content herein.

References

Arkin, R. (2009). *Governing lethal behavior in autonomous robots*. Chapman and Hall/CRC.

Badre, D., & D'Esposito, M. (2007). Functional magnetic resonance imaging evidence for hierarchical organization of the prefrontal cortex. *Journal of Cognitive Neuroscience*, *19*, 2082–2099.

Beavers, A. (2011). Moral machines and the threat of ethical nihilism. In P. Lin, G. Bekey and K. Abney (Eds.), *Robot ethics: The ethical and social implications of robotics*, 333–344. Cambridge MA: MIT Press.

Bello, P. (2011). Shared representations of belief and their effects on action selection: A preliminary computational cognitive model. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 2997–3002).

Bello, P. (forthcoming). Cognitive foundations for a computational theory of mindreading. *Advances in Cognitive Systems*, *1*.

Bello, P., Bignoli, P., & Cassimatis, N. (2007). Attention and association explain the emergence of reasoning about false belief in young children. *Proceedings of the 8th International Conference on Cognitive Modeling* (pp. 169–174). University of Michigan, Ann Arbor, MI.

Bello, P., & Guarini, M. (2010). Introspection and mindreading as mental simulation. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 2022–2028). Austin TX.

Berns, G. S., Bell, E., Capra, C. M., Prietula, M. J., Moore, S., Anderson, B., Ginges, J., & Atran, S. (2012). The price of your soul: neural evidence for the non-utilitarian representation of sacred values. *Philos Trans R Soc Lond B Biol Sci*, *1589*, 754–762.

Bringsjord, S., Arkoudas, K., & Bello, P. (2006). Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems*, *21*, 38–44.

Cannon, E., & Woodward, A. (2012). Infants generate goal-based action predictions. *Developmental Science*, *15*, 292–298.

Cassimatis, N. (2006). A cognitive substrate for human-level intelligence. *AI Magazine*, *27*, 45–56.

Cassimatis, N., Bello, P., & Langley, P. (2008). Ability, parsimony and breadth in models of higher-order cognition. *Cognitive Science*, *33*, 1304–1322.

Egan, A. (2008). Seeing and believing: Perception, belief formation, and the divided mind. *Philosophical Studies*, *140*, 47–63.

Feit, N. (2008). *Belief about the self: A defense of the property theory of content*. New York: Oxford University Press.

Goldman, A. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. Oxford: Oxford University Press.

Govindarajalu, N., & Bringsjord, S. (2011). Logic-based simulations of mirror testing for self-consciousness. *Proceedings of the First International Conference of IACAP Celebrating 25 years of Computing and Philosophy (CAP) conferences: The Computational Turn: Past, Presents, Futures?*. Aarhus, Denmark.

Guarini, M. (2011). Computational neural modeling and the philosophy of ethics. In M. Anderson and S. Anderson (Eds.), *Machine ethics*, 316–334. Cambridge University Press.

Kauppinin, A. (2007). The rise and fall of experimental philosophy. *Philosophical Explorations*, *10*, 95–118.

Knobe, J., & Nichols, S. (2011). Free will and the bounds of the self. In R. Kane (Ed.), *Oxford handbook of free will: Second edition*, 530–554. New York: Oxford University Press.

Langley, P., Laird, J., & Rogers, S. (2009). Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, *10*, 141–160.

Mill, J. S. (1979). *Utilitarianism*. Hackett Publishing Company (original work published in 1861).

Mueller, E. (2006). *Commonsense reasoning*. San Fransisco: Morgan Kaufmann.

Rim, S. (2011). *Distance-dependent focus on causal antecedents versus causal consequents*. Doctoral dissertation, New York University.

Scally, J., Cassimatis, N., & Uchida, H. (2011). Worlds as a unifying element of knowledge representation. *AAAI Fall Symposium Series* (pp. 280–287).

Swallow, K. M., Zacks, J. M., & Abrams, R. A. (2009). Event boundaries in perception affect memory encoding and updating. *Journal of Experimental Psychology: General*, *138*, 236–257.

Trope, Y., & Liberman, N. (2010). Construal level theory of psychological distance. *Psychological Review*, *117*, 440–463.