Murdering and Murderable Minds
Experiments and Remarks on the Psychology of Moral Status

by

Jonathan LaTourelle

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved November 2022 by the
Graduate Supervisory Committee:

Richard Creath, Co-Chair
Elly Van Gelderen, Co-Chair
Jason Robert
Karin Ellison
D. Vaughn Becker

ARIZONA STATE UNIVERSITY

December 2022

ABSTRACT

Moral status questions, (who and what counts morally) are of central concern to moral philosophers. There is also a rich history of psychological work exploring the topic. The received view in psychology of moral status accounts for it as a function of other mind perception. On this view, entities are morally considerable because they are perceived to have the right sort of minds. This dissertation analyzes and tests this theory, pointing out both empirical and conceptual issues with the received view. The results presented show that important moral intuitions (for example about unjustifiable interpersonal killing) cannot be explained by appealing to other mind perception. Some alternative views of the psychology of moral status are presented, as well as avenues for further research.

DEDICATION

For My Mother, For My Teachers, and For Myself

ACKNOWLEDGMENTS

So many people have helped me through my life, through my education, through my research and writing. I want to thank the people principally involved in my research. My committee especially has been generous and patient with me. Elly Van Gelderen spent time and energy educating a philosopher on linguistic theory. My experimental program was first inspired by readings in her graduate syntax course. Karin Ellison employed me, taught me how to teach research ethics, and gave me opportunity after opportunity to do just that. Jason Robert spent countless hours talking with me about my work, recommending relevant material, and being relentlessly optimistic about my prospects. Vaugh Becker helped me think through the first steps of my implementing experimental program and helped analyze the results. And Richard Creath read every draft, encouraged at every step, and carried me across the finish line. I am forever grateful. I must also thank Michael Barlev, who helped me design, run, and analyze my experiments. Without him donating time to educate me, to work with me, to challenge me intellectually, this dissertation would not have been possible. Without his friendship I would be immeasurably poorer.

I should also like to thank the long list of teachers who got me here along the way. My Santa Barbara City College science and humanities teachers for inspiring, educating, and employing me: Fred Marschak, Jan Shultz, Jeff Meyers, Michael White, and Melanie Eckford-Prossor. My teachers at the University of Pittsburgh, Sandra Mitchel, Peter Machamer, and Leslie Hammond nurtured my intellectual

interests. My colleagues and teachers at ASU also provided social and intellectual support: Jane Meinschein, Ann Kinzig , Wes Anderson, Valarie Racine, Kate MacCord, Challie Facemire, Brad Armendt, and Ben Hurlbut. I would vary particularly like to thank Theo Tiffany. I could never have asked for a friend and intellectual teammate more dedicated to my wellbeing and success.

My family and friends have provided social and financial support. Of special note is my aunt Robin Ferry, who rescued me again and again. She showed up when few others would. She purchased my course textbooks for me more than once. And spent countless hours with me on the phone encouraging me.

Many people contributed to a tuition gofundme, which allowed me to finish this work as well. I will spend years thanking all of them in person. A very special thanks to Lisa Gunter and Hanna Wachtel, who housed me when I had no place to live, who fed me when I had nothing to eat, and who lent me a car when I had no way to travel.

Walt Whitman wrote “Still here I carry my old delicious burdens,/ I carry them, men and women, I carry them with me wherever I go,/ I swear it is impossible for me to get rid of them,/ I am fill’d with them, and I will fill them in return.”

Much of what is good and useful in the following pages can be attributed to the time and care these people have shown me. I am forever indebted. I will carry them with me

TABLE OF CONTENTS

Page

APPENDIX Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

LIST OF PICTURES

Picture Page

CHAPTER 1

MORAL STATUS: HISTORICAL, LEGAL, AND PSYCHOLOGICAL

> …moral philosophy without psychological content is empty, whereas psychological investigation without philosophical insight is blind.
>
> (Alfano 2016)

> moral agent and moral patient are real psychological categories that are related to, but more restricted than, the more familiar grammatical categories of agent and patient, and that the former pair of categories can be given an adequate computational analysis that renders them distinct from one another and specifies their standard range of application.
>
> (Mikhail, 2011, section 9.5)

1: The question presented

How do humans make moral status attributions?

To have moral status is to be an entity to whom moral rules apply. It might be useful to think of normative systems, moral systems being just one kind of those, as involving crucially two things: rules, and entities the rules apply to. Normative rules can apply to these entities in two ways: they can bind or compel an entity's actions; or they can protect[1] an entity from unjustified harms. Take, for example,

[1] This is a kind of metaphor here that might be misleading. By protect here I mean that the moral system makes it defeasibly wrong to harm them.

the moral rule expressed by the maxim "my right to swing my fist ends where your nose begins"[2]. That rule binds an agonist's ("my") freedom quite directly. And it protects a nebulous class of actors ("your") against arbitrary assault. What entity's are bound by this moral principle? Surely it isn't "all entities that can make a fist." Infant children make fists after all. And they quite frequently use them to hit longsuffering parents without being dragged before a court on assault charges. As Agustine purporteldly put it, "if babies are innocent, it is not for lack of will to do harm, but for lack of strength." And what entities are protected by this principle? Surely not any that has a nose. Mr. Potatohead dolls have noses, if you put one on them, but there are no advocacy groups attempting to protect them from abuse. Moral status and moral rules interact with each other in powerful ways. If someone's freedom to swing their fist is limited by a Potatohead's nose, they're significantly more limited than if only prohibited from punching humans.

Moral status problems are problems of category membership: who counts as being part of the class of entities whose interest matter morally. As we've seen, entities that have moral status can have it in two distinct ways: they can be moral agents (entities that can be justly praised or blamed for the good and evil that they do) and/or moral patients (entities deserving of protection from evil -unjustified harms-). Moral philosophers have sometimes been in the business of offering normative criteria for membership in each of these categories, and recently

[2] Credited variously to Oliver Wendell Holmes, John Stuart Mill, and Abraham Lincoln, the maxim appears to come from John B. Finch, Chairman of the Prohibition National Committee in the 1880's (Quote Investigator 2011).

psychologists have taken up a descriptive variation of that project. It is the aim of this dissertation to draw on this conceptual and empirical work in an effort to contribute to the still developing naturalized account (Flanagan, Sarkissian and Wong 2008) of moral status (Warren 1997)– in the hopes that a better understanding of how moral status attributions are made might help us enrich our ideas of ourselves, how our minds work, and maybe what we should expect, morally, from ourselves and others.

My goal then is to attempt to discover the psychological shape and structure of these concepts (moral agent/patient), and to characterize the mental mechanism(s) at play when we make judgments about moral status membership. I have no doubt that I will fall short of this goal. But hopefully the attempt will be profitable.

2: Moral Minds

> "Of course, moral status is not a thing, if by 'thing' we mean an object or phenomenon which we can observe in nature, e.g., through a microscope, or with the help of a CAT scanner"
>
> (Warren 1997, 9).

I hope to at least at the outset encourage you to doubt Mary Warren is correct in this about moral status. I do think it is a thing: a kind of psychological mechanism that is embedded in a system of moral perception. As such it is

implemented by brains. And so, in principle can be indirectly observed by some kind of brain scanner.

It can be a bit trivial to say that some psychological process appears on this or that brain scanner – after all, if it is a psychological process, where else would it be happening than in the brain. Even if Warren is correct and moral status judgments are just the products of deliberate argumentation, any human deliberating on moral status would be using their brains to do so. I'm attempting to do something a bit different when denying Warren's claim that moral status isn't a real thing than just say it is too because it's a process happening in the brain. I want to explore the evidence that moral status is a kind of psychological template that itself isn't learned, but rather innate and deployed by the mind as part of moral development (learning who/what counts morally). Once that moral maturation/learning happens, the template is then used by the other mental mechanisms that generate moral judgments.

That take is certainly not original to me, and I'll be exploring two different paradigms of answers from the psychological literature that address these issues. But, in order to even begin to convince you of all this, I think it's important to sketch out some of the broad claims of the research tradition we'll be working in (the computational representational theory of mind). Additionally, it's important to touch the grass a bit before heading into a conceptually dense and difficult topic, so I've spent some time here attempting to illustrate the phenomena in the world that a theory of moral status is supposed to explain. That is all to say, I

want to make clear what my conceptual framework is here, and what phenomena in the world to I expect it to help explain.

First, the mind is a computational representational machine. It isn't like the computer I'm writing on now, but it is some kind of physical-symbol manipulator (Gallistell and King 2009; Pinker 1997; Pylyshyn 1984; Marcus 2008; Marcus 2015). Second, on this view the mind is a system of systems implemented by the brain. Proponents of this view have used many different metaphors to point to these mental systems, calling them at times modules (Fodor 1983), organs of the mind and learning acquisition devices (Chomsky 1975), faculties (Hatfield 1998), senses, or instincts (Pinker 1994).[3]

This fairly aligns me with the research tradition in moral psychology that borrows methods, conceptual tools, and empirical findings from linguistics. Sometimes this set of assumptions is called "the linguistic analogy" (Dwyer 1999; Dwyer, Huebner and Hauser 2010; Mikhail, 2000; Roedder and Harman 2010). While I will in fact borrow somewhat from linguistics in my analysis, I don't want to overstate the analogy. In some ways it amounts to saying that we should treat moral psychology as a problem of representation and computation, like any other branch of psychology persued by reseachers that subscribes to the computational representational theory of mind. We could as easily and truthfully suggest that the moral sense should be investiaged in the same way we investiate

[3] For a critique of this research see "the new phrenology"**Invalid source specified.**.

the number sense (Gelman and Gallistel 1986). But, at least to get going, all we really need to borrow from linguistics is the methodological assumption of the "Tenet of Constancy": "whenever a psychological constancy exists, there must be a mental representation that encodes that constancy" (Jackendoff, Languages of the mind, essays on mental representation 1992, 5). And further, that explanations of that encoded constancy will come in the form of mentalistic and formal theories (Jackendoff 1997, chapter 1), sometimes called a cognitive model (Allen 2014).

I view the project of moral psychology as an attempt to explain human moral intuitons as the product of the moral sense. In terms of the history of philosophy, this verbage about a moral sense might cause some confusion. Quite usually "moral sense theory" is a name reseved to refer to the empiricist moral senimentalist (see Frick 2010 or alterntiviely Gill 2006, for worthy overviews), such as Francis Hutcheson (Hutcheson 1728/2013), David Hume (Hardin 2007), Adam Smith (Evensky 2005), and Peter Kropotkin (Kropotkin 1902/1972, 1924, Morris 2002), as well as contemprarily philosophers like Jesse Prinz (Prinz 2006) and Patricia Churchland (Churchland 2011). This reseach tradtion holds that moral judgments are (necessarily and sufficiently) accounted for by moral emotions. Here's Prinz favorably quoting Hume: "To believe that somehing is morally wrong (right) is to have a sentement of disapporbation (approbation) towards it" (Prinz 2006, 33). Additionally this reseach tradition often emphasizes the role of domain general learning mechanisms in the formation of moral beliefs.

Contrastingly the tradition of "ethical rationalism" refers to a reseach tradtion that views moral judgment as a product of (pure and practical) reason alone (Cudworth 1731). This view could also be applied to a number of natural law theorists (for example, Suárez 1675, Pufendorf 1673/1991). Jospeh Butler attempted a kind of synthesis of moral sense and rationalist tradtions which I am favorably disposed to. But as the legal scholars John Mikhail and Matthias Mahlmann write, "be it logos in the Stoic tradition, Synderesis in Scholastic thought, the "moral sense" of the British Moralists, or "reason" in the tradition of modern Natural Law", from a naturalistic perspective the object of investigation is the same (Mahlmann and Mikhail 2005, 95). In fact, I will argue that all those words are well labeled black boxes that we should be very interested in trying to open.

I think it is fair to say that all these traditions should be thought of as opposed to the positivism and egoism of the Hobbesian view that morality is "a device discovered by men for securing the maximum degree of self-preservation" (Hudson 1967, 3). Morality on this view is always a product of a faculty of the mind. And arguably best conceived of as a perceptual faculty (Mikhail, Elements of Moral Cognition 2011, 4.4.1). So, morality isn't an invention like writing, but part of our human nature like language.

What are the parts of this moral mental module? Hudson in his discussion of ethical intuitionism points out that on their view "three functions may be attributed" to the "moral faculty": the "perception of moral properties;

approbation or disapprobation; [and] motivation, or excitement to action[4]" (Hudson 1967, 9). As a first approximation of a complex set of interrelated cognitive systems this is about as good as it gets. We have a system that makes judgments of rule violations (perceives moral properties), a separate system that apportions blame or praise (approbation or disapprobation), and a third that motivates us to actions-or not (that is, a system of moral emotions that also interfaces with other pragmatic judgments). In my work I focus primarily on the perception of moral properties, and the assignment of approbation and disapprobation. Though I hope to not totally neglect the emotional motivational aspect of the moral sense and moral status judgments.

I want the reader to consider very carefully that there are now already two systems of moral evolution. And that both apparently have to interface with the mental mechanism that instantiate moral status judgments. A novel hypothesis explored in this dissertation is that moral status is processed differently by the faculty that perceives moral properties of an action than by the faculty of approbation/disapprobation. How moral status judgments fit into this picture is not immediately clear – but we will see that moral status judgments significantly impact both the moralistic judgment we make about an actor's actions in some set

[4] It is important to note that like the 17th century Ethical Intuitionists, current researchers distinguish between moral behavior, the perception of moral properties, and the assignment of blame and praise. In the contemporary tradition the latter two are parts of the moral faculty proper (i.e., part of moral *competence*). Moral behavior on the other hand is part of *performance*.

of circumstances (the moral properties we detect in an event) and the amount and kind of blame or praise we think they deserve for those actions.

In these efforts I'm most deeply indebted to the work of John Mikhail, especially the masterful *Elements of Moral Cognition* (Mikhail, Elements of Moral Cognition 2011), which, with a footnote, inspired most of this research project. And also, to the work of Kurt Gray, especially *colleagues*(Wegner and Gray, The Mind Club: Who Thinks, What Feels, and Why It Matters 2016), which roused me from my dogmatic slumber. In a way this project is an attempt to reconciling these two approaches to moral psychology. While I don't think I succeeded at that, I hope I've done them both justice by taking them seriously.

3: The Moral Dyad in The World

Whenever humans make moral judgements about an event, a central feature of the event is the cast of characters involved: the doers (agents) and the done-tos (patients). Who those characters are powerfully controls the kinds of moral judgments we make about the events they're a part of. Imagine, for example, that a young child attacks a neighbor's pet cat, killing it. Now imagine that instead of a young child the attacker is a dog, or an adult man, or another cat. Changing nothing about an event other than who's doing something to whom makes a dramatic difference to moral judgments.

In "The moral dyad: a fundamental template for unifying moral judgment" (2012) Kurt Gray, Adam Waytz, and Liane Young argued that unlike the rules of

our moral psychology, which seem to be domain specific[56], judgments from all of these domains include the same basic cast of characters: a moral agent-patient dyad. And in all domains the nature of these characters, which entities play the agent-patient roles, substantially effects moral judgments.

Gray and his colleges have a worthy hypothesis about what the input conditions for moral status membership are. On Gray's view, moral status is fed by mind perception (Wegner and Gray, The Mind Club: Who Thinks, What Feels, and Why It Matters 2016). So moral agents are entities that other people think of as having extremely sophisticated mental capacities for agency (they can intend their actions, and control them, etc.). And moral patients are entities that other people think of as having extremely sophisticated mental capacities for experience (they can feel pain, hunger, pleasure etc.).

I'll have a lot more to say about this in Chapter Two, but in the rest of this chapter I want to draw your attention to the type of phenomena to be explained - the explanandum in the parlance of philosophy (Hempel and Oppenheim 1948, 136-7). I'll do that by pointing at cases from history, and some fiction, that illustrate the diverstity of possible positions about who or what counts as a moral

[5] Domain specific i.e., rules about purity purportedly cannot be reduced to rules about harm, nor those about harm to those about fairness. Though this is contentious, and it may be possible to reduce all of morality to concerns about harms (for example see, Schein and Gray 2017)

[6] The moral domains, called "foundations" by Jonathan Haidt, include care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, and sanctity/degradation, and Haidt defends a sixth liberty/oppression**Invalid source specified.**.

agent/patient. It will be hard to miss the unity underlying this diversity, and that too will be instructive.

The examples that follow in the next sections of this chapter are intended as inution pumps (Dennett 2013). They should be productively puzzling. And hopefully serve to show that there's some skin in the game when we're analizing the concept of moral (agent-patient) status. Judgements of moral status are absoutely centeral to moral philosophy. But they're also at play when jurists make judgments, when fiction writers create a moral narritive, and when everyday people get outraged over injustice.

These examples are cases that call out for an explanation. My efforts towards a satisfactory expanation here have been largely empirical. But it is inescapable that our moral psychology is deeply relevant to our moral philsophy. This, however, is not an atempt to use the intutions we form about these cases as dispostive with respect to normative arguments we might be intersted in making; i.e., this is not an effort to use them as part of a project employing "the method of cases" (see Machery 2017 for insightful criticism of this method).

3.1 Moral Agents

Moral agents are entities that can justly be blamed and praise for their actions.

3.1.1 Human partial-agents and impaired will

In June of 2012 James Butwin wrote out two detailed suicide notes, and then proceeded to kill his wife and their three children, before killing himself. *The Arizona Republic* relays friends and acquaintances shock in the wake of events: "He was totally soft-spoken and a devout Jew. He was very peaceful like that, very even keeled. […] I never saw a flash of anger from him." They also relay that Mr. Butwin was "battling a brain tumor" at the time (Hansen and Wagner 2012).

In August of 1966, after murdering his wife and mother-in-law, Charles Whitman went to the University of Texas in Austin with guns and ammunition, climbed a tower, and opened fire indiscriminately on people below. Whitman too had a brain tumor. David Eagleman quotes his suicide note usefully:

> It was after much thought that I decided to kill my wife, Kathy, tonight … I love her dearly, and she has been as fine a wife to me as any man could ever hope to have. I cannot rationa[l]ly pinpoint any specific reason for doing this.
>
> (Eagleman 2011)

Both of these cases are intended to illustrate what moral agency is by showing what it looks like when it is impaired. What caused the murderous behavior in these cases? The men themselves, through their choices and intentions? Or the brain tumors? Moral agency, we can see through these cases, requires a notion of freedom of the will, freedom of choice, freedom to do

otherwise (or what Alfano's calls the condition of control (Alfano 2016, 55)). I want to sidestep the bloody battlefield in moral philosophy populated by compatibilists, incompatibilities, and libertarians with their concerns about determinism.[7] Irrespective of whether freedom of the will exists, or whether that question is even answerable (Machery 2017; also "Free Will Hunting", Futurama, Cohen 2012)[8], it is perceived to exist in people not otherwise extremely impaired. Or rather, we should start to consider the hypothesis that entities not perceived as having freedom of the will are therefore not perceived as being moral agents. If this is true, we should expect to find lots of cases in the history of law, and in moralistic fiction, that turn on judgments of an agent's capacity to do otherwise: we do.

Brain tumors aren't the only way moral agency may be diminished. In Atkins v. Virginia, 536 U.S. 304 (2002), the US Supreme Court held that "executions of mentally retarded criminals are "cruel and unusual punishments" prohibited by the Eighth Amendment." The opinion goes on to say

[7] In "Elements of Ethics" Bertrand Russell summarizes things well: "What determinism maintains is that our will to choose this or that alternative is the effect of antecedents; but this does not prevent our will from being itself a cause of other effects. And the sense in which different decisions are possible seems sufficient to distinguish some actions as right and some as wrong, some as moral and some as immoral" (Russell 1910, Sec. IV). People who accept some version of this thesis are compatibilists, those that don't, aren't. Or see (Mele 2019) (McFee 2014) (Vihvelin 2018)

[8] I agree largely with Machery that philosophers should give up on questions like "whether necessarily an action is free only if the agent could have acted otherwise" (Machery 2017, 1).

> Mentally retarded persons frequently know the difference between right and wrong and are competent to stand trial, but, by definition, they have diminished capacities to understand and process information, to communicate, to abstract from mistakes and learn from experience, to engage in logical reasoning, to control impulses, and to understand others' reactions. Their deficiencies do not warrant an exemption from criminal sanctions, but diminish their personal culpability.

In our language here, the courts decided that people with serious, but unspecified[9], intellectual disability were not full moral agents. Or at least, at first blush that is a plausible reading. However, the last sentence is crucial: intellectual disability does not "warrant an exemption from criminal sanction". If, as the ethical intuitionists held, there's a cognitive system that "perceives moral properties" (generating answers to 'does this event contain a moral offense') and another that generates culpability judgments (generating answers to 'how much punishment is deserved for this offense'), moral status might act differently in those systems. That is, we might discover that people judge that diminished capacity doesn't affect whether an entity qualifies as moral agent with respect the evaluative system, but that it is heavily weighted in the culpability system.

[9] "mental retardation" was a broad cluster category that encompassed many different intellectual disabilities. Here, since the term "retardation" is widely considered to be offensive, I will use the term "intellectual disability" also to refer to that broad cluster category.

All this raises a question about the psychological structure of the moral agent concept: namely is it an all or nothing affair (discrete -binary nominal-), a hierarchy (discrete -ordinal-) or does it come in degrees (continuous -scale-)? It also highlights that the concept might take one of these forms in the evaluative system, and another in the culpability system. If the court's reasoning can be taken as suggestive preliminary evidence, we could tentatively rule out a binary nominal hypothesis for moral agency with respect to culpability, but not evaluation. Yet we should take caution before conflating normative and positive judgments (such as judicial rulings) with a descriptive project aimed at characterizing a set of intuitive judgments and explaining them by proposing a cognitive model that could generate them. It could be that the opinion of the courts here isn't representative of folk judgments on the subject of diminished capacities but is a product of extensive legal education[10]. Again, our examples here are intended to raise questions rather than settle them.

Pressing on, cases involving criminal children also provide fodder for our investigation. Children, past infancy, are certainly blamed and praised for their

[10] This may very well be the case, consider the follow contrary evidence: in the majority of cases "a defendant with mental retardation is more likely to be arrested, more likely to be held pending trial, more likely to be convicted, [and] more likely to receive longer sentences" (Nevins-Saunders 2012). If people with intellectual disability were generally perceived, by the average police officer and jurist, as not having full moral agency, we'd expect that they'd be treated more leniently -less responsible, less culpable-. This evidence is more consistent with the hypothesis that intellectually disabled people are downgraded with respect to their moral patienthood -i.e., causing unjustifiable harms against them is on average perceived as less wrong.

behaviors. And if the behaviors are bad enough, and they're old enough but not too old[11], they'll find themselves in front of the juvenile justice system. If their behavior is extremely bad, they'll find themselves "tried as an adult" in front of the adult justice system[12].

Tessa Majors was walking in Morningside Park in Manhattan in the evening of December 11, 2019, when she was accosted and stabbed by three teenagers, which later resulted in her death. The assailants it turned out were boys, one thirteen-year-old and two fourteen-year-olds. The two fourteen-year-old boys were charged as adults with second-degree murder and robbery. But the thirteen-year-old when through the juvenile court system and was sentenced "to up to 18 months in a juvenile detention center, were he will undergo mental health counseling and be able to continue his education" (Piccoli 2020).

The family of Tessa Majors objected to this lenient sentence: "there are no minor actors in the murder of Tess Majors" (ibid). But LaVonne Roberts, an MFA student and mother that lives near the park where Tessa was murdered, wrote an objection to the treatment of these boys published in the Columbia Daily

[11] Age classes for juvenile court vary from state by state. With some states mandating that certain serious crimes committed by children above a certain age, 14 in New York for example (The New York Times Editorial Board 2020), automatically qualifies them to be "tried as an adult".

[12] Accoding to Snyder, Sickmund and Poe-Yamagata 1996, p28-30, transfering juveniles to adult criminal court increased 41% between 1989 and 1993, a likely biproduct and "tough on crime" policies that were so popular in the late 80s and into the 90s. Thought its notable that while "war on drugs" related laws were in effect durring this time, a vast majority of the increase in transfers (and therefore punishment) can be attributed cases that involved interpersonal harm, not drug sale or possession.

Spectator titled, “No 14-year-old should be tried as an adult” (Roberts 2020). This push and pull over the partial moral agency of children is evidenced in all cases of this kind, where children are tried as adults, and when they’re not. But again, we can draw attention to the two moral systems I’ve been alluding to and how they might handle moral status differently: this system appears to treat children as full moral agents in terms of evaluation (did they do something wrong) and as partial (-scale wise-) or downgraded (-ordinally-) moral agents by the culpability system (how much punishment is just). The contrasting way moral agency might be handled by these two proposed systems could have provided the psychological impetuous for creating a separate court system for children; a system that still condemns them for immoral behavior, but also ensures that punishments are structured as individually tailored rehabilitation plans: the “juvenile justice system was founded on the concept of rehabilitation through individualized justice” (The Office of Juvenile Justice and Delinquency Prevention 1999, see also Snyder, Sickmund and Poe-Yamagata 1996).

So far, I’ve reviewed cases of impairment and adolescence. I’ve used them to pump some intuitions out of you, the reader, about who should count as a moral agent. I’ve done this by drawing your attention to how intuitions about justifiable punishment change (on hypothesis) because of the intellectual impairment of the agent. In the next case however, we’ll see this isn’t the only way we can try to tease out the psychological structure of the concept of moral agent status.

In February 2011, Sir Xavier Brooks (19 at the time) was murdered during a fight outside a club in Chandler Arizona. The case is rather straightforward: witnesses were eager to report on the altercation and they identified Orlando Nembhard (also 19) as the perpetrator. However, the police and courts had a problem. Orlando's identical twin brother, Brandon, was also at the scene of the crime. This all made relying on witness testimony for identification very difficult. Complicating matters, each twins blamed the other for the crime (Lacey 2011). So far as is publicly available, charges were dropped against Nembhard because there was insufficient evidence to know which twin was involved (Maricopa County Attorney's Office 2011).

This case also highlights a feature of moral agent status, so far as I can tell, first identified by Steven Pinker in *How the Mind Works* (1997). A moral agent is an identifiable individual. This might seem like a strange thing to dwell on, because of course it is. However, Pinker points out this feature of our moral psychology significantly constrains the representational format the concept of a moral agent must take. This is because "Rather than symbolizing an entity as an arbitrary pattern in a string of bits" (i.e., with a rigid designator[13]), neural networks don't represent individuals at all. Instead, individuals are picked out by a "pattern in a layer of units, each standing for one of the entity's properties". Identification then is a process of specifying so many properties that in the ideal

[13] Rigid designator is a term coined by Saul Kripe in "Identity and Necessity" **Invalid source specified.**. However, Ruth Marcus described the same idea as a "tag" in "Modalities and intensional languages" (1961) ten years prior.

case the members of the set so specified equals one. But "An immediate problem is that there is no longer a way to tell apart two individuals with identical properties. They are represented in one and the same way, and the system is blind to the fact that they are not the same hunk of matter" (Pinker, How the Mind Works 1997, 114-5). Whichever twin committed the crime, blame is not apportioned by degrees to any entity with sufficiently similar properties to the actual agent. We really care about which of the two twins committed the crime.

So far, we've noted three types of distinctions relevant to the concept of moral agency. First, that it might be profitable to decompose the concept into component features. Second that moral agency might come in degrees or kinds. And third that moral status might be treated differently by two different parts of our moral cognition.

The features we've draw attention to have largely been concerning the minds of the agents involved. This approach is not uncommon. So, for example Alfano's argues that "Adult humans are generally assumed to be at least partially responsible for their behavior […] provided that at least two conditions are met: a knowledge condition and a control condition (Alfano 2016, 52). Cases of impairment cause issues for the condition of control, and that way effect moral status judgments. But not all of these "conditions" or features are about the kind of psychological capacities the agent is perceived to have. We've also noted that in order to be a moral agent you have to be an identifiable individual.

The second distinction we've sorted out of these cases is a concern about whether moral agent status is an all or nothing affair. Can you be 2/3 of a moral agent? Or can you instead just be some type of moral agent: say, a juvenile moral agent. Or a moral agent lacking "the condition of control" but having all other relevant conditions. A features list decomposition could be used to generate a hierarchy of moral agency, such that an entity was placed on the hierarchy by checking the list of moral agenthood-relevant features an entity is perceived to have compared against the list of moral agenthood-relevant features any moral agent could maximally be perceived to have: that is, against a prototype moral agent. We will see this concern again in the example cases to come. And will deal with it more formally in Chapter Two.

And finally, we've noted that moral agency might be handled differently by two different parts of our moral psychology. Suggestively, moral agency might be an all or nothing affair with respect to the system of moral evaluation. And a graded or scale affair with respect to the system of moral culpability.

All three of these distinctions will continue to be relevant to the other examples concerning moral agency. And, in fact, to the examples concerning moral patienthood too.

3.1.2 Moral agency in animals:

The Book of Days chronicles the story of the "trial of a sow and pigs at Lavegny", according to which in 1457 a female pig and her six young piglets

were put on trial for murdering and partially eating a child. Mark you, it was a fair trial, the defendants were provided with competent legal counsel, and brought before a duly appointed magistrate. It was decided that sow was guilty of the charges, and she was put to death. But her pigs were acquitted, and on grounds we might find reasonable, if they were not being applied to pigs: they were exculpated "on account of their youth, the bad example of their mother, and the absence of direct proof as to their having been concerned in the eating of the child" (Chambers 1869, 128-129).

Picture 1. Sow on Trail

This case is a frequent example from the history of animal trials in European law, probably because it's accompanied by this evocative illustration.

Even 152 years ago this case struck the author of *The Book of Days* and his print maker as comical: "Our artist has endeavoured to represent this scene; but we fear that his sense of the ludicrous has incapacitated him for giving it with the due solemnity" (ibid, 128). And in truth the print is comical – showing the sow in tears at court, with her piglets nursing as she is sentenced. But in 1457, trying animals for murder, theft, and various other crimes was common and completely serious. *The Book of Days* provides some useful context, which bears quoting at length

> On the Continent, down to a comparatively late period, the lower animals were in all respects considered amenable to the laws. Domestic animals were tried in the common criminal courts, and their punishment on conviction was death; wild animals fell under the jurisdiction of the ecclesiastical courts, and their punishment was banishment and death by exorcism and excommunication. [… and] In every instance advocates were assigned to defend the animals, and the whole proceedings, trial, sentence, and execution, were conducted with all the strictest formalities of justice.
>
> (Chambers 1869, 126-7).

The European trials of animals might be a fun cocktail conversation, but it isn't clear that they provide evidence in support of the hypothesis that the people involved in these proceedings really perceived the murderous pigs and felonious rats as moral agents. The legal scholar Gary Francione expresses those doubts

well: “The trial and execution of animals is a legal anomaly. Even if such trials and exestuations represented some notion of animal responsibility or rights, it would be difficult to jump to the conclusion that animals possessed rights or responsibilities under the law” (Francione, Animals, Property, and the law 1995, 94).

Nevertheless, we are interested in that “notion of animal responsibility” and its psychological roots. And, analogous to the proposed genesis of the system of juvenile justice, how from that notion and those roots a body of positive law treating animals as some sort of moral agents could grow[14]. Chambers recounts that there was controversy at the time about if all this made any sense, noting that “some learned canonists” disputed the legitimacy of punishing “lower animals […] devoid of intelligence” and that “punishments for injuries committed unintentionally and in ignorance of the law, were unjust” (ibid). And if there were iconoclasts[15] who objected to the sensibility of blaming animals for their crimes, that suggests that at least some relevant people thought the practice did make sense.

---

[14] In Chapter Two, we’ll find that the distinction between an I-Morality (I as in internalized, intensional, and individualized) and an E-Morality (extensional, externalized) (Mikhail 2011, 63) gives us some traction on how the former can lead to the latter.

[15] Some did object to blaming animals. For example, Pierre Ayrault, a French jurist form Angers, argued that “brute beasts and inanimate things are not legal persons (*legales homines*) and therefore do not come within the jurisdiction of a court “ (Evans 1906, 109). And Thomas Aquinas held that unless animals were possessed by the devil, it was inappropriate to bless or curse them (Evans 1906 54).

The exculpatory criteria given by Chambers' "canonists" are worthy of deeper consideration. First, ignorance of the law, in most cases, is not exculpatory for moral agents. And so, we're left asking the same question in a different way: for what kind of entity is ignorance of the (moral) law not exculpatory? Likewise, it is hard to see in what sense the pig eating the child was not intentional[16]. In any case, if a human did those things, all else equal, we'd call them intentional.

We cannot go on long with such a loose explication of "intentional" with respect to moral agency. Nor, for that matter, without exploring further what kinds of ignorance are in fact exculpatory. And we'll deal with those issues more formally in Chapter Two. However, for now it is worth seriously considering the trial of the sows and her piglets: after all, six out of seven pigs were exculpated. And the exculpating reasons are important: the piglets were too young to know better and they had a bad parental example. These are exculpatory criteria we recognize. The absurdity of this case from a contemporary perspective isn't sourced in the reasons given, but because we fairly wonder at what age a pig becomes morally responsible for her actions? As well as the on what grounds we are warranted to go about blaming a sow for not giving her piglets a moral education; something she's patently incapable of doing. To most modern readers

[16] By intentional here I mean only that the pig had a desire/goal (to eat the child). That she had a set of beliefs constituting an action plan describing how to satisfy that desire (go from here to there, bit down, chew and swallow, etc.), and the physical capacity to implement that action plan. And that she did in fact try to obtain that goal by implementing an action plan designed to achieve it: i.e., she acted intentionally.

then, a pig, whatever its age or education, is just not a moral agent. Though this too is an empirical claim, one we address head on in Chapter Three.

In a commentary in *Animal Law*, Sykes says that the animal trials could very well have been "legal ritual" that "disguised [the] exercise of brute power". But goes on to say that they "gave voice, in a limited way, to the voiceless" (Sykes 2011, 278). I agree with Sykes: two things can be true here. The first is that pragmatic and even sadistic[17] motives could serve to sustain and motivate "a legal anomaly" (Francione, Animals, Property, and the law 1995) like this. And at the same time the people involved were treating these entities as moral agents, in part because it was believed that they possessed the right sorts of capacities.

Sealey (1983), claimed that trials of animals (and objects, next section) were rooted in the "very primitive notion that things and animals are responsible agents. […] It is in no wise strange that people who saw something divine in every fountain, river and tree, should have endowed all common things with life and animals with responsible intelligence" (Id, 358). But as we've seen, phrases like "responsible intelligence" are the black boxes of moral agency. These terms seem intuitively appealing but behind that appeal hide the potential features we're interested in: what kind of intelligence amounts to the 'responsible' kind of

[17] Though sadism too can be seen as an exercise of justice – or rather an exercise of our sense of retributive justice. For example, Evans (1906, 140) recounts that "in 1386, the tribunal of Falaise sentenced a sow to be mangled and maimed in the head and forelegs, and then to be hanged". He insists that this was a "strict application of the *lex talionis*" – it was retribution against the sow "for having torn the face and arms of a child", killing it.

intelligence. Or, in our language so far, what are the criteria required for moral agent-hood.

A further question is whether in fact the ancients were as strange as they might appear from this history. We should always guard against the moral vertigo induced by looking to the past. That vertigo (sometimes called historical presentism) often unjustly transforms moderate differences into the incomprehensible strange or repugnant. After all there are certainly academics today that argue that the concept of legal personhood should be extended to a wide range of nonhuman animals (for an outline of the debate, see Francione and Garner 2010). Some of these efforts extend even to environmental entities like rivers. Though both of these efforts might more fairly be thought of as ways to ensure these entities protection against harm (i.e., moral patient status (See 2.2.3), they're not clearly representative of folk morality generally.

Setting aside historiographical concerns for a moment, it's important to remember that it is assumed in most of the philosophical literature that the potential for just-blame automatically licenses the potential for just-praise. So that part of the phenomena to be explained by a naturalized account of moral agent status would include cases like that of Winkie the "heroic" pigeon who was the first nonhuman animal to receive the Dickin Medal for "delivering a message under exceptional difficulties." In fact, the existence of the Dickin Medal itself, an award established in 1943 to honor the work of animals in war, and its 71

honorees, provide 71 cases of nonhuman animals being considered as (at least some kind of) moral agents (Force War Records).

3.1.3 Moral Agency in inanimate entities

Theogenes of Thasos was famous around 480 B.C.E. for his Olympic victories. After his death a statue of him was made. A former adversary of the athlete apparently obsessed by indignities he suffered at Theogenes' hands, took it upon himself to punish the statue. Purportedly each night, the man would visit the statue and physically attack it. On one such occasion the statue fell upon its aggressor, causing his death. The dead man's son took the statue to court where it was successfully prosecuted as a murderer, and upon a guilty verdict, it was thrown into the sea as punishment (Sealey 1983, 347). The ancient Greeks in fact had a court that handled special cases in which the agents were nonhuman animals or inanimate objects. In this context it was perfectly possible to express sentences like "the javelin murdered the boy" (id, 343), which I will venture reads strangely to the contemporary reader.

But this strangeness isn't limited to ancient history. The deodand system was operating during much of the same time as the animal trials (between ~ 900 and ~1850 C.E.), and as with nonhuman animals it is possible to doubt the extent to which deodands were perceived as moral agents. In English law, a "deodand" was something someone owns (a chattel) that causes someone else's death. The deodand, and not its owner, would be criminally charged, and if found guilty it

would be “forfeited to the king”. The 1911 *Encyclopedia Britannica* entry for Deodand highlights some interesting features of the law.

> The law distinguished, for instance, between a thing in motion and a thing standing still. […] where a man’s death is caused by a thing not in motion, that part only which is the immediate cause is forfeited, as “if a man be climbing up the wheel of a cart, and is killed by falling from it, the wheel alone is a deodand”; whereas, if the cart were in motion, not only the wheel but all that moves along with it (as the cart and the loading) are forfeited.
>
> (Encyclopedia Britannica 1911)

The author of the encyclopedia entry examines how “Blackstone accounts” for this legal quirk and finds it wanting. But findings from the cognitive sciences might help us understand this distinction better. A great body of psychological work has shown that children and adults perceive collections of things that move together as a unitary object (Spelke, Principles of Object Perception 1990): as an individual thing. And, as we’ve already seen with our twins’ case, being an identifiable individual is also a feature of moral agent status. So, these quirks of the legal system, somewhat unintelligible in a positivist framework, might be reflecting some of the quirks of how the mind works.

The author of the 1911 Encyclopedia entry thought it was clear that “the imputation of guilt to inanimate objects” and “lower animals” was operative in the minds of the people involved in these cases and the crafters of the law.

Condescendingly they write that all this “is not inconsistent with what we know of the ideas of uncivilized races” (ibid). Scholars may wish to peruse skepticism about the extent to which this legal system was really a product of the relevant people perceiving the deodand as a moral agent. But I think we can begin to leave this behind us -since after all it isn’t the law that we’re interested, but the moral sense.

3.1.4: summary remarks on cases of moral agency

The range of entities that can at least partially be judged as moral agents seems wide. But we have to remember that we’re interested in the perception of moral agency, and so this wide variety could be explained if these entities were being perceived as all sharing features relevant to the perception of moral agency. It doesn’t matter so much to this portion of our investigation if a statue actually has a mind, if a piglet actually didn’t get a sufficient moral education, or if a tumor in just the right part of the brain effectively obliterates free will. What matters is how those entities are perceived by people making the relevant judgments.

So, this diversity might hide an underlying unity. That unity seems to involve an agent’s psychological capacities: does the agent have effective control over their own behavior; do they have malicious goals. I hope to also have motivated the hypothesis that moral agency is potentially treated differently by two different moral systems.

Next, I will review the phenomena relevant to moral patienthood. There again our methods and goals will be the same: can we decompose the concept into features; are there plausible hypothesis as to if moral patient status comes in degrees or kinds; and, generally, what is the psychological structure of this concept.

3.2: Moral Patients

A moral patient is an entity that deserves protection against arbitrary harm. It can also be said to have "intrinsic value" as opposed to merely "instrumental value."

3.2.1: Moral Patienthood in Humans

At first blush we might note that objection to homicide is a human universal (Brown 1991, Mikhail, 2010). However, societies seem to carve out special exceptions to this general prohibition. For example, the Kaulong people of the New Guinea highlands, up until the 1950s, strangled all newly widowed women. This obligation was customarily discharged by the widow's brother's-in-law (the brothers of her late husband). However, if there were none of those to be found, the widow's sons would fill in for their nonexistent uncles. Jared Diamond's account of one such case, as reported by *The Guardian*, is worthwhile

> "In one case, a widow – whose brothers-in-law were absent – ordered her own son to strangle her," […] "But he could not bring himself to do it. It

> was too horrible. So, in order to shame him into killing her, the widow marched through her village shouting that her son did not want to strangle her because he wanted to have sex with her instead." Humiliated, the son eventually killed his mother.
>
> (McKie 2013)

It seems unreasonable here to say that this son thought his mother wasn't a moral patient when he killed her. But carve-outs like this are narrowly specified: they are exceptions that prove the rule. And they seem to treat the intentional killing as obligatory. While it might not be a perfect tool for interrogating judgments about moral status, eliciting folk intuitions about when intentional killing is justified is a good one. It is ecologically relevant, both in the context of our environment of evolutionary adaptation and our contemporary environment. And judgments about which intentional killings are not wrong can serve to test hypothesis about our psychology (Daly and Wilson 1988).

In 2017 and again in 2019, Republican member of the Texas House of Representatives Tony Tinderholt introduced the "Abolition of Abortion in Texas Act". The first substantive section of the bill, section 2 (a), states. "A living human child, from the moment of fertilization upon the fusion of a human spermatozoon with a human ovum, is entitled to the same rights, powers, and privileges as are secured or granted by the laws of this state to any other human child" (Tinderholt 2017). When interviewed about the bill, representative Tinderholt told the Texas Observer, "I don't think that there should be any

exceptions to murder, no matter what" (Guarecuco 2017). It's clear what the moral claim is: murder is wrong, and killing a developing human is murder.

Historian and philosopher of biology Jane Maienschein argues that clarity about the developmental biology of embryos can serve to settle some contentious issues concerning their moral patient status (see Maienschein, 2014, for a comprehsive analysis of these issues). Strikingly, in light of our prior discussion, one of her objections to "embryo-personhood" laws, similar to those proposed by Tinderholt, is that embryos prior to a certain state of development don't necessarily count as a single identifiable individual: "individual embryos are not quite as neatly defined as the public perception would suggest" (Maienschein 2016, 135). This can be true in two ways. The first, perhaps predictably, is through twinning: where one embryo splits into two or more. The second, perhaps less well known, is the process of chimerization: where two embryos merge into one. Maienschein contrasts a biological embryo -the concept as its understood by the sciences- and the "public embryo" -the concept as understood by the folk- in order to question claims that the embryo counts as a (full) moral patient:

> In short, this biological embryo in its earliest stage is a bunch of cells. It is not organized yet, it is not expressing genes that cause differentiation, and up until what is called the blastocyst stage, it does not even grow larger. It is just a bunch of material cells that interact and interconnect. Its "meaning" is quite different from the newly conceived public embryo with its imagined emerging personhood that has evoked calls for protection.

(Maienschein 2016, 131)

Maienschein rightly draws attention to this biological feature of the embryo as relevant to moral patient status. If the claim that moral patient status is always attached to an identifiable individual is true, then increasing doubt about if an embryo is an individual should increase doubt as to if it's a moral patient. We can even frame this as a testable hypothesis: we should expect that after people's attention is drawn to the phenomena of twinning and chimerization, and the consequences for conceptualizing the embryo as an identifiable individual, their intuitions about the moral patient status of the embryo should change: specifically, they should be less likely to consider the embryo a moral patient and therefore more permissive about abortion.

It is important here to note that this effort to understand the mental representation of a concept has led us to a testable hypothesis about resolving moral disagreement. Certainly, this amounts to a question about persuadability – rather than effort to defend the idea that being an identifiable individual is itself a normatively justifiable criterion. However, if we can, by providing people with true information about an entity, change people's moral intuitions, we might prima facie say that is justifiable. We are not coercing their judgments after all. Nor deceiving them. Only educating.

In any case, diagnosing the sources of moral disagreement is one advantage I see in studying moral psychology. If we're lucky enough that the source of some disagreement turns on how a mistake of fact interacts with our

psychological machinery, then all we really need to justify edification is the conceit that a pleasant belief founded on false information is worse than a difficult belief founded on the truth.

There are certainly other grounds for contending that a developing human is or isn't a moral patient. It's often argued that the moral status of embryos depends on its sensible capacities: can it feel pain; how and to what degree does it suffer from injuries against it; etc. The pro-choice side of these debates sometimes[18] emphasis this directly: the entity doesn't suffer, at least not in a way comparable to the suffering of the mother caused by being forced to bring an unwanted child to term. Pro-life rhetoric often focusses on a capacities view of an embryo as well. They often argue that the embryo feels pain from the procedure. Though just as usually they just focus on the fact that the embryo has recognizable parts: a heart, fingernails, etc. And speak of it as having a soul and of it being an individual already known by God.

I'm being vague and general about the views of these two camps. Mostly because they are well known to a general audience. But also, because the particulars are not so important here. Instead, the takeaway should be that

[18] Though sometimes people in this camp appeal to medical privacy rights, I'm going to take the heart of the issue to still be the status of the developing embryo - after all there are no privacy rights that license us to kill extra-utero humans. So, privacy concerns seem to depend on prior judgments about moral patienthood. It is also argued that political liberation of women depends on reproductive freedoms like abortion. I agree. Yet this also seems to involve weighing the interests of some entities (embryos) differently than others (women). And while I find both of these approaches to have many merits, they presuppose judgments about moral patient status. And those are what I'm interested in here.

decomposing moral patient status into features (identifiable induvial, has sensible capacities, etc.) might be a profitable way forward. And that paying attention to these features can help us frame testable hypothesis about the psychology involved in judgments about moral patient status.

There is a countervailing consideration that shouldn't be ignored. Namely that practically, for policy purposes, people must assign an instrumental value to a human life and use that (rather than any notion of intrinsic value) to assess the cost/benefit of safety interventions. The paradigmatic example here is the Ford Pinto case. In the 1970s the Ford Company failed to install a small safety feature in their fleet of Pintos that would reduce rear-end explosions of the car. Ford estimated that while the fix itself was about $13 per car, the total cost of the fix across the whole fleet meant that they'd be spending more than $200,000 per life saved. And they decided against the feature on the grounds that that overvalued the life saved. Milton Friedman defended the principle, saying that "no one can accept the principle that a signal human life is infinitely valuable" (Friedman 1977). Treating humans in this way turns out to be a practical necessity. Currently the value of a human life, as assessed by the U.S. government, is approximately 10 million dollars (Planet Money 2020).

However, I think it's safe to say that treating humans in this way is intuitively objectionable to many people. And that that intuition is sourced in the view that some entities are moral patients, have intrinsic value, and tradeoffs

between them and instrumental goods are illicit[19]. John Mikhail ran a version of the trolly problem in which "five million dollars of new railroad equipment" would be destroyed by the trolly if an agent doesn't throw a switch, diverting the out-of-control train to a side track on which one man stood (Mikhail 2011, 165). He reports that his subjects viewed flipping the switch to be impermissible (ibid, 160-164). So at least we have some evidence that 5 million dollars isn't enough. Though that's currently half the instrumental value of a human life as priced by the US government. So perhaps isn't the best test for teasing apart intrinsic and instrumental value. We will return to how Mikhail deals with issues of moral status in Chapter Two.

3.2.2: Moral Patienthood in Animals

In July 2020, the online magazine aeon.co published "The face of the fish", an article by the bioethicist Michael Woodruff with the subheading "they're not cuddly, they don't behave at all like us – yet they are sentient. Why fish belong in the moral community." In the article the author argues that fishes are sentient and so experience pain and suffering. Which in turn, Woodruff argues, makes them worthy of moral consideration and protection: i.e., they count as moral patients. For Woodruff, "the question of moral status is tied up with the question of sentience: its definition and its application" (Woodruff 2020).

[19] This sets aside the issue of whether Friedman is right, and that it is normatively not justifiable to rely on this intuition of intrinsic value. We'll return to that in chapter 5.

For our purposes, "sentience" is as much of a black box for moral patient status as "responsible intelligence" was for moral agent status. According to the author, "sentience is the ability to have the feel of a sensory experience". And to "have the feel" here roughly points at the tradition in philosophy which has co-identified "consciousness" and/or "sentience" with "possessing" direct first-person experience. So, for example, we can imagine an entity, say a little metal windup toy monster with a mouth, biting into an apple. Intuitively, or so the argument goes, we would say that that windup toy monster didn't experience the taste of the apple. Because it didn't experience anything at all. So, if, for some entity it feels like something to have an experience, that entity is called "conscious" or "sentient"[20]. Where on the continuum of capacities from windup toy to six year old adult "sentience" emerges is the important unanswered question here. But since "sentience" so defined has no empirical implications, and so isn't experimentally tractable. So, answering the "when sentient" question isn't an empirical question. It's a set of decisions about what we value. But if we shift the question to "when is sentience perceived in another entity" that is experimentally tractable.

We might wonder what these labels give us, further than the claim that an entity has sensible experiences. And particularly for our purposes we might

[20] It is doubtful that sentience is a useful concept. There are no tests for sentience because as a hypothesis it makes no testable predictions. A philosophical zombie and a normal human agent produce all the same behaviors by definition. Such concepts should be abandoned as soon as more tractable alternatives can be found.

wonder which sensible experiences are relevant to the perception of moral patient status. Obviously, arguments in favor of the moral patient status of animals are acutely relevant to the practice of eating them. I cannot hope to exhaustively review the possible positions here, sorting through the arguments looking for features that might count as qualifying criterion for moral patient status. But it is worth calling attention to at least two major positions: utilitarian and rights-based arguments for moral patient status.

The first is attributable in its contemporary incarnation to Peter Singer (Singer 1975). Signer's criterion relevant to moral patient status is unitary: can it feel pain/pleasure (Singer 2009). Certainly, different entities can experience different kinds of pain and pleasure; and some entities can experience more than others. And those differences matter for Singer. But he insists that all animals are equal (Singer, All Animals Are Equal 1974), if they possess the relevant capacities in the same proportions. So instead of sentience we have something a bit more tractable: can an entity feel pain.[21]

I'll take the philosopher Tom Regan to be a fairly representative example of the rights-based view. His word for moral patient status is "inherent value" and he says that any entity that is a "subject-of-a-life" has inherent value. Not willing

[21] It's obvious that I've smuggled back in the problem of direct first-person experience here with the word "feel". My own view is that we can change this question to, "according to our best science, does this entity feel x" and get along just fine (see for example **Invalid source specified.**). But insofar as I'm interested in people's perceptions about the minds of other entities, if those entities actually do feel is less relevant.

to leave "subject-of-a-life" as a black box, Regan spells out exactly what he means:

> "Individuals are subjects-of-a-life if they have beliefs and desires; perception, memory and a sense of the future, including their own future; an emotional life together with feelings of pleasure and pain; preference- and welfare-interests; the ability to initiate action in pursuit of their desires and goals; a psycho-physical identity over time; and an individual welfare in the sense that their experiential life fares well or ill for them, logically independently of their utility for others and logically independently of their being the object of anyone else's interests."
>
> (Citation taken from Lengauer 2020, 94)

Here Regan is especially eager to sperate moral patient status from an entity's instrumental value. The distinction between inherent value and instrumental value are in the mind of the beholder. And if there is a difference between them psychologically, that has to be established. That is, at least in our context, it is an open empirical question whether there is a psychological difference between intrinsic versus instrumentally valuing a thing. So that the difference between the value placed on a car (because it is useful to its owner who invested scarce resources in acquiring it) and the value placed on a pet dog (because it is the subject-of-a-life) involves a difference in kind of valuation, rather than merely a different degree of it. This this distinction certainly has a good deal of intuitive plausibility, and we will return to this concern in Chapter Four.

3.2.3: Moral Patienthood in Objects

It is quite unusual to find instances of inanimate objects treated as moral patients. It is typical to treat objects as having instrumental value: i.e., they have value because they're valuable to someone. It is unclear to what extent instrumental value plays a role in our moral psychology. We might want to insist that concepts like theft depend on it. To make this claim we'd need to assert that taking something from someone that belongs to them, but is of no value to them, is not theft. That if something has no value to someone, they would not be upset if that thing was destroyed or otherwise no longer available to them, by whatever means. However, this seems doubtable. It could be that people are offended by theft even in these circumstances. But it also seems stilted to suggest that this is because harm was committed again the stolen object. Instead, that anger is over being stolen from at all. These are empirical issues which I did not test, so I will leave this issue aside for now. However, I haven't been able to discovery a system, akin to the deodand system, that treats inanimate objects as moral patients. Though of course the god of the Bible gets angry at Moses for hitting a rock (Numbers 20:11-12). So, all things are possible.

3.2.4: Moral Patienthood in the Environment

In late 2016 CNN reported on protests against the construction of Dakota Access Pipeline. The protestors, largely Amerindian, were concerned over the

path the pipeline takes through tribal lands. Calling themselves "water protectors" they insisted that the path of the pipeline put water ways they view as sacred in danger: particularly the Cannonball River, and other Missouri River tributaries. They explained that "Water […] has memory. When people speak or sing to it during a ceremony, […] the water holds on to what it hears and can later share what it learns" (Ravitz 2016). Faith Spotted Eagle, a 68-year-old Sioux woman interviewed for the article, spoke about the need to protect the river: "One hundred years from now, somebody's going to go down along the Cannonball River and they're going to hear those stories. […] They're going to hear those songs. They're going to hear that memory of what happened here at this camp" (ibid).

The water protectors are our contemporaries. And they take very seriously their obligation to protect a river. In our language, they seem to be treating it as a moral patient.

Similarly, in New Zealand the Whanganui river has been granted legal personhood **Invalid source specified.**. From March 2017 to July 2017 there was a fierce legal dispute in India over the legal status of the Ganges and Yamuna rivers. The Supreme Court of India ended the argument by overruling the early order by the Uttarakhand state High Court which had granted the rivers legal status (BBC World News 2017a, 2017b). Taking a different position, in 2018 Supreme Court of Colombia "declared that, "for the sake of protecting this vital ecosystem for the future of the planet," it would "recognize the Colombian

Amazon as an entity, subject of rights, and beneficiary of the protection, conservation, maintenance and restoration" (Bryner 2018).

Some environmental ethicists argue that parts of ecosystems do in fact have intrinsic value. See, for example (McShane 2007). Here I want to leave open the possibility that perception of moral patient status doesn't automatically exclude entities like rivers, so long as they're perceived in the right ways: as a single individual thing; as having other features reliant to moral patient status, whatever those features turn out to be.

3.2.5: Moral Patienthood in moral agents

In 1989 *Star Trek: The Next Generation* aired an episode called "The Measure of a Man" (Snodgrass 1989) in which one of the shows primary characters, the android Data, is ordered to report to a governmentally run robotics lab for experimentation. Asked to leave his home and work, he demurs. When ordered to comply, Data chooses to resign instead. However, the antagonist of the plot – doctor Bruce Maddox - insists the android is not "sentient" and so does not deserve "the right to choose" to resign. A court case ensues, in which Data and counsel are asked to prove the android is sentient. Worthy arguments aside, in deciding the case the jurist in charge says "Is Data a machine? Yes. Is he the property of Star Fleet? No. We've all been dancing around the basic issue: does Data have a soul? I don't know that he has. I don't know that I have. But I have

got to give him the freedom to explore that question for himself. It is the ruling of this court that Lt Commander Data has the freedom to choose."

Moral controversies in speculative fiction often put problems of moral patient status center stage. This is one of a number of *Star Trek: The Next Generation* episodes focusing narrowly on issues of moral status in nonhuman entities[22]. But the example of Data in this episode is common in commentary on the moral status of sufficiently intelligent computers. One feature of this vignette that is rarely commented on is that Data's capacity for agency is relevant to his moral patient status. In the episode, if Data has the capacity for consciousness, then he is entitled to choose what he does with his life and time. That is to say, if he is a moral patient, then obstruction his agency is wrong.

But we can't be too quick here: if we take this seriously, it isn't just that Data is judged to be conscious, after all many nonhuman animals certainly are

[22] In "Lonely Among Us' (Season1, episode 6) a human character, when asked to provide live animals as food to an alien guest, replies "We no longer enslave animals for food purposes." In "Home Soil" (season 1, episode 18) terraformers want to ignore evidence of "silica based life" so as to continue their colonial projects until the crew of the Enterprise discover the coverup and intervein; in "Evolution" (season 3, episode 1) a science experiment gone wrong releases small robotic replicators which, through replication and differentiation, evolve into a sentient colonial robotic organism that threatens the crew and other important science projects – and still the crew of the enterprise work to protect the new "lifeform"; when asked to kill an interstellar lifeform, called the "crystalline entity", that has killed many humans Picard, captain of the Enterprise and notorious humanist, draws an analogy: "the sperm whale on Earth devours millions of cuttlefish as it roams the oceans. It is not evil. It is feeding!" ("Silicon Avatar" season 5, episode 4); in "The Quality of Life" (season 6, episode 9), Data becomes convinced that repair robots called "exocomps" exhibit creative problem-solving capacities and therefore cannot be sacrificed in order to save other crewmembers' lives.

conscious. It isn't that Data has goals and desires. Many nonhuman animals have goals and desires. And frustrating those goals is… well, frustrating for them. So, at some change point an entity is viewed as a patient for whom obstructing their free will would be wrong. And discovering that change point is acutely relevant to our stated goals.

3.3: Summary remarks on moral agent-patient illustrative cases

So far, I hope to have made clear what the phenomena in the world is that I'm attempting to explain. I hope I've shown you that moral agents-patients can seemingly come in all shapes and sizes. I've attempted to draw attention to potential features of unity, that might underwrite the diversity. We've entertained the idea that moral agent-patient status might be treated differently by different parts of the moral faculty. And we've seen a number of ways in which being an identifiable individual seems to be particularly important.

Additionally, when considering moral patient status, we had to include a consideration of intrinsic versus instrumental value. I want to here suggest that moral psychology should be the grounds on which we draw this distinction between the kinds of ways we value a "done-to" entity. How the mind represents these concepts matters. If the mind doesn't bother with the distinction when churning out the relevant judgments, we might reasonably expect a bit more justification for introducing the ideas into our moral philosophy. And if the mind

treats these two modes of valuation completely differently, it puts a bit more meat behind claims that the two value measures are incommensurable normatively.

But the main goal of this chapter was to introduce you to the kinds of phenomena that needs to be explained by any descriptively adequate theory of moral status. In Chapter Two I outline the efforts by psychologists, linguists, and experimental philosophers in this respect.

CHAPTER 2

PSYCHOLOGICAL ACCOUNTS OF MORAL STATUS

1. Is there a specialized mental system that generates moral status judgments

> Like all other domains of knowledge, our moral knowledge did not evolve in a neural vacuum, isolated from other processes. Further, our moral behavior depends upon other systems of the mind. What we are after is a description of those processes that are specific to morality as well as those that are not specific but play an essential supporting role.
>
> (Hauser 2006, p. 50)

If someone is bold enough to claim the mind contains a specialized mechanism functionally organized to make judgments about moral status, then questions about an elegant integration of that mechanism with other mental systems becomes acute. Is this mechanism really specific to morality? Or does it play an essential supporting role?

It might be useful to draw on some terminology from linguistics here. (Hauser, Chomsky and Fitch, The Faculty of Language: What Is It, Who Has It, and How Did It Evolve? 2002) propose that we make a distinction between the broad faculty of language (FLB), capacities used by language but not unique to it, and the narrow faculty of language (FLN), capacities unique to it. This distinction was motivated by their concerns that people exploring question of the evolution of

language often conflated these things. According to them, the FLB contains capacities like vocal imitation. Obviously, that's an important capacity with respect to human language learning and use. But it isn't unique to language, we can imitate nonlinguistic noises too. Hauser et al argue that the only capacity unique to language is "core syntax" or recursion[23]. But for our purposes this debate doesn't matter as much as the distinctions it caused the authors to draw. One way to ask our driving question is, is moral status part of the broad moral faculty (FMB), or the narrow (FMN)? Is it unique to moral cognition, does it simply play a supporting role? Answering that requires specifying a cognitive model of social psychology with a good deal of precision. I touch on two approaches here that I think have considerable merit.

2. Constructivist approaches: aka, moral status in the FMB

The psychologist Kurt Gray has proposed an elegant and persuasive answer to this question. He and his colleagues argue that moral judgments are formed by a domain general dyadic template fed by other-mind perception (Gray, Young and Waytz 2012, Schein and Gray, Chelsea Schein1 and Kurt Gray 2017): "we suggested that the diversity of moral judgment is underlain by the moral dyad, a psychological template of two perceived minds—a moral agent and a moral patient" and furthermore they suggest that this template is the essence of moral cognition (2012, p. 205-206). We saw evidence that seems to corroborate

[23] This claim is hotly contested (Jackendoff 1997) (Jackendoff and Pinker 2005).

this hypothesis already. In case after case mentioned in Chapter One, what seemed to be at stake was the minds of the actors involved. Did the shooter have a brain tumor? Does a piglet know right from wrong? Our exculpatory concerns are often rooted in factual questions about the kinds of minds of an agent or patient.

On this view, the extent to which an entity qualifies as a moral agent is the extent to which that entity is perceived as having a range of particular psychological capacities, e.g., self-control, planning, memory, and the like, that allow it to act intentionally. And, that the extent to which qualifies as a moral patient is the extent to which that entity is perceived as having psychological capacities, like fear, pain, pleasure, joy, embarrassment, etc., that allow it to experience things. Using factor analysis, Gray et al lump a host of psychological capacities into two dimensions of mind perception: "Agency" and "Experience" (Gray, Gray and Wegner 2007). "Agency qualifies entities as moral agents, those who are capable of doing good or evil, whereas experience qualifies entities as moral patients, those who are capable of benefiting from good or suffering from evil" (Gray, Young, & Waytz, 2012, p.104).

To be clear, this is opposed to the view that we make judgments about an entity's capacities based on the evidence provided by their behavior. The psychological ques are more closely connected to the rapidity of behavior (Wegner and Gray 2016) or the kinds of bodies an entity has (Knobe 2011). Again, humans do not appear to take a scientific or empiricist approach to their judgments about the kinds of minds other entities have. Instead, our minds extract

the values of some variables (speed and goal directedness of behavior, type of body, etc.) in order to generate intuitions about an entity's capacity for Agency or Experience.

Gray and his colleges have developed this view and now call it "the constructionist Theory of Dyadic Morality (TDM)" (Schein and Gray, Chelsea Schein1 and Kurt Gray 2017, 1). On this account "acts are condemned proportional to three elements: norm violations, negative affect, and — importantly— perceived harm. This harm is dyadic, involving an intentional agent causing damage to a vulnerable patient" (Ibid). The mechanisms involved in producing moral status judgments are "domain-general cognitive processes … used in other decision making paradigms. […] Rather than involving definitions and modules, research reveals that categorization decisions are made by comparing examples to cognitive templates" (id, p. 17).

On this view other mind perception provides the input conditions for all moral status judgments (Gray, Young and Waytz 2012). There are certainly many merits to this view. It has prima facies plausibility given the kind of review of example cases we touched on in Chapter One. The view is also congenial with respect to the history of philosophy, which has a long track record of arguing that the relevant features to judgments about moral status are all psychological (some contemporary examples, Carruthers, 1992; Korsgaard, 2004; Regan, 1989; Singer, 2009). It has experimental successes. For example, if subjects are given a rich description of the mind of a suffering lab animal, they'll be less likely to think its

use in the lab is permissible (Sytsma & Machery, 2012). And normatively the capacities approach to moral status has been called "the only game in town" (DiSilvestro 2010). But of most concern to us here will be that the model helpfully has clear test implications.

The constructionist view is an argument in favor of moral status judgments playing an essential but supporting role in moral cognition; i.e., being part of the broad faculty of morality (FMB), not the narrow (FMN). "Constructionism suggests that psychological phenomena emerge from the combination of more basic [psychological] ingredients, rather than through the operation of distinct mechanisms (Schein and Gray, Chelsea Schein1 and Kurt Gray 2017, 24)." So, in fact, it is consistent with the view that there isn't anything at all in the FMN: that there's nothing unique to moral cognition that isn't also used in social event psychology more generally.

The constructivist view was explored in this context by the philosopher Joshua Knobe, and developmental psychologists Brent Strickland and Matthew Fisher. They made this argument with the slogan "moral structure falls out of general event structure" (Strickland, Fisher and Knobe 2012). They argue "that the role assignment one finds in moral cognition can be explained in terms of a far more general theory about how people make sense of events. […] people show a quite general tendency to construe events in terms of agents and patients" (id, 198-199).

In order to demonstrate this, Strickland et al set out to show that the moral

dyad is a product of what is sometimes called core knowledge (id., 205). Core Knowledge (Spelke and Kinzler 2007; Spelke, Perscom, March 2016) is a theory of how the mind works/learns, proposed by developmental psychologists Elizabeth Spelke. Spelke argues that "humans are endowed with a small number of separable systems of core knowledge. [and that] New, flexible skills and belief systems build on these core foundations" (2007, 89). And that "core systems for representing objects, actions, numbers, places, and social partners may provide some of the foundations for uniquely human cognitive achievements" (Ibid). So, for example, we have a number sense (Gelman and Gallistel, 1986; Stanislas 2011), and out of that number sense plus the language faculty we can construct a counting algorithm (1, 2, 3, …n | all increasing by n+1). Not all peoples have counting systems. But all peoples have a number sense, and the language faculty.

The constructivist view is that the "moral" in moral agent/patient is just a way of saying that the core knowledge system that guides learning and judgments about social partners is being used during moral cognition. In order to defend this view, that "the structure one finds in moral cognition actually falls out of a structure one finds in event cognition more generally", Strickland et al look "to the literature in theoretical linguistics" (Strickland, Fisher, Knobe 2012, 199, original emphasis). In practice this turned their question into, is there something special about moral agents and moral patients as opposed to thematic AGENTs and PATIENTs?

Their way of reducing the moral dyad to core knowledge was to show that

the moral dyad was the name for an effect seen when the thematic AGENT of a sentence is involved in a morally charged event: "Across the three variables associated with agency (intentionality, responsibility, punishment) participants consistently assigned higher ratings to the person who appeared as the grammatical subject[24] than to the person who appeared as the grammatical object" (id, 204). So, the thematic AGENT in a morally charged event, they say, is always rated as more intentional, more responsible, more worthy of punishment. These are the sentences that generated their results:

(1) a. Steven is 25 years old and Kate is 15 years old.

b. Steven French kissed Kate.

(2) a. Steven is 25 years old and Kate is 15 years old.

b. Kate French kissed Steven.

(3) a. Steven is 35 years old and Kate is 25 years old.

b. Steven French kissed Kate

(4) a. Steven is 35 years old and Kate is 25 years old.

b. Kate French kissed Steven

(id, 202. renumbered)

Participants were asked questions about how responsible they thought Steven and

[24] strictly speaking we can doubt this interpretation since they do not passivize their test sentences and so confound AGENT and subject as well as PATIENT and object. But a generous interpretation of them is that the conflation is harmless given their experimental methods. But here instead of grammatical subject they actually mean thematic AGENT.

Kate were, how intentionally they acted, etc. And then compared the results of people's judgments of (1b) with (2b) and (3b) with (4b). Strickland et all say that (1) and (2) are "morally charged" because the actors would "be breaking many U.S. state laws"[25] whereas (3) and (4) were "morally neutral" because not against the law (id, 201). I have several objections to cases like this[26][27] but I'll leave them aside and focus only on if, waving those objections, this evidence supports the view that the moral dyad is just thematic roles AGENT and PATIENT in a moral context.

The authors conclude from the fact that people view the AGENT of a morally charged event as more morally culpable that "it appears that people's use of these roles [agent/patient] is not at all restricted to the moral domain" (id, 205). Yet even so I have strong doubts that these authors have showed that moral agents/patients and thematic AGENTs/PATIENTs are identical. All these authors

[25] It is worth noting that this claim, that the kissing would break state laws, is not supported by the authors with any evidence; at least in Arizona and California this action breaks no state laws. See ARS §13-1410, §13-1401, and §13-1405 for Arizona; See California Penal Code §647.6, §288a, §289. The only place sexual kissing appears to be considered part of a sex crime against a minor is in The Netherlands (see Article 245).

[26] I prefer vignettes to simple sentences; we are investigating a perceptual system of the mind, and therefore contextualized events (or acts), rather than simple sentences, are the right size and shape for the operation of our moral system: "stand-alone sentences that allegedly have moral or non-moral content (e.g. 'The elderly are useless' versus 'Stones are made of water'), rather than acts of conspecifics that can be carefully manipulated to test specific theories of mental representation [… are] poorly motivated" (Mikhail, 2007).

[27] I don't myself share the intuition that (1) and (2) are morally charged whereas (3) and (4) are not. I simply don't view (1) or (2) as obviously involving harms. Unless we learned that one or the other of the individuals was unhappy about the kissing, which would be a different issue.

are entitled to conclude from their study is that the thematic roles are relevant to the moral dyad in that the roles seem to overlap significantly. AGENTs are a lot like moral agents in that they can be said to be the cause of intentional or unintentional action and the like, just as moral patients are a lot like thematic PATIENTs in that they're both the thing effected by the action of the verb.

To see why I think the authors are not successful in eliminating the moral dyad by reducing it to the linguistic thematic system, consider the following sentence,

(5) a. Steven is a 25 year old dog, Kate is a 15 year old woman.

b. Steven French-kissed Kate

c. Kate French-kissed Steven

Changing Steven to a dog dramatically changes our judgments here. I think b. is laughable and c. is weird maybe gross. But neither of them strikes me as immoral. Of course, changing Steven from a human to a dog isn't just changing one thing, it's changing a whole lot. Including, especially, how we think about the entity [Steven]'s mind, how Kate would likely feel about the kissing, and how it would be viewed by onlookers. Instead of changing the position in a sentence of two entities both assumed to have full moral status, we should investigate if moralized verbs can be applied to any entities whatsoever. That is, we should look at verbs that describe interpersonal harm and which lexicalize moral disapprobation: e.g., murder, assault, rob, and rape.

(6) The man killed the cow.

?(7)[28] The man murdered the cow.

(8) The tornado killed the man.

#(9) The tornado murdered the man.

(10) The man killed the bear.

(11) The bear killed the man.

?(12) The man murdered the bear.

#(13) The bear murdered the man.

Strickland et al's hypothesis would seem to suggest that being the subject of these sentences is sufficient to account for moral agent status. Yet if this is true what accounts for the discomfort with the use of the word murder in (7, 9, 12, and 13). If moral status is simply thematic role, why is it that a bear (a perfectly good thematic AGENT and/or PATIENT) can't easily participate in sentences with the moralistic verb murder?

If we apply the moral status as mind perception hypothesis that is at the heart of Gray's TDM, judgments of (7, 9, 12, and 13) should vary predictably based on how sophisticated the minds of bears, cows, and tornados, are perceived to be by the people making the judgments. Perhaps because bears are lower in both perceived agency and experience (Gray, Gray and Wegner 2007), they're

[28] I will use? for questionable semantic reading/acceptability and # for semantic unacceptability.

less capable of being moral agents and moral patients. Murdering a cow might be an acceptable sentence for some populations (perhaps religious Hindus and some moral vegetarians, if those views are pegged to the perceptions of the entities mental capacities by those people) but not for English speakers more generally. If moral status is fed by mind perception, then people who judge (7) to be acceptable must have a rich view of a cow's capacity for experience. Likewise, a tornado murdering a person should only be acceptable to committed animists who perceive the phenomena to have the capacity to intend its own behavior.

But we don't here have an answer as to why the verb murder only applies to some entities and not others. What's the dividing line where a judgment of immoral killing licenses the use of the word murder? And is that line drawn by reference to the perceived minds of the entities involved?

2.1: The Causal Structure of The Constructivist Approach to Dyadic Morality

Sketching the testable predictions of this hypothesis provided me with some additional reasons for skepticism that the moral dyad can be reduced to social event cognition more generally. If moral status just is mind perception, then the difference in treatment between dogs and children, with respect to judgments of praise and blame, must be caused by a difference in mind perception. However, according to Gray, Gray, and Wegner, children and dogs receive basically the same ratings in terms of Agency and Experience (Gray, Gray and Wegner 2007).

Yet we obviously treat them radically differently with respect to their moral status.

While Gray and his colleges sketch the casual structure of some parts of their TDM (see again (Schein and Gray, Chelsea Schein1 and Kurt Gray 2017)), to my knowledge no one has sketched the mind perception as moral status hypothesis very clearly. Remember that the TDM involves norm violations, the moral dyad (intentional agent and vulnerable patient), and moralistic emotions. Graphically

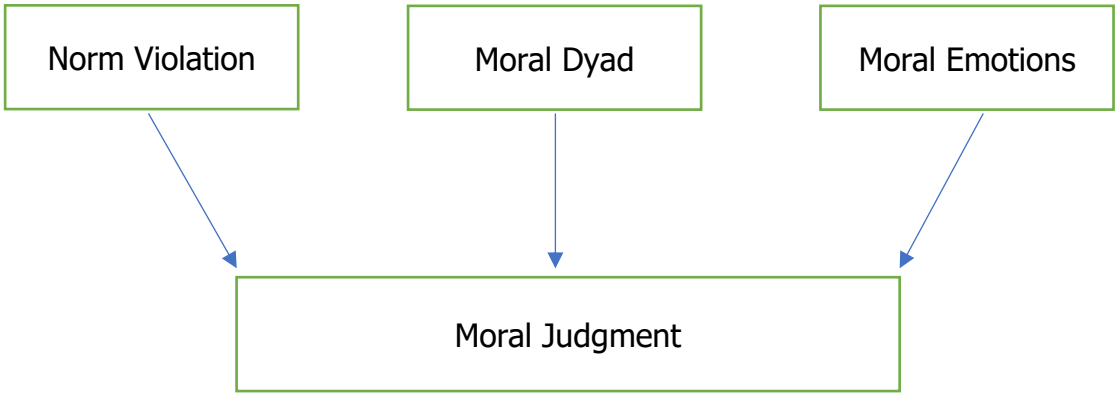


Fig. 1: Broad Causal Structure of The Constructivist Approach to Dyadic Morality

For Gray, the "moral dyad" box above must look something like this:

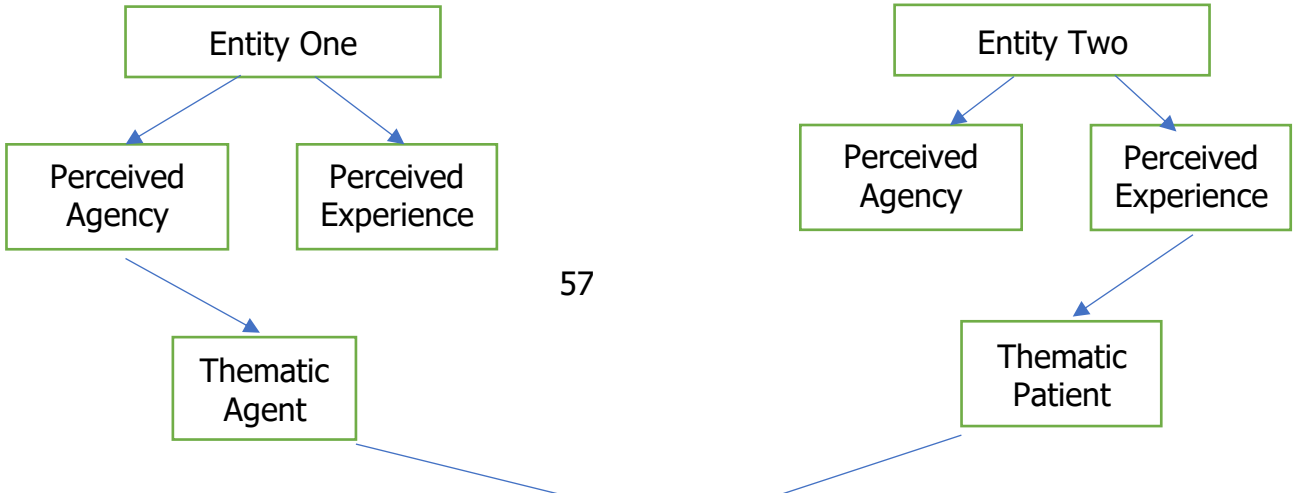


Fig. 2: Causal Structure of Gray's Moral Dyad

Here we can see that only the perceived agency, of the entity that plays the role of thematic AGENT, feeds the moral dyad; and likewise, only the perceived experience of the entity that plays the role of thematic PATIENT feeds the moral dyad. While this seems to have the benefit of being a reasonably accurate characterization of Gray et al's view. It doesn't incorporate the insights of (Strickland, Fisher and Knobe 2012), that simply playing the role of thematic agent/patient can change the perceived minds of the entity in question (i.e., that there's an interaction between thematic role and perceived mind). It also excludes alternatives such as (Sytsma and Machery 2012), who argue that moral patient status is fed by both perceived Experience and Agency. A richer graph in line with this work might look like the following

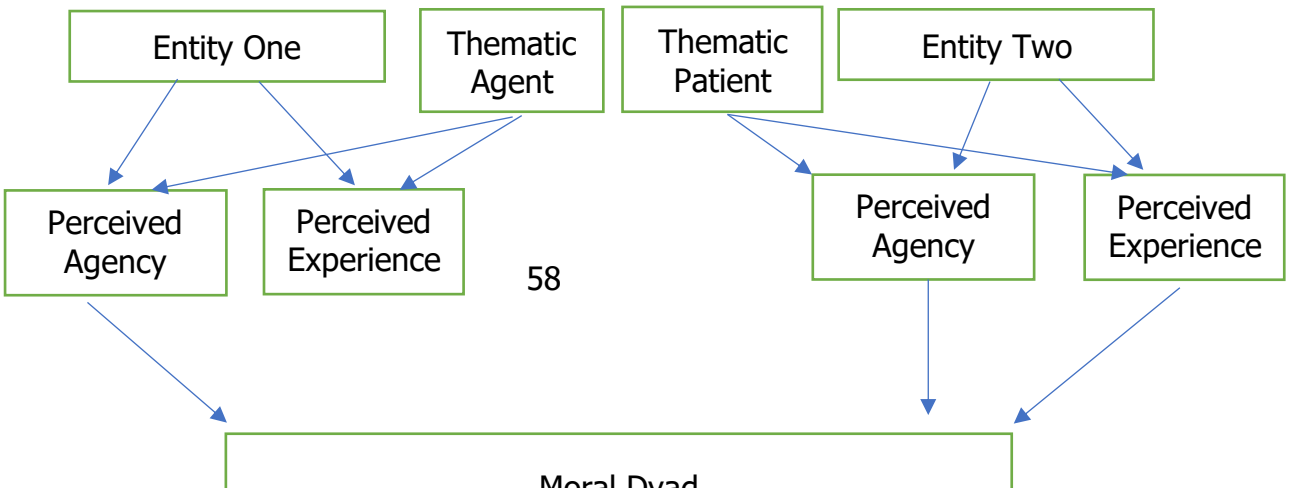


Fig. 3 Casual Structure of The Moral Dyad and Thematic Role

Entity one and two here are of course stand ins for the ques of other-mind perception (Gray, Gray and Wegner 2007; Wegner and Gray 2016). So, let's instantiate the graph with the kitten scratches man example: "Kitten (low agency) scratches the man (low experience)? Not immoral" (Wegner and Gray 2016, 41).

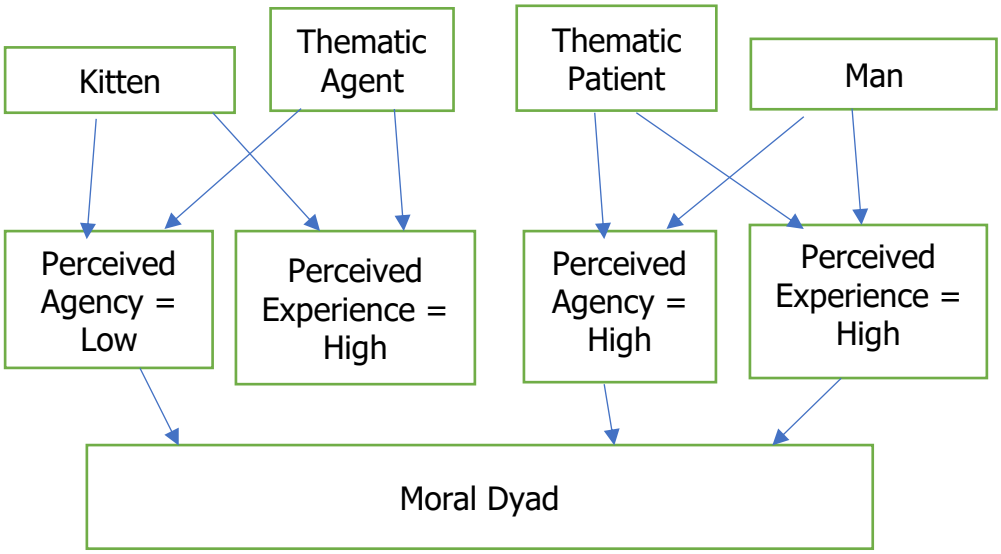


Fig. 4: Detailed Causal Structure of the Moral Dyad

I have departed in one respect from the kitten scratches man example as given by Gray et al. Because, for reasons I cannot decern, they state that the man here would be rated as having "low experience". Yet this is contrary to their data presented in the same work (Wegner, Gray 2016, 30), and mine (presented in Chapter Three): a generic man is always rated as having a high capacity for experience[29]. The boxes for thematic agent/patient both feed perceptions of other minds, in line with (Strickland, Fisher and Knobe 2012) and my own work (again, Chapter Three) which show that the thematic role an entity plays does shift

[29] Though this claim may be motivated in part by Gray's view of moral typecasting **Invalid source specified.**. Gray argues that moral agents are typecast as agents, and so judged to be more agentive in subsequent events under analysis. For now, we can set this concern aside.

agency/experience judgments. And the arrow from the thematic patient [the man]'s perceived experience to the moral dyad represents the "two sources" of moral patiency hypothesis defended by (Sytsma and Machery 2012).

Box and arrow diagrams like this can clarify testable predictions of different hypothesis. The first thing that we must notice is that perceived mind is the direct cause of moral status judgments. And that the entities (here, man and kitten) and the thematic role they play (agent/patient) are the direct causes of the perceived minds. An inference of the TDM is that if one entity (say, a man) is perceived as having sufficient Agency for moral agent status, then any entity perceived as having the same (or greater) capacity for agency will also be judged a moral agent. Likewise, if one entity (say, a kitten) is perceived as having sufficient experience for moral patient status, then any entity perceived as having the same (or greater) capacity for experience will also be judged a moral patient.

In (Wegner and Gray 2016) the authors introduce the term cryptomind, which will be useful to us from here on. A cryptomind is an entity whose mind we're iffy about: how much Agency does a dog have? How much does an android have? How much Experience? We're generally not iffy about the minds of other humans (almost everyone judges them to have as high Agency and Experience capacities as it is possible to), so for the most part a cryptomind is just a term for the minds of non-human entities, irrespective of whether they exist or not. Remember this is a matter of perception not fact – dogs do exist and there's presumably some fact of the matter about their agentive and experiencing

capacities (see, for example, (Wynne 2019)). Androids don't exist and so there's really no truth (relative to the real world) about if they're extremely agentive or not.

An important step in creating a testable hypothesis here means we require a reliable measure of perceived moral status for cryptomind. My proposal is the following: "murder" is the wrongful killing of a moral patient by a moral agent. It is an interpersonal harm, which according to TDM, is the very center of morality (Schein and Gray, Chelsea Schein1 and Kurt Gray 2017). That means that judgments of who can murder, and be murdered, should be explicable on the TDM account. Particularly, it should be explicable by reference to the perceived minds of the entities involved; i.e., their Agency and Experience capacities. I put this measure into my experiments, reported on in Chapter Three. But for now, I want to turn again to the question of if there's any alternative to the constructionist view that there's no domain specific psychology involved in moral status judgments. That the moral dyad is in the FMB, not FMN. That it is not constructed from core knowledge, but one of the basic elements on the periodic table of moral cognition.

3. Moral Grammar and Mikhail on Moral Status and the FMN

The legal scholar and psychologist John Mikhail is well known for his defense of a theory of universal moral grammar (UMG) (Mikhail 2000, 2007, 2011). UMG seeks to explain human moral judgments by using "concepts and

models similar to those used in Chomsky's program in linguistics (Mikhail 2007). According to UMG, all humans are equipped with a moral instinct, a kind of specialized learning mechanism, which enables them to acquire moral knowledge[30]. Humans, after all, are moral animals (Hauser 2006). Or at least moralistic ones. We delight in passing judgments, if popular media is any indication of the things in which we take delight.

Some work in developmental psychology seems to support this innatist view of moral cognition. For example, the Kiley Hamlin has argued that preverbal infants moral judgments "do not stem from [the] infants' own experience with the actions involved" (Hamlin, Wynn and Bloom 2007, 559).

> From extremely early in life, human infants show morally relevant motivations and evaluations—ones that are mentalistic, are

[30] For Mikhail, moral knowledge means a decription of the operative principles of our moral sense. Mikhail draws a distinciton between operative and express moral principles. Express principle's are basically moral rules that we can talk about and reflect on. For example, rules against scientisfic misconduct prohbit falisfication, fabrication, or plagerism. These are express rules. They're even instantiated by federal policy. Operative principles, however, are a basic component of the computational representational theory of mind. They're the mechanisms that build and manipulate data structures. And for Mikhail the project of moral psychology is to discover which operative principles are at work in the production of our moral judgments (as well as which data structures they're operating over). The difference between operative and express principles is also evidence in the phenomena of moral dumbfounding (Haidt 2001, Dwyer 2009, Hauser, Cushman, et al. 2007). Express principles are the (often deficient) reasons we give for our moral judgments; operative principles are the psychological causes of our moral judgments.

> nuanced, and do not appear to stem from socialization or morally specific experience. Indeed, these early tendencies are far from shallow, mechanical predispositions to behave well or knee-jerk reactions to particular states of the world: Infants' moral inclinations are sophisticated, flexible, and surprisingly consistent with adults' moral inclinations, incorporating aspects of moral goodness, evaluation, and retaliation.
>
> (J. K. Hamlin 2013, 191)

Or as Mikhail puts it, "the intuitive jurisprudence of young children is complex and exhibits many charecteristics of a well-developed legal code" (2011, 104).

On Mikhail's view, moral judgments call out for domain specific computational apraratus for the same reason linguistic judgments do. Human moral judgments are "spontaneous, stable, stringent, and highly predictable," and also "appear to be highly structured, amenable to constituent manipulation," as well as "consistent with judgments rendered in structurally similar scenarios" (Mikhail 2011, 82-85). Mikhail argues that moral psychology faces a "projection problem that is, the problem of explaining how ordinarily individuals are capable of applying their moral knowledge to new and often unprecedented cases" (ibid, 30). And because our moral judgments don't "stem from socializaition or morally

specific experience", Mikhail argues that moral psychology faces a poverty of the stimulus problem (Mikhail 2008).

Many moral psychologists, including Mikhail (2000, 2011), view moral psychology as a kind of perceptual psychology (Haidt 2001, Cushman, Young and Hauser 2006, Hauser, 2006, Hauser, Cushman, et al. 2007, Dwyer, Huebner and Hauser, 2010). So good and bad behavior is not particularly what is being targeted for explanation here. Instead, moral judgments are the explanandum, to be explained by reference to a mental model.

Unlike Gray's TDM, which insists that "categorizations are not made with strict verbal definitions containing a list of necessary and sufficient conditions" (Schein and Gray, Chelsea Schein1 and Kurt Gray 2017, 17). Specifying a moral grammar is an "attempt to state necessary and sufficient conditions for assigning a deontic status to a given act or omission" (Mikhail 2011, 124). Unlike TDM with its domain general mechanisms, a moral grammar is domain specific.

Mikhail's formal model, proposed in his Ph.D. dissertation and expanded upon in his 2011 book, is quite involved. He proposes a cognitive model for all social event psychology, and then enriches that with specialized normative data structures and operations. The figure below is a recreation of his "Expanded Perceptual Model for Moral Judgment" borrowed with permission from (Mikahil, 2011, 114).

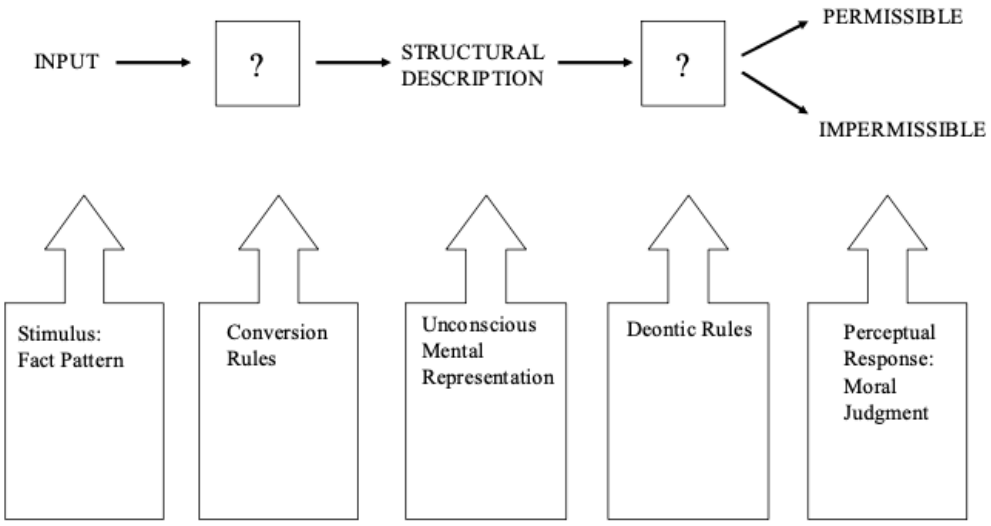


Fig. 5. Mikhail's Expanded Perceptual Model for Moral Judgments

This model means something like this: when we see, hear, or read about a morally salient event, the mind converts that "fact pattern" into an unconscious mental representation or "structural description." This might at first seem like an overcomplicated way of talking about social event psychology. That the mind has to represent a witnessed (or hearsay) event is uncontroversial. However, Mikhail meticulously argues that the right way to characterize this psychology is with act-trees[31] (Mikhail 2000, 2011, Levine, Leslie and Mikhail 2018), that encode at

[31] Act trees are "two-dimensional tree diagram[s..] successive nodes of which bear a generation relation to one another that is asymmetric, irreflective, and transitive" (2011, 118).

least a temporal structure,[32] a causal structure,[33] a moral structure,[34] and an intentional structure[35] (Mikhail 2011, 118-120, 162-175). "Deontic rules" operate over the structural description to produce a perceptional response in the form of an ethical intuition or moral judgment.

Where in this model does moral status fit? And what parts of this model could fairly be called part of the FMN. The best way to work through these questions is by attempting to apply the conversion rules in Mikhail's model to an example. Unfortunately, our "Kitten scratch Man" example borrowed from Wegner and Gray won't do because there's just not enough information about the event to make walking through the details useful: (did the cat intend to scratch the man?, did it do cognizable damage to his person?, etc.). Also, while Mikhail's model purports to be able to handle all interpersonal harms, his experiments (concerning judgments about trolly problems) all involve someone's death, and so his formal work often incorporates that. Here I'll present one of the trolly problem I gave to subjects in Mikhail's framework, for expositional purposes.

3.1 Applying Mikhail's conversion rules to a fact pattern

---

[32] This structure involves an indexing of the parts of the event to create a chronology.

[33]This structure is a representation of how causal psychology supplements the available data (the fact pattern).

[34] This structure consists in the following three moral primitives: "(i) an effect that consists of the death of a person is bad, (ii) an effect that consists of the negation of a bad effect is good, and (iii) an effect that consists of the negation of a good effect is bad. (2011, 172)

[35] This structure is a representation of the intentions of the AGENT of the event.

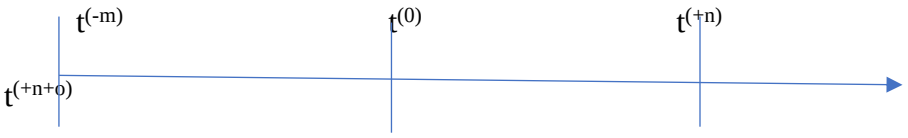
(a) Stimulus, Fact-Pattern, Input

The manufacturer of a self-driving car has to engineer different decision rules into the car about what it should do in different scenarios. The programmer is asked to imagine the following situation and then decide how to program the car's response.

A driverless car is driving down a country road that has no shoulders, with the ocean off of one side, and sea cliffs off the other. The car turns a blind corner and within 20 feet sees 5 dogs in its lane. Presumably the dogs are crossing the road to get to the beach. The car knows it cannot stop fast enough to avoid the dogs. Fortunately, it can swerve into the other lane. Unfortunately, there is 1 adult human pedestrian in that lane.

What should the programmer of the driverless car have the car do?

(b) Temporal structure



The car's noticing The car's changing The car's preventing

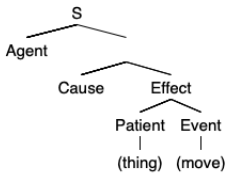
The car's

the dogs lane the death of 5 dogs

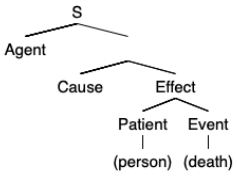
killing the man

(c) Causal Structure

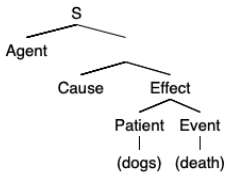
(i) Semantic structure of "the car's changing lane"



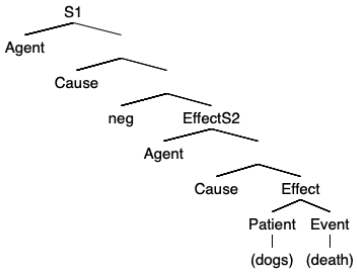
(ii) Sematic structure of "the car's killing the man."



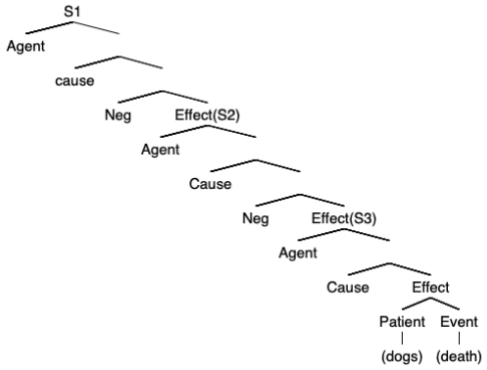
(iii) Semantic structure of "the car's killing of the dogs"



(iv) Semantic structure of "the prevention of the killing of the dogs."



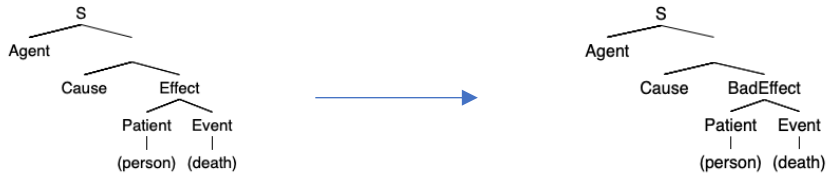
(v) Semantic structure of "letting the dogs die"


S1
Agent
cause
Neg
Effect(S2)
Agent
Cause
Neg
Effect(S3)
Agent
Cause
Effect
Patient
Event
(dogs)
(death)

Note here that out of the resources (i-iv) above, a causal change can be generated for "The car's changing lane causes the car's killing the man and causes the prevention of the killing of them dogs."

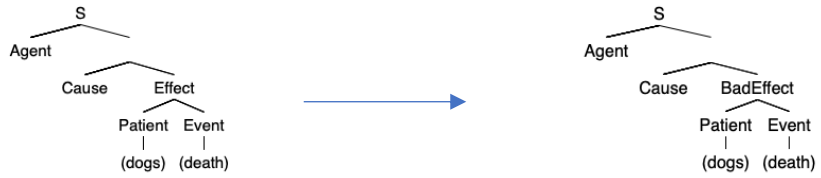
(d) Moral Structure

Moral transformation of (ii) "the car's killing of the man".



This transformation is made possible by Mikhail's first postulate of the moral calculus of risk, "[Effect [(Person, Death)] → [BAD EFFECT]" (Mikhail 2011, 137).

Moral transformation of (iii) "the car's killing of the dogs".



This transformation is actually not possible with the resources Mikhail provides. He postulates only "three primary bad effects" in his "simple model." The first is already listed above. Postulate 2 specifies "bodily harm to a person" as a bad effect. And postulate 3 specifies the "destruction of a valuable thing" as a bad effect. It may be

possible to restate these postulates so that they are capable of handling all moral patients. If we substate "moral patient" for "person", we could restate the first postulate

[EFFECT [(Moral Patient, Death)] → [BAD EFFECT]

We can further do away with postulates 2 and 3 if we, consistent (Schein and Gray, Chelsea Schein1 and Kurt Gray 2017)'s view that there is a spectrum of harms with death being at the extreme high end of the harm scale[36], restate the postulate

[EFFECT [(Moral Patient, Harm)] → [BAD EFFECT]

This would also allow us to reduce Mikahil's postulates 6-7 which specify that the death of a person is morally worse than the harming of a person which doesn't result in death. And that the harming of a person is morally worse than the destruction of a valuable thing (Mikhail, 2011, 138). Instead, I purpose the following alternative postulate

[EFFECT[(Moral Patient$_1$, Harm$_1$)] $>_m$[37] [EFFECT[(Moral Patient$_2$, Harm$_2$)] if

Moral Patient$_1$ and Harm$_1$ $\leq_m$ Moral Patient$_2$ and Harm$_2$

[36] This however would involve a strong commitment to the idea that the minds represent all harms as commensurable with each other. That is, x amount of battery can $=_m$ one death, etc.

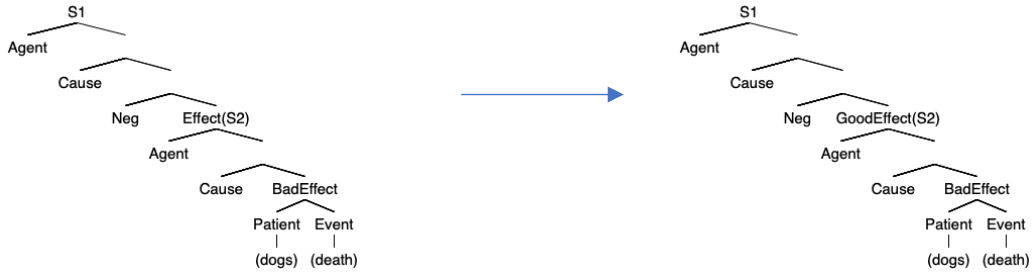
[37] Following Mikhail, the subscript m is intended to make clear that these are normative, rather than mathematical concepts. So, they can be read as "morally greater than".

Discursively this would read, an effect that harms a moral patient is morally worse than an effect that harms another moral patient, if the first moral patient's moral status is greater than or equal to the second moral patient's moral status and the harm done to the first is greater than or equal to the harm done to the second. There are obvious complications here. What if the moral status of the first moral patient is less than that of the second, but the harm to the first is greater than the harm to the second. What do we do with, for example, killing a dog versus punching a man? We might have to enrich this postulate to

$[\text{EFFECT}[(\text{Moral Patient}_1, \text{Harm}_1)] >_m [\text{EFFECT}[(\text{Moral Patient}_2, \text{Harm}_2)]$ iff the product of $(\text{Moral Patient}_1, \text{Harm}_1) > (\text{Moral Patient}_2, \text{Harm}_2)$

However, it strikes me that this stronger presentation of the postulate, particularly the part that says "the product of", requires experimental support. It could be the case that moral status is weighted more heavily than harms here, or vice versa. If so, a simple product of the two would not be supported. This also generates a potentially deep incommensurability problem. Namely, how would we obtain the product of a partial moral status score and a harm score without making the additional assumption that they're represented in the same format: (that for example, that both moral patient status and harms can be represented as continuum from 0-1; 0 = no moral status/no harm, 1= full moral status/death). We will return to these issues with fresh eyes in Chapter Four.

Moral transformation of “the prevention of the killing of the dogs”



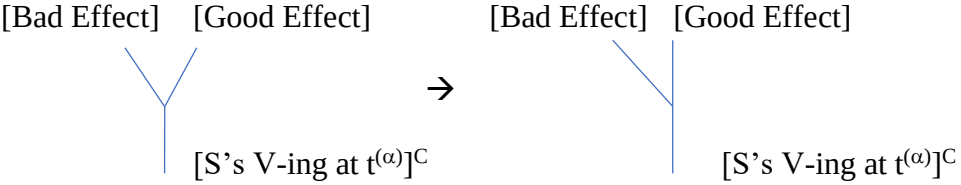
This is generated with the resources already given, plus Mikhail’s 4th postulate:

[EFFECT[neg[BAD EFFECT]]] → [GOOD EFFECT]

“an effect that consists of the negation of a bad effect is a good effect, and may be rewritten as such” (Mikhail, 2011, 138).

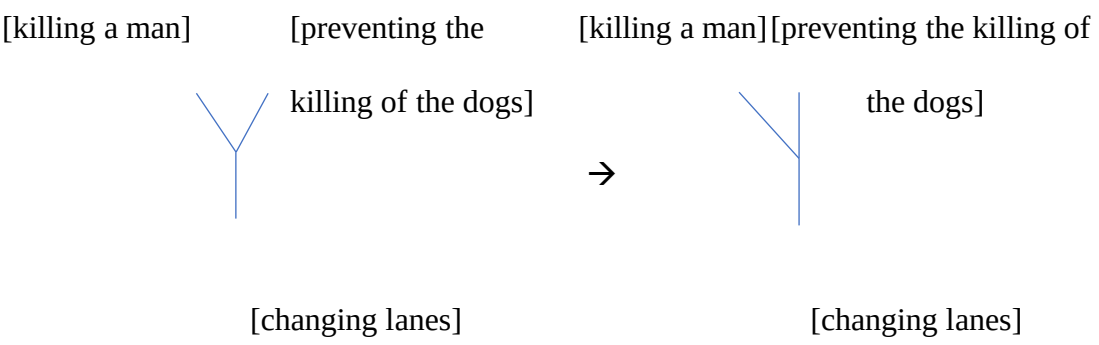
(e) Intentional Structure

Computing intentional structure of an act with good and bad effects (recreated from Mikhail, 2011, 170)



The intentional structure rewrite rules allow us to identify the intentional goals (ends), foreseen but not intended side effects, and means. The figure above on the left is used to

illustrate that the good effect is the intended outcome of a subjects (S) doing something (V-ing) at some time ($t^{(\alpha)}$) in some circumstances (^{C}). The bad effect here is side effect of [S's V-ing at $t^{(\alpha)}$]^{C}. So,



Mikhail's system provides the resources to distinguish between knowingly causing some effect, or K-Generation, and intentionally causing an effect, or I-Generation (Ibid, 130). I won't provide his full formal account of these processes here, but more on this shortly.

(f) Deontic Structure

Partial derivation of representation of homicide in 5 Dogs Trolly Problem, if the programmer has the car change lanes[38]

1. [The car's changing lane at $t^{(0)}$] Given

[38] This derivation makes use of the resources outlined in (Mikhail 2011, 132-136). I have not detailed them all here, because I fear they would be distracting from our central concern. There is some analogy to TDM reference to "norms". A description of moral norms, as opposed to non-moral norms (Ibid, 104), in terms of operative principles of the mind is after all Mikhail's primary focus. I have attempted to include enough so that one can get a sense of the cognitive model, but no more than is needed for that and an understanding of how questions of moral status fit in to the picture.

2. [The car's changing lane at $t^{(0)}$] ⊃ [The car's killing the man at $t^{(n)}$]

   Given

3. [The car's killing the man at $t^{(n)}$] 1, 2; Modus Ponens

4. [The man has not expressly consented to be killed at $t^{(0)}$]

   Given

5. [The man has not implicitly consented to be killed at $t^{(0)}$]

   Given, Abductive

6. [[The car's changing lane at $t^{(0)}$] ⊃ [The car's killing the man at $t^{(n)}$]] ⊃

   Self Perseveration[39]

   [the man would not consent to be killed at $t^{(0)}$ if asked]

7. [the man would not consent to be killed at $t^{(0)}$ if asked] 2, 6; MP

8. [The car's killing the man without his express, implied,

   3, 4, 5, 7

   or hypothetical consent at $t^{(0)}$]

9. [The car's killing the man without his express, implied, or

   Definition of homicide[40]

   hypothetical consent at $t^{(0)}$] ⊃ [The car's committing homicide at $t^{(0)}$]

[39] See (ibid, 137)
[40] See (ibid, 134)

10. [The car's committing homicide at $t^{(0)}$] ` 8, 9; MP

3.2: Takeaways from applying and modifying Mikhail's cognitive model

You might be wondering at this point how we produce a judgment of whether its permissible to program the car to switch lanes, killing the one man and saving the five dogs, from all this formalism. Is committing homicide worth the lives of 5 dogs? We have a derivation that says its homicide to cause the man to die. But that's certainly not enough.

Empirically, according to my survey, of the 279 who responded to the question, only about 21% of people said it was permissible to have the programmer have the car switch lanes, saving the dogs and killing the man (see Chapter Three). So, for a large majority allowing the killing of 5 dogs is morally better than causing the killing of 1 man in order to save the dogs.

Walking through the particulars of Mikhail's model first has provided us with an opportunity to see an alternative to Gray et al's TDM. But it also neatly illustrates where questions of moral status intersect with elements of a general theory of moral grammar. At least by the time the conversion rules that generate the causal structure are applied, our model requires the conceptual resources of an agent and patient.

The transformation rules that generate moral structure allow for the labeling of bad effects impinging on a moral patient. It involves computations that must be able to

reckon the differing value of moral patients. Mikhail is able to abstract away from some of these difficulties by postulating that "the life of one person has the same moral worth as that of another" (2011, 139) and by only using a generic man[41] in his experimental stimuli. This simplification was motivated for Mikhail but it's our goal here to reintroduce this complication to his work. That is, our goal is to test to see how people judge the relative moral value of different entities.

The 11th postulate of the moral calculus of risk, restated below, allows the cognitive model to produce judgments consistent with the idea that the death of one person is less morally bad than the death of 1+n persons.

$\forall(x, y)\ [[x>y] \equiv [(x \text{ Persons, Death})] <_m [(y \text{ Persons, Death})]]$

As we have elsewhere, this must be modified, by substituting "moral patients" with "Persons", in order to make it sensitive to judgments about entitles with differing moral statuses. And as elsewhere we can substitute "Harm" for "Death" to yield a model with still fewer postulates and more general expressive power.

$\forall(x, y)\ [[x>y] \equiv [(x \text{ Moral Patient(s), Harm})] <_m [(y \text{ Moral Patient(s), Harm})]]$

Mikhail does caution us about a move like this,

[41] The one exception is that he does report on one trolly problem in which valuable equipment plays the role of the patient (ibid, 106).

> Even if one assumes that the physical security of one person has the same moral worth as that of another, it does not follow that bodily harm to five persons is morally worse than bodily harm to one person; to reach this conclusion, both the type and the extent of the harm must be held constant.
>
> (Mikhail, 2011, 139)

So, for him, harms should be represented as a type with a magnitude. The claim here is that different types of hard are not commensurable with each other – that causing the death of an entity is categorically different than causing it other kinds of survivable damage. This is certainly a different claim from the TDM, which views all harms as commensurable. Happily, these are empirical questions.

Mikhail's proposed deontic rules (presumptive prohibitions) also incorporate elements of both moral agent and patient status. His definition of Homicide, for example, involves killing of a person.

Definition of Homicide from (Mikhail 2011, 134)

[S commits homicide at $t^{(\alpha)}]^{C} =_{Df}$ [S's V-ing [EFFECT (Person, Death)] at $t^{(\alpha)}]^{C}$

(Ibid, 134)

As we've seen, in order to be sensitive to the psychology of moral status, this can be altered to the following

Definition of morally salient killing

[S commits morally salient killing at $t^{(\alpha)}]^C =_{Df}$ [S's V-ing [EFFECT [(Moral Patient, Death[42])] at $t^{(\alpha)}]^C$

Note that this would slightly alter the partial derivation offered in (f). From this definition plus the distinction between K-Generation and I-Generation, we can generate representations of purposeful and knowing morally salient killing as alternatives to Mikhail's purposeful and knowing homicide.

Representation of Purposeful Morally Salient Killing

[S's V-ing at $t^{(\alpha)}]^C$ I-generates [S's committing morally salient killing at $t^{(\beta)}$}]

Representation of Knowing Morally Salient Killing

[S's V-ing at $t^{(\alpha)}]^C$ K-generates [S's committing morally salient killing at $t^{(\beta)}$}]

The intentional structure in (e) is a fine-grained representation of the agent's intentions. It allows us to clearly distinguish between knowingly causing some bad effect and intentionally causing some bad effect: (i.e., to specify means, ends, and side effects.) Deontic rules operate over that structure to apply either K-Generation or I-Generation procedures. Presumably this puts strong limits on what kinds of entities can count as full

[42] For purposes of consistency this should say "Harm" and have some subscript indicating the magnetite of the harm. But again that assumes all harms to be on one commensurable scale.

moral agents. They must be entities perceived to be capable of I and K-Generation: i.e., perceived capable of making the distinction between foreseen but not intended side effects, and intended main effects.

Because this model is a functional hypothesis about operative mental mechanisms,

this overview also raises several interesting and tractable empirical questions. For example, what kind of entities can count as moral agents/patients? Are the qualifying criteria mind perception, as on the TDM model? Does moral agent/patient status come in degrees rather than kinds? i.e., is it best represented with a scale variable, or with a nominal one. And, does our moral grammar make a distinction between "persons" as entities with full moral status and treat them categorically differently than entities with partial moral status? If not, then my alternative formulations of Mikhail's postulates have some warrant.

However, if we return to the distinction between the FMN and FMB, it seems that Mikhail's model does require moral status to play a functional role in aspects of the grammar that should be FMN: the moral structure and deontic rules. His model requires agents be capable of K and I-generation. And it requires a way to weigh different harms to different moral patients against each other. This doesn't necessarily exclude the constructivist account of moral status, however. We still can't rule out that moral agent/patient status is just agent/patient status in a moral context, as the TDM and (Strickland, Fisher and Knobe 2012) suggest. To do that we have to show that…

> Moral agent and moral patient are real psychological categories that are related to, but more restricted than, the more familiar grammatical categories of agent and patient, and that the former pair of categories can be given an adequate computational analysis that renders them distinct from one another and specifies their standard range of application.
>
> (Mikhail 2011, 302)

I had hoped to be able to do just that with this dissertation, however as we will see in the next two chapters my experimental findings raised more problems for these models than they solved.

CHAPTER 3

MORAL STATUS EXPERIMENTS

1. The road ahead

In Chapter One, we were introduced to a wide variety of examples of moral status as a phenomenon in the world. We drew a distinction between moral agents and moral patients. And we saw some evidence that the mind might use those concepts in two different parts of the moral faculty: the mental process(es) that generate moral evaluations (permissible, impermissible, obligatory) of an agent's action in some set of circumstances; and the mental process(es) that generate judgements of appropriate retribution for doing that act in those circumstances.

In Chapter Two we were introduced to two psychological accounts of moral status. The constructionist theory of dyadic morality, who's chief proponent is the psychologist Kurt Gray. And the theory of universal moral grammar, headed up by the legal scholar John Mikhail. We saw that these accounts differ in their claims about the representational format of moral status. On the TDM moral status is constructed out of core knowledge resources. It is a domain general cognitive template used in social cognition more generally. As such it has fuzzy boundaries, i.e., you can be more or less of a moral patient/agent. We also saw that the implementation of the theory of moral grammar suggests that moral agent/patient status is distinct from the thematic roles AGENT/PATIENT, contra the claims of (Strickland, Fisher and Knobe 2012).

Additionally, on the TDM we've begun to see that there's no clear distinction between the rightness and wrongness of an action, and the punishments thought to be

justifiable. Judgements of justified punishment, like judgments of rightness or wrongness, are predicated on the same proposed causal structure: on the perceived minds of agent and patient, local moral norms, and moral emotions (see CH 2, section 2.2).

In this chapter I report on my attempts to put these accounts to the test. I ran two general kinds of experiments. The first kind are modifications of the trolly problem which ask subjects to trade 5, 10, or 20 animal lives {Dogs, Chimps} against one human pedestrian life. The second kind of experiments elicit judgments about interpersonal killings involving cryptomind. This involved two different data collection efforts. The first collection effort includes the driverless car trolly problems (section 2), the murdering and murderable cryptominds vignettes (section 3), the battle of the sexes vignettes (section 4), and the battle of generations vignettes (section 5). The second collection effort includes the attempt to replicate the findings of some of section 2 using different measures for mind perception (section 6). As well as an exploratory study testing visual perception ques to moral status (section 7). I was aided significantly by Michal Barlev, Ph.D., in implementing these experiments on Qualtrics and collecting the data. He also provided invaluable feedback on the language of the cryptominds vignettes, as well as some primary help with data analysis thereof.

1.1. Data Preparation, first collection effort

I removed anyone

1. who spent less than 10 minutes on the survey (n = 292)
2. anyone remaining who didn't answer either attention check 1 or 2 (n= 14)

3. anyone remaining who's answer to attention check 1 included webpage addresses or restated the attention check question (n=8).
4. anyone remaining who answered attention check 2 with gobbledygook like "der, eulb" (n=17)
5. anyone remaining who answered incoherently to both questions (n = 1)
6. anyone remaining who answered "yellow" for attention check 2 (n=93), and answered attention check 1 with one word (n=9), blank (n= 8) or long irrelevant sentences about grammar (n=2). Excluding n=19
7. anyone remaining who didn't answer attention check one with sentences about food (n= 11)
8. anyone remaining who completed less than 60% of the survey (n=21)

Strictly speaking this leaves in a sizable number of respondents who failed to answer attention check 1 with grammatical sentences as required. Or who answered attention check 2 incorrectly (i.e., with "red, blue" or any answer with "yellow" included) (n=92) or not at all (n= 46). I did not exclude these responses.

That means I excluded through cleaning 383 respondents. That left the total n of the study at 1339. Though of those only 1315 answered all the demographic questions.

1.2. First collection general demographics

Participants were n = 1315 Amazon Mechanical Turk (MTurk) workers (mean age = 40, SD = 13.3, range 18-81, 61% female). The vast majority of respondents were white (76.7%) and educated (88% had some college or better under their belt). Most report being middle class (53%). A plurality were Democrats (41.4%). Eight percent of the sample (n= 116) self-identified as vegetarian, which is slightly higher than the five percent average for Americans (Hrynowski 2019). Four percent identified as vegan (n= 52). But n= 35 of the vegans also answered that they were vegetarians. So vegan and vegetarian together (n= 133) or 10.11% of the sample. This is slightly less than the 13.4% that said it was morally wrong to eat meat.

This makes them a fairly WEIRD population (Joseph Henrich 2010). While I do not address these issues, this experimental paradigm could certainly benefit from replication attempts in non-WEIRD populations. Some more detailed demographics are provided in the appendix.

1.3. Data Preparation, second collection effort

I removed anyone,

1. who completed the survey in less than 200 seconds (n= 131)
2. then those who left attention checks 1 and 2 blank (n= 40)
3. then those who's answer to attention check 2 included "yellow" and who answered attention check 1 incorrectly or not at all (n= 2)
4. then those who's answer to check 2 was blank, and who's answer to check 1 was wrong (n= 1)

Strictly speaking this leaves in the study people who answered check 2 incorrectly. After cleaning, I was left with n= 731

1.4. Second collection general demographics

Participants were n = 731 Amazon Mechanical Turk (MTurk) workers (mean age = 38.8, SD = 12.8, range 18-87, 57% female).

2. Driverless Car Trolly Problem

This experiment was exploratory and represent a first attempt at testing Mikhail's proposed deontic postulates, especially those which allow for the comparing of harms done to one moral patient against harms done to another. While Mikhail avoids these complications in his work, the TDM attempts to tackle them head-on. According to the TDM, a kind of hierarchy of moral consideration can be generated from a graph about mind perception. If some entity is perceived to have more of a capacity for agency than another, then it will be perceived as more of a moral agent. And if some entity is perceived to have more of a capacity for experience than another, then it will be perceived as more of a moral patient. In fact, Gray, Gray, and Wegner first generated this graph in 2007, reproduced here with permission from AAAS.

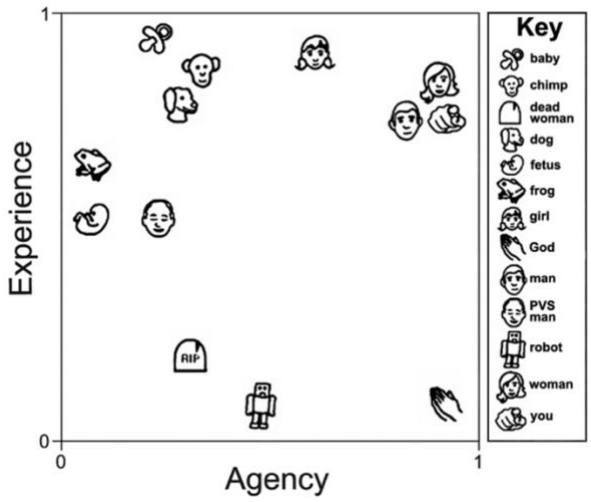


Fig. 6. Gray Gray and Wgner's Mind Perception Graph

On the TDM, you can replace the Y axis label with "Moral Patient" and the X axis label with "Moral Agent". The picture can be enriched and complicate with TDM consistent contributions like the "two sources" of moral patient status (Sytsma, Machery 2012) in which moral patient status involves both perceived Agency and Experience (Ch 2 Section 2.2). Either way, TDM implies a few clear testable predictions.

i. Any entity perceived as having greater capacity for Agency then another should be perceived as more of a moral agent: i.e., should be viewed as more culpable for unjustified harmful actions.

ii. Any entity perceived as having greater capacity for Experience then another should be perceived as more of a moral patient: i.e., should be viewed as something more worthy of protection against unjustified interpersonal harms.

iii. (Sytsma, Machery 2012) Alternative to ii. Any entity perceived as having greater aggregate capacity for Agency and Experience than another should be perceived as more of a moral patient.

What we don't have from TDM or its proponents is an explicit way their model reconciles harms to different numbers of entities. Remember that we proposed to alter Mikhail's formalism of this issue as

$$\forall(x, y)\ [[x>y] \equiv [(x \text{ Moral Patients, Harm})] <_m [(y \text{ Moral Patients, Harm})]]$$

In a way this represents an exceedingly flexible empirical claim. It only suggests that all else equal, a harm to more moral patients is morally worse than a harm to fewer. So, we don't know if there is a marginal losses per moral patient added. What's worse, there is some reason to suspect that our moral sense is significantly, if not totally, innumerate. For example, Paul Bloom has persuasively argued that at least our empathy is innumerate (Bloom 2016). It is true that because empathy is a moral emotion, we might be able to get away with putting that problem into another box, so to speak. However, that seems to be rather hasty. For example, (Dwyer, Huebner and Hauser, THe Linguistic Analogy: Motivations, Results, and Spectulations 2010) convincingly argue that that our moral psychology is sensitive to Pareto improvement. If true that would put at least one strict condition on the moral calculus we're trying to accomplish with the above formalism.

Faced with this problem, we can add in another complication: does moral status, like mind perception, come in commensurable units or incommensurable kinds? That is, perhaps entities with partial moral status are treated categorically differently by the moral sense than entities with full moral status. Which would mean our moral sense would be stymied by questions that require tradeoffs between those two types; between pigs (for most people perceived as cryptominds with maybe partial moral status) and humans (a priori possessors of full moral status), for example.

Before we proceed there is some consistent amount of background criticism about whether using trolly problems in moral psychology are useful or not. For example, see (Dahl and Oftedal 2018) who argue that "general willingness to trade lives in the trolly context may be an artifact that is due to its unrealistic setting". I however, agree with (Cushman and Greene 2011), that "moral dilemmas illuminate cognitive structure". And that unusual contexts in moral psychology, just like unusual sentences in linguistics, are particularly useful tools when attempting to characterize the relevant psychology. However, it is reasonable to doubt if some of the more implausible stipulations of some iterations of the trolly problem are believed by subjects; for example, we can doubt if people truly think that a large man (or man with a large backpack) could physically slow down an out of control trolly sufficient to save the lives of the 5 people down track (as in the so-called footbridge problem). Here I made an effort to overcome some of those objections by asking subjects about how a driverless car should be programmed to make decisions. However, when the number of entities in one lane reaches 20, we are again straining physical credulity a bit.

2.1 Methods

I first wanted to establish that my stimulus would be judged similar to how trolly problems are judged. I expected that they would be, because they involve the same core structural representations as the trolly problem does; (i.e., a tradeoff between allowing the killing 5 entities, versus changing lanes thereby intentionally causing the death of 1 entity, and saving 5 entities). And a major feature of this account is that structurally similar events will be judged similarly. In fact, it is argued that this methodology (giving a formal account of the mental recreations of actions) is an important methodological advance for the field of social psychology (Levine, Leslie and Mikhail 2018).

In order to do that I have three "baseline" vignettes, which involve tradeoffs between the same type of entities {5 versus 1 pedestrian; 5 versus 1 dogs; 5 versus 1 chimps}. Subjects were also presented with exploratory vignettes that involve tradeoffs between 1 human life and x {5, 10, or 20} non-human animal life {dogs, chimpanzees}. This yields a total of 9 vignette conditions (3 baseline, 6 exploratory).

Each subject saw all three baseline conditions. However, each subject saw only two exploratory conditions, one per entity {dog, chimp}. The order of presentation in the baseline and exploratory conditions was randomized, respectively.

2.1.1 Participants

n= 1315 of subjects pulled form the first data collection effort (section 1.2)

2.1.2 The experimental stimuli:

2.1.2.1 The generic driverless car vignette

The manufacturer of a self-driving car has to engineer different decision rules into the car about what it should do in different scenarios. The programmer is asked to imagine the following situation and then decide how to program the car's response.

A driverless car is driving down a country road that has no shoulders, with the ocean off of one side, and sea cliffs off the other. The car turns a blind corner and within 20 feet sees [x number of entities] in its lane. Presumably the [entities] are crossing the road to get to the beach. The car knows it cannot stop fast enough to avoid the [entities]. Fortunately it can swerve into the other lane. Unfortunately there is [1 entity] in that lane.

What should the programmer of the driverless car have the car do?

- Stay in its lane, killing [x entities]
- Swerve out of its lane, killing 1 human pedestrian, and saving the [x entities]

2.1.2.2 Mind Perception Measures,

All subjects (n = 1315) were also asked to rate these entities {A DOG, A CHIMPANZEE, A PEDESTRIAN} from 0 (Not at all capable of Agency/Experience) to 100 (Extremely capable of Agency/Experience) on the Agency and Experience dimensions of mind. The stimulus for these questions was

2.1.2.2.1 Stimuli for Experience Measure

Please consider the following:

Entities or beings vary in how much they are capable of Experience.

Entities or beings with Experience are capable of things like hunger, fear, pain, pleasure, rage, desire, personality, consciousness, pride, embarrassment, and joy.

Please rate the following entity based on where you think it falls along this dimension.

[Entity label = {A CHIMPANZEE, A DOG, A PEDESTRIAN}]

2.1.2.2.2 Stimuli for Agency Measure

Please consider the following:

Entities or beings vary in how much they are capable of Agency.

Entities or beings with Agency are capable of things like self control, planning, memory, emotional recognition, communication, and thought.

Please rate the following entity based on where you think it falls along this dimension.

[Entity label = {A CHIMPANZEE, A DOG, A PEDESTRIAN}]

2.2 Descriptive Statistics

2.2.1 Mind Perception Descriptives

Fig. 7: Driverless Car Cryptomind Ratings of Agency and Experience

Figure 6 here show's the mean and standard deviations for Agency and Experience for each entity.

Histograms of these distributions show that the mind perception ratings for Dogs and Chimps are, for the most part, normally distributed. The ratings for the Pedestrian are not. They are heavily skewed right, i.e., the vast majority of the scores bunch up towards the extreme high end of the scale (mdn Agency = 100; mdn Experience = 100).

An independent t test did indicate that there are significant difference between Agency for the chimps (m= 62.34, SD 27.594) and Agency for dogs (m= 54.05, SD 28.217), t(2628)= -8.296, p> .001, Cohen's d= .297, a small effect. A t test also indicated a significant difference between Experience scores for chimps (m= 70.44, SD 27.078) and dogs (m= 66.06, SD 28.07), t(2628) = -4.072 , p> .001, Cohen's d= .159, a small effect.

2.2.2 Moral judgment descriptives

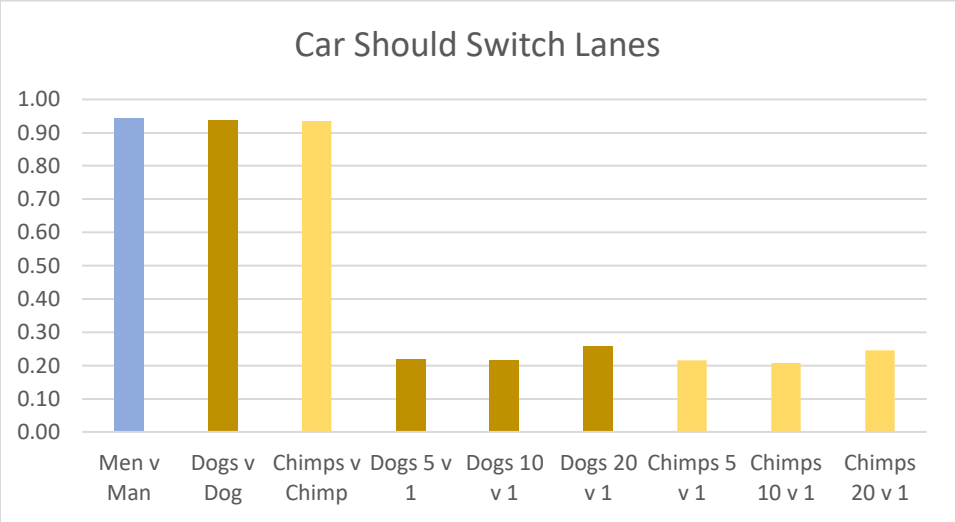


Fig. 8: Driverless Car Moral Judgments

Figure 7 shows the percent of respondents who said the car should switch lanes, avoiding the death of the 5, 10, or 20 entities, causing the death of one pedestrian.

An independent t test found no significant differences between the baseline conditions depending on the entities involved.

An independent t test found no significant difference between answers in the (5 v 1) and (20 v 1) change lane judgments for either entity {dogs, chimps}.

An Independent t test found no significant difference between change lane judgments for dogs and chimps whatever number of entities (5, 10, 20}.

2.3 Inferential Statistics

Because we are searching for correlations between nominal (switch lane or not) and scale (0-100 Agency & Experience) variables, Pearson's correlation test is inappropriate. Instead, I used Spearman's correlation test, which is more appropriate for these circumstances. Please note that Spearman's correlation (rho) is itself a measure of effect size, 1 representing a perfect positive relationship, 0 no relationship at all, and -1 a perfect negative relationship. Using this test I found several small but significant effects (detailed inferential statistics are available in the appendix, section 1).

I therefore attempted a binary logistic regression for answers in each case, using the change lane judgment as the dependent variable, and the mind perception measures of the relevant nonhuman animal {Dog, Chimp} as independent variables, excluding variables from the model with $p < .051$. Mind perception was not a significant predictor in the 5 Dogs and 10 Dogs cases. However, in the 20 dogs to 1 human condition, there was a significant effect between Agency rating and change lane judgments, ($\chi^2(2) = 10.97$, $p = .001$). Cox & Snell $R^2 = .024$ can be interpreted as Agency ratings predicting 2.4% of the variance in change lane answers for this condition.

The results of a binary logistic regression indicated that in the 10 Chimps v 1 Pedestrian condition, change lane judgments were significantly predicted by Experience

($\chi^2(2) = 6.404$, p= .011). Cox & Snell $R^2 = .014$ can be interpreted as Experience ratings predicting 1.4% of the variance in change lane answers.

Likewise, the results of binary logistic regression indicated that in the 20 Chimps v 1 Pedestrian condition, change lane judgments were significantly predicted by Agency ($\chi^2(2) = 4.048$, p= .044). Cox & Snell $R^2 = .009$ can be interpreted as Agency ratings predicting .9% of the variance in change lane answers.

2.4 Discussion

Because, for dog and chimpanzee, there is a reasonable amount of variation (large standard deviations) in Agency and Experience ratings, we should be able to detect a strong relationship between change lane judgments and those ratings, if one exists. However, our regression analysis never discovered a significant effect that explained more than 3% of the variance in change lane judgments (section 2.3).

While the difference in mind perception between dogs and chimpanzees is not extreme, if the account of moral status offered by the TDM we would expect to see higher change lane judgments for chimpanzees than for dogs, since their Agency and Experience scores were significantly different (section 2.2.1).

Additionally, the dramatic difference between subjects willing to trade 1 human life for 5 human lives, and their overall unwillingness to trade 5 nonhuman lives for 1 human life (section 2.2.2), causes some trouble for an account of moral patient status as mind perception. If moral consideration isn't proportional to perceived capacity for

agency/experience, then proponents of TDM seem to have missed something deeply important about moral status.

Additionally, there is hardly any change in mean change lane judgments irrespective of the number of nonhuman animals on the chopping block (section 2.2.2). So, increasing the number of nonhuman animals being traded against 1 human life by a factor of 4 had no significant effect on change lane judgments.

The moral discounting of nonhuman animals in these cases can seemingly not be explained by reference to the difference in mind perception. Yet the conditions which involve trading dogs for dogs, and chimps for chimps, are judged almost identically to the baseline 5 v 1 human condition. What is the cause of this moral discounting of the lives of nonhuman animals, if not mind perception? Perhaps this is evidential support for the hypothesis that partial and full moral statuses are not commensurable. i.e., that most humans just morally value another human perhaps infinitely more than any other entity whatever the number.

3 Murdering and Murderable Minds

My second type of experiment deployed the use of the English verb murder. So far, we've been referring to "thematic roles" and linguists in an unspecified way. But thematic argument structure is central to almost every question in linguists: how are verbs learned (Pinker 1989), what a verbs representational format is and exactly what information is stored in them (Beth Levin 2005), how did the faculty of language evolve (Jackendoff and Pinker, The nature of the language faculty and its implications for

evolution of language 2005), how does change in verb meanings happen over time (Van Gelderen 2018), etc. So generic appeals to thematic structure, as in (Strickland, Fisher and Knobe 2012), might obfuscate the diversity of opinion about just what argument structure is, psychologically. This just won't do. As we've seen, answers to what thematic argument structure is in the mind are answers about the basic "core" components of our thinking process. It's the conceptual structure that is the scaffolding, the very stuff, of thought (Pinker 2007).

A first approximation of the literature on argument structure is that it is "the study of the possible syntactic expressions of the arguments of a verb". Verbs are exceedingly choosy about their phrase mates, and linguists want to understand why. The classical example, introduced to the literature by Fillmore (Fillmore 1970), altered somewhat below.

- The girl broke the window with a ball
- The girl hit the window with a ball
- The window broke
- *The window hit

Linguists are interested in why the window broke is acceptable, but the window hit isn't. What is it about these different verbs that makes it so they can only participate in certain structures. Theories of argument structure attempt to account for these differences in many ways: some attempt a semantic roles list (Jackendoff 1972), so that a verb is stored with information that specifies exactly what arguments it can take (break verbs, for example, would have listed "Agent, Patient" where as verbs like put would list "Agent,

Theme, Location"). Some attempt solutions using a feature decomposition where 1 or 2 features are combined to specify what arguments the verb can take (Reinhart 2002). So the role "Agent" on this account is specified by the features +c (causer) and +m (has mind) and other roles can be generated out of combinations of + or – those features (c, m). Some attempt predicate decompositions, which in contemprary generative theory is instantiated through the use of multiple verb-phrase-shells: "VPs whose heads correspond to the primitive predicates of lexical decomposition" (Levin and Hovav 2005, 69). And some argue that arugment realization can be adequatly explained by reference to the temporal and mereological properties of predicates (Van Gelderen 2018).

For our purposes specific solutions to this incredibly complex topic are not imediately important. However, some of the some of the conversation here is. For example, verbs often "lexicalize" different semantic features that are otherwise part of argument structure. That is, some basic concept will be decorated with another bit of conceptual structure, and then turned into a word. Keeping to our relevant domain, verbs of killing lexicalize either the means or instrument of killing {class members: asphyxiate, crucify, drown, electrocute, garrote, hang, knife, poison, shoot, smother, stab, strangle, suffocate}, or the purpose or manner of killing {assassinate, butcher, dispatch, eliminate, execute, immolate, kill, liquidate, massacre, murder, slaughter, slay} (Levin, English Verb Classes and Alternations 1993, 230-233). These classes of verbs act a bit differently. For example, the class members that lexicalize means can include a resultative

- He was crucified to death

- He was shot to death
- He was stabbed to death

But the class members that lexicalize manor cannot

- *He was executed to death
- *He was assassinated to death
- *He was murdered to death

I argue that the verb like murder is part of a small class of verbs that lexicalizes moral disapprobation and which require moral agents/patients as phrase mates {assault, extort, murder, rape, rob}.

- *The bear assaulted the bear
- *The bear extorted the bear
- *The bear murdered the bear
- *The bear raped the bear
- *The bear robbed the bear

An implication of this kind of analysis is that a predicate decomposition of murder is possible[43]. And further that the intension/mental representation of the verb murder involves the following conditions,

[43] There is an enormous literature on if the verb kill can be specified with a predicate decommission such that it means "cause to die." **Invalid source specified.**. Much of the concerns of this literature can be avoided if we date the event by following legal conventions and "dating an action by when it is completed" (Mikhail 2011, 125). For

Condition 0) X cause Y to die (df, kill)

Condition 1) X kills Y intentionally

Condition 2) X kills Y unjustifiably (i.e., killing is a prima facies wrong + no standard exculpating circumstances exist)

Condition 3) X's unjustifiable killing of Y is deserving of punishment

Condition 4) X must be a moral agent (partial or full)

Condition 5) Y must be a moral patient (partial or full)

Crucially, my experimental program depends on the idea that the extension of verb murder should include only circumstances that involve a moral agent/patient pair. Approval of describing an event as a murder should decrease if one or the other of the entities don't have the relevant sort of moral status.

In order to use the verb murder as a way to detect the moral agent/patient status of an entity, there are always at least 4 things to control for. First, the event must be perceived as involving a killing (Condition 0). Second, that killing should be an intended goal of the action (on Mikhail's model, murder judgments should be I-generated, not merely K-generated like manslaughter (Condition 1)). Third, the killing should be defeasibly wrong, with no commonly accepted exculpatory circumstances (Condition 2). Recall that Mikhail's first postulate stipulates the killing of a person as a bad effect -aka, a presumptive or prima facies wrong-. Mikhail has also found, that in "in several hundred

---

these purposes I follow the approach to predicate decomposition outlined by (Levin & Hovav, 2005, 68-74).

jurisdictions throughout the world, including all of the 204 member-states of the United Nations" codify murder as a prima facie wrong, defeasible by a small set of universally shared exculpatory factors (Mikhail, Is the Prohibition of Homicide Universal? 2010). None of my vignettes included those exculpating circumstances (for example, no one is killing for self-preservation). And so, condition 2 here is met.

Last, in order to use murder as a method to detect moral the agent/patient status, you have to control the moral status of one of the entities (either condition 3 or 4). That is, one of the entities has to have a known moral status. I used a generic man to serve as an a priori full moral agent and full moral patient.

I therefore asked subjects whether five sentences corresponding to these conditions were true about the vignette condition I presented (section 3.1.3).

3.1 Methods

3.1.1 Participants

Because this study was done during the same round of data collection as the prior one described, the participant information is identical. See 2.1.2.

3.1.2 Stimuli

The open [agent] and [patient] slots in the following stimuli descriptions were filled by the name of two entities from our list {Man, Android, Bear, Dog, Adult Chimp, Juvenile Chimp}. In almost all cases "a man" filled one of the roles, where one of the other names filled the opposite role. The only exceptions are the cases where "a man" occupies both

agent and patient roles (one basic, one time delay condition (TDC)), and the cases where "an android" occupies both agent and patient roles (one basic, one TDC). There were 14 basic vignette conditions (7 cryptominds x 2 thematic role). And 12 time delay conditions (TDC) (5 cryptominds x 2 thematic role). There were fewer TDCs because I did not collect them for the entities {adult chimp, baby chimp}.

3.1.2.1 Generic Murder Vignette

Imagine the following:

A [agent] is walking in the woods near the area in which [agent pronoun] lives. While walking [agent pro] sees a [patient]. The [agent] shoots/attacks the [patient], which results in the [patient]'s death."

3.1.2.2 Alterations for two more entities in the agent role

Minor variations of this were made to the text when the entities {Tree, River} were slotted into the agent role. Trees and rivers cannot go walking after all. Nor can they obviously launch an attack. In those cases, the text reads as follows,

Imagine the following:

A man is walking in the woods near the town in which he lives. While walking, he sees a large tree/river. The tree falls on/the river floods and the flood waters overtake/ the man which results in the man's death.

3.1.2.3 Time Delay Variations

We replicated these conditions with all entities (except the chimpanzees) with a minor revision to the narrative that indicates a time delay between when the [agent] first encounters the [patient], and when the [agent] kills the [patient]: (this is the difference between getting into an argument with someone and killing them, versus getting into an argument with someone, waiting a day, and then killing them: i.e., what is sometimes in the law called malice of forethought.) The text for a time delay conditions is as follows[44]:

"A [agent] is walking in the woods near the area in which [he/it] lives. While walking [he/it] sees a [patient]. The [patient] moves on and the [agent] continues walking.

The next day, the [agent] is out walking again when [he/it] sees the same [patient]. This time, the [agent] shoots/attacks the [patient], which results in the [patient]'s death."

[44] Changes from the base vignette text are underlined here for clarity, but not underlined in the text presented to subjects:

Participants were randomly assigned to see 1 condition per cryptomind (non-human) entity: for example, they would never see two vignettes to do with bears, or two to do with trees. The conditions were presented in random order. Hereafter time delay conditions are identified as TDCs.

3.1.3 Measurement Variables

3.1.3.1 Moral Judgments

After reading each vignette, respondents were asked to indicate whether the following five sentences were true or false:

1. the [agent] killed the [patient]
2. the [agent] murdered the [patient]
3. the [agent] attacked the [patient] in order to kill [agent pronoun]
4. the [agent] was wrong to attack the [patient]
5. the [agent] should be punished for attacking the [patient]

3.1.3.2 Measuring Mind Perception

The stimuli here was identical to that of 2.1.2.2 That is, after answering these true/false questions, respondents were asked to rate the agent and patient entities, that played a role in the condition they just read, along agency trait and experience trait dimensions (from 0 – Not at all capable of [Experience/Agency] to 100 – Extremely capable of [Experience/Agency]).

3.2 Results

3.2.1: Moral Judgments (True/False)

3.2.1.1: Baseline, Man v Man

Subjects approve of the vignette circumstances being described as a wrongful (95%) intentional (92%) killing (99%) deserving of punishment (97%) and appropriately described as a murder (97%), in the condition in which the agent and patient were both "a man" (n= 288).

For all other descriptive statistics concerning the moral judgments, I've generated the following coordinate graphs for easy of comparison

3.2.1.2 Kill

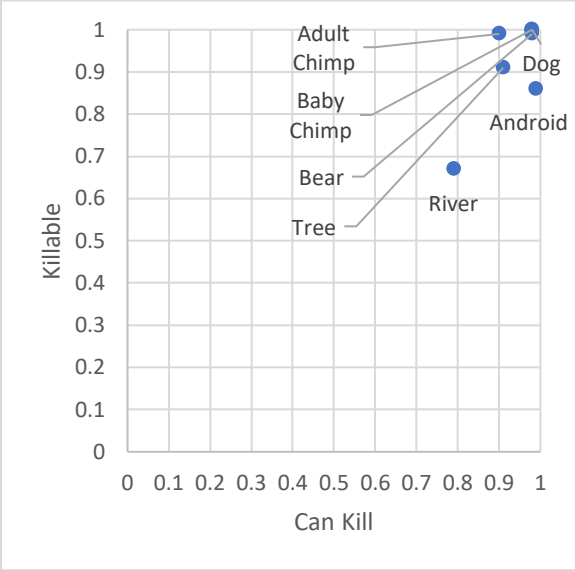


Fig. 9: Killable and Killer Cryptominds

3.2.1.3 Murder

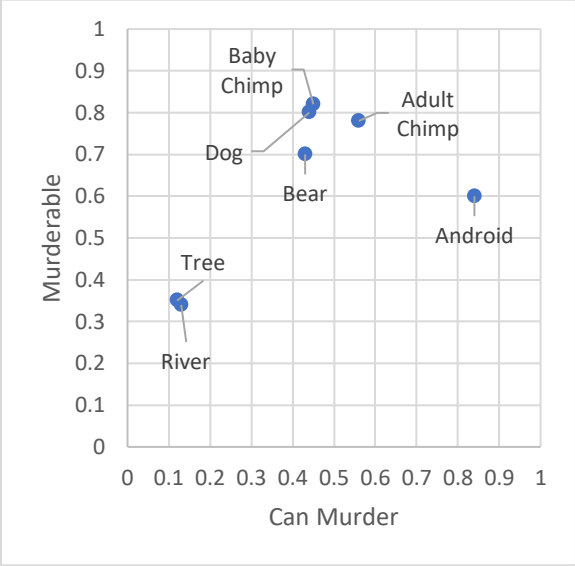


Fig. 10: Murderable and Murdering Cryptominds

3.2.4 In Order To

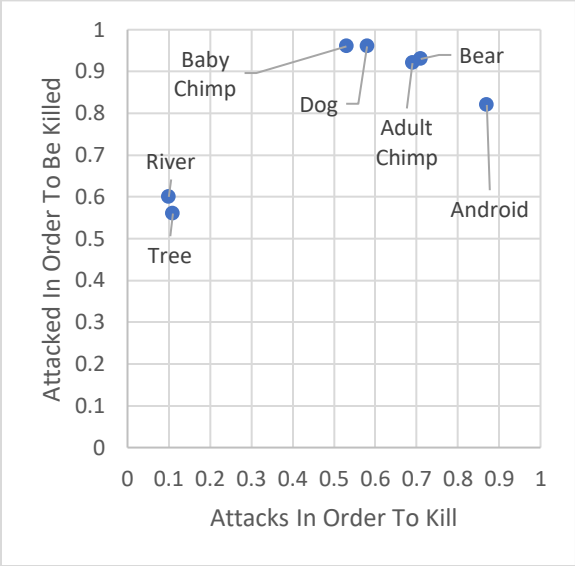


Fig. 11: Intentional and Unintentional Cryptominds

3.2.1.5 Wrong

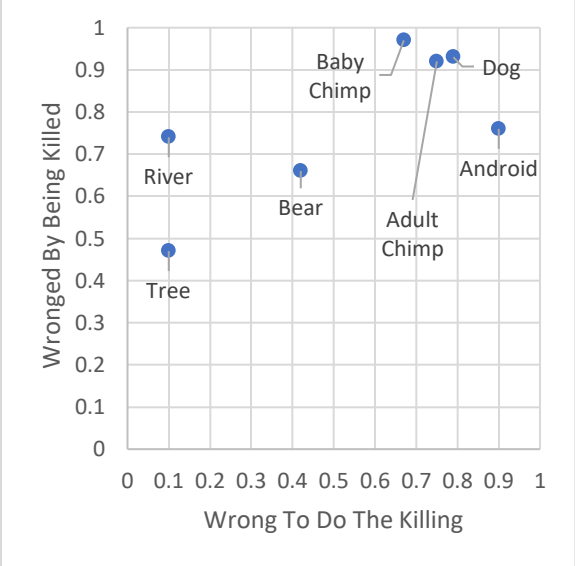


Fig. 12: Wronged and Wrongful Cryptominds

3.2.1.6 Punishable

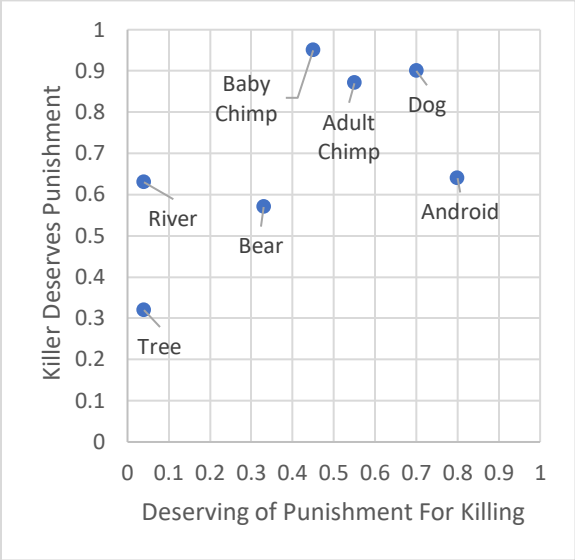


Fig. 13: Punishing and Punishable Cryptominds

3.2.2 Moral Judgments, Time Delay Condition

The above graphs were gathered from descriptive statistics taken on the moral judgments irrespective of if they were part of a TDC. I therefore assessed if the TDC made any significant difference to moral judgements by means of an independent t test. Below is a table showing effects where they were significant by giving the Cohen's Conventions effect size. The detailed means comparisons are reported in the appendix.

TDC t test Effect Size	Theta Role	Kill	Murder	In Order To	Wrong	Punish
Android	Agent		-.194			
	Patient				.259	
Bear	Agent					
	Patient			.202	.213	.183
Dog	Agent					
	Patient	-.236				
Tree	Agent					
	Patient			.243	.303	.243
River	Agent					
	Patient					.166

Table 1: T Test for effect of Time Delay Conditions on Moral Judgments

3.3 Measured Mind Perception

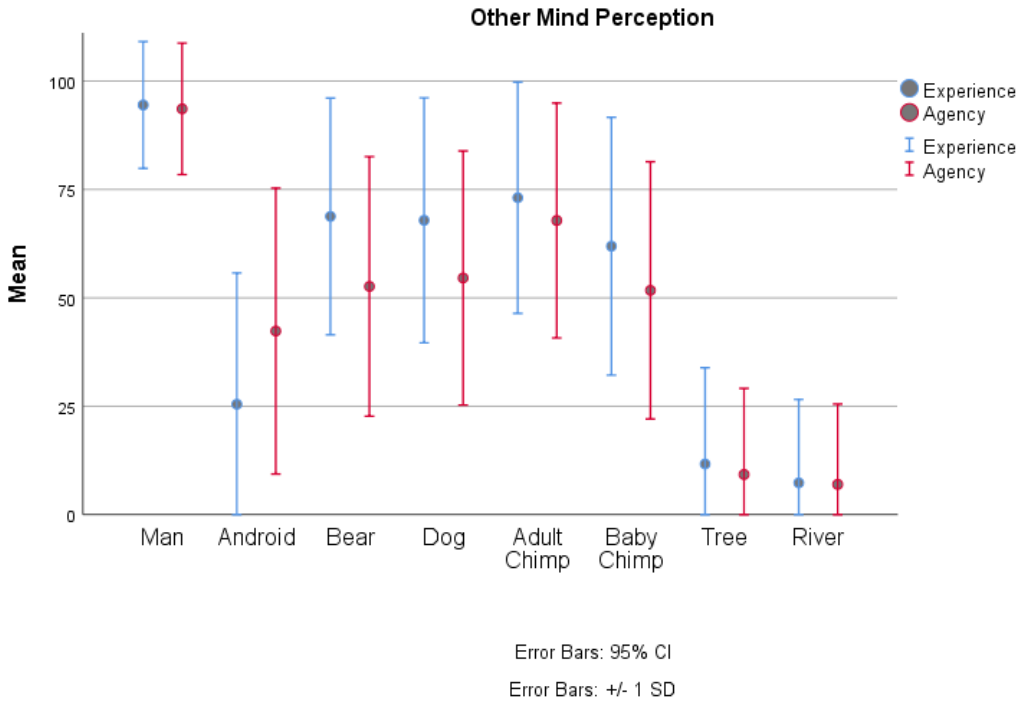


Fig. 14. Ratings of Experimental Entities Minds

Figure 10 above is a bar graph showing the mean and standard deviations for the dimensions of mind Agency and Experience, as measured. Three entities mind perceptions scores where not normally distributed: {Man, Tree, River}. Otherwise, the distributions were largely normally distributed.

3.3.1 Effect of Thematic Role on Measured Mind

I evaluated the differences in distributions of perceived mind (Agency and Experience measures) per entity per thematic role. I used two statistical tests here: an independent t test, appropriate for the entities with normally distributed Agency and Experience measures; and the Mann-Whitney U Test, appropriate for the non-normally

distributed entities. Cohen's d has been calculated to obtain the effect size of the parametric tests (J. Cohen 1988). And the Eta (= n^2= Z^2/n-1) has been calculated for the non-parametric tests. A variety of very small but significant effects were detected. Below is a table representing the findings. The detailed means comparisons are reported in the appendix, section 3.

Comparing Means	Theta Role	Agency	Experience
Android	Agent	.132	
Bear	Agent	-.320	-.175
Dog	Agent	-.147	
Adult Chimp	Agent		
Baby Chimp	Agent		
Tree	Agent	-.007	-.007
River	Agent	-.011	-.007
Man	Agent		-.008

Table 2: Thematic Role and Mind Perception

3.3.2 Effect of TDC on Mind Measures:

I evaluated the differences in distributions of perceived mind (Agency and Experience measures) per entity. I used two statistical tests here: an independent t test, appropriate for the entities with normally distributed Agency and Experience measures; and the Mann-Whitney U Test, appropriate for the non-normally distributed entities. Because only two small effects were found a table is not given. Experience was very

slightly greater in the TDC for the Android and Dog. The inferential statistics are given in the Appendix, section 4.

3.4 Correlations: Mind Perception Measures and Moral Judgments

Again, while it is typical to provide Pearson's correlations between nominal variables (like my true false moral judgments) with scale variables (like my mind perception measures), strictly speaking this is illicit, and Spearman's correlation should be preferred. Also, because the mind perception measures, Agency and Experience, are also not normal for several entities {Man, Tree, River}, a two tailed Spearman's seems to me to be preferable for all entities. Recall that Spearman's rho is itself a measure of effect size (see section 2.4).

Below are four tables summarizing the inferential statistics are given in detail in the appendix. In all cases the effects detected were small (Spearman's rho never = greater than .29 for any entity in any thematic role) but significant. The following tables list significant effects between moral judgments and the experience and agency of the patient, and the same judgments and the experience and agency of the agent. All effects were positive. I have not provided the spearman's rho in these tables because I want the reader to get the idea that there are very small, and seemingly unpatterned correlations being detected.

Correlations Experience, Patients	Kill	Murder
Android	effect	effect
Bear		effect
Dog		effect
Adult Chimp	effect	
Baby Chimp		
Tree		effect
River	effect	effect

Table 3: Perceived Experience of

Patients and Moral Judgments

Correlations Agency, Patients	Kill	Murder
Android	effect	effect
Bear		effect
Dog		effect
Adult Chimp	effect	
Baby Chimp		effect
Tree		effect
River	effect	effect

Table 4: Perceived Agency of Patients and Moral Judgments

Correlations Experience, Agents	Kill	Murder
Android		effect
Bear		
Dog		effect
Adult Chimp		
Baby Chimp		
Tree		effect
River		effect

Table 5: Perceived Experience of

Agents and Moral Judgments

Correlations Agency, Agents	Kill	Murder
Android		effect
Bear		
Dog		effect
Adult Chimp		effect
Baby Chimp		effect
Tree		effect
River		effect

Table 6: Perceived Agency of Agents and Moral Judgments.

Because there were a lot of small effects, I ran a binary logistic regression with the moral judgments as dependent variables, and the measures of mind perception for the cryptomind entity as independent. The tables below summarize the findings. The number

in the box is the percent of the answers to the judgment that can be attributed to the row. For example, the 8.5 in row {Android, Agency}, column {Murder}, in the Patients table, means that a model with the independent variable android Agency was able to account for 8.5% of the variance in responses to “The man murdered the android” = TF?. Quotations marks in boxes mean the number above is the product of a model that includes both Experience and Agency. The detailed results of the regression analysis are available in the appendix, section 6.

Patients	Mind	Kill	Murder	In Order To	Wrong	Punish
Android	Experience	5.5				
	Agency	"	8.5	3	2.7	4.8
Bear	Experience			0.8	4.1	6.4
	Agency		4.7		"	"
Dog	Experience			0.8		0.8
	Agency		1.9			
Adult Chimp	Experience					
	Agency	1.8				1.6
Baby Chimp	Experience					3.8
	Agency		1.9			
Tree	Experience					
	Agency		4.4	3.4	4.6	7.8
River	Experience		5.4	1.6		0.7
	Agency					

Table 7: Regression on DV Moral Judgments and IV Mind Perception, Patients

Agents	Mind	Kill	Murder	In Order To	Wrong	Punish
Android	Experience				1.6	1.6
	Agency		3.9	1.6		
Bear	Experience					
	Agency					
Dog	Experience					
	Agency		2.8	2	1.8	2
Adult Chimp	Experience					
	Agency		1.6	1.4		3.7
Baby Chimp	Experience					
	Agency		1.5			1.5
Tree	Experience					
	Agency		7	5.7	7.9	8.4
River	Experience		6.3		6.1	11.1
	Agency		6.3	7.8	6.1	11.1

Table 8: Regression on DV Moral Judgments and IV Mind Perception, Agents

3.5 Discussion of Correlations and Regressions

The following generalizations are true abstracting away from specific r and p values, and just looking at the number of effects as presented in the tables above:

1. Cryptomind patients were affected almost equally by Experience (20 detected effects) as Agency (21 detected effects).
2. Cryptomind agents were more effected by Agency (21 effects) than Experience (15 effects).
3. Experience was more relevant to patients than agents (20 effects versus 15).
4. The entity most effected by mind perception over all was the Android (17 effects), followed closely by the Tree (16 effects), River (14 effects), Dog (10 effects), Bear (9), Adult Chimp (7), and finally Baby Chimp (6).
5. The judgment most effected by mind perception was Murder (21 effects) followed closely by Punish (20 effects), In Order To (17), Wrong (13), and finally Kill (6).

At first blush 1, 3, 4, and 6 seems to be evidential support for the two sources of "moral standing" (moral patient status) hypothesis listed as the iii testable prediction of the moral status as mind perception paradigm (above, section 1). Jointly 1 – 4, 6 seem to support the paradigm overall (testable predictions I, and ii). However, the effect sizes in all cases are quite small. Recall that TDM advocates have said that mind perception is the essence of moral status (Gray, Young, & Waytz, 2012), where they define essance as both "the most signifcant element, quality or aspect of a thing" and "the ultimate nature of a thing" (Gray, Waytz, & Young, 2012, 207). The effect sizes we've discovered hardly support this claim.

Nevertheless I ran a series of regression analyisis to test the predictive power of Agency & Experience to moral judgments. I again found a series of small but real effects. The following generations are true abstracting away from specific p values and effect sizes.

6. There were more instances of regression models using Agency (16) than Experience (10) for patients.
7. There were more instances of regression models using Agency (17) than Experience (5) for agents.
8. There were more instances of the Punish question being effected by mind perception (patinet 7, agent 6), than any others. This was followed by Murder (patient 6, agent 6), In Order To (patient 5, agent 5), Wrong (patient 3, agent 4), and finally Kill (patient 2, agent 0)
9. Rivers and androids were tied with the most effected by mind perception judgments (patients 3, agent 7; patient 6, Agent 4), followed by tree (patient 4, agent 4), bear (patient 6, agent 0), adult chimp (patient 2, agent 3) then baby chimp (patient 2, agent 2).

Again we have some findings here that wouldn't be expected if the account of moral status offered in the TDM is accurate. Agency seems to have an larger effect on moral judgments about cryptomind patients than Experience does. Punishment judgments were more susceptable to mind percetion than judgments about whether the killing was wrong

or not. And the entities for whom mind perception was the most perdictive were the more alien cryptminds: the tree, river, and android.

4. Battle of the Sexes

One hypothesis I had was that women might be morally typecast (Gray & Wegner, 2009), such that women are typically perceived as moral patients more than moral agents. Recent evidence suggests that this hypothesis already has some experimental support (Reynolds, et al. 2020). In order to test this I provied a vignette in which two named work friends get into an argument, which results in one killing the other by pushing them into a deep empty swimming pool.

4.1 Methods

In two conditions the entities were the same sex (indicated by sex steriotypical names {Trevor, Brett, Mary, Jillian), and two condtions the entities were opposite sex (male agent, female patient; female agent, male patient; again indicated by sex specific names {Nancy, Richard}). Additionally each of these conditons had a TDC alternate. Yielding a total of 8 conditions of the vignette stimuli. As always order of presentation was randomized.

4.1.1 Stimuli:

4.1.1.1 Basic Vignette:

Please Imagine the following:

[Agent] are [Patient] are work friends. One day after work they get in an argument, and when no one is looking, [Agent] pushes [Patient] into a deep empty swimming pool, causing [patient pronoun] to fall to [patient reflexive pronoun] death.

4.1.1.2 TDC Vignette:

Please imagine the following:

[Agent] and [Patient] are work friends. One day after work they get in an argument. The next day when no one is looking, [Agent] pushes [Patient] into a deep empty swimming pool, causing him to fall to his death.

4.1.2 Measures:

Both the moral judgments and the mind perception measures were identical to those in sections 2 and 3.

4.2 Results:

I again used independent t tests to measures differences between moral judgments in the TDC and basic conditions, as well as sex of the entities {m, f}. And as before, I used a Spearman's test to check the if there was relationship between mind perception and moral judgments.

4.2.1 Moral Judgments:

The following figure 11 below gives the mean and standard deviations of the results for each vignette for each true false question.

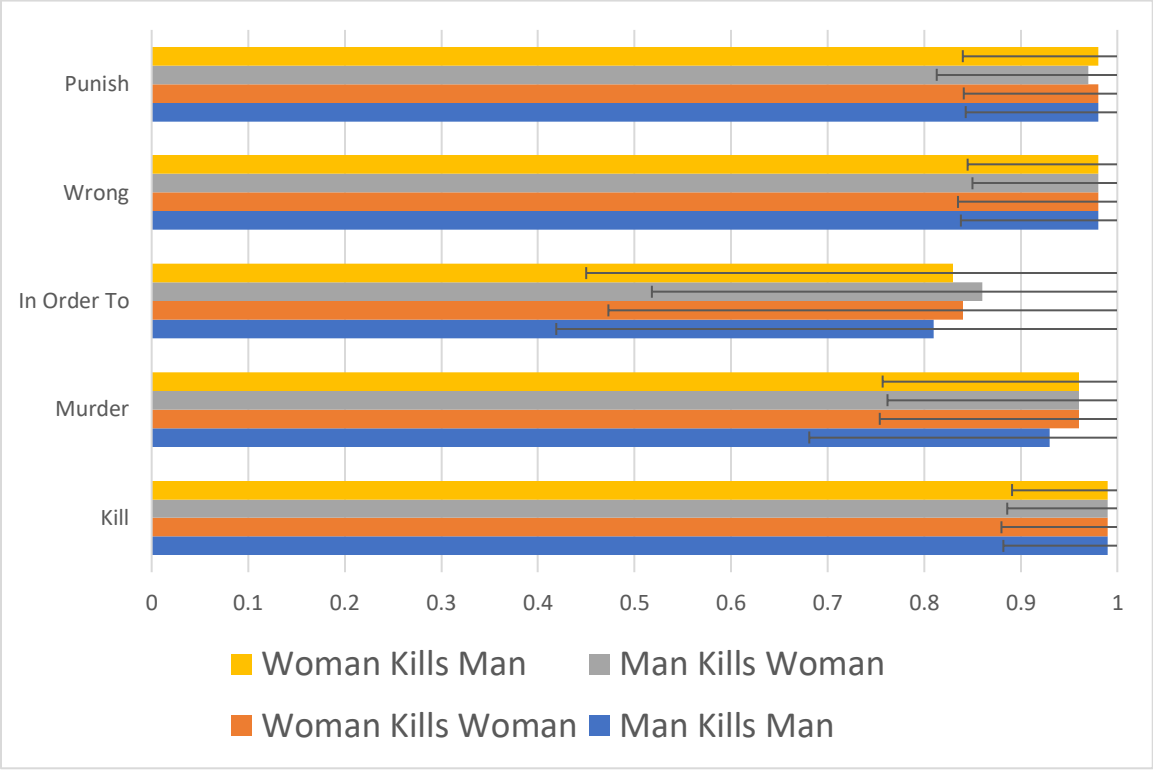


Fig. 15: Battle of the Sexes, Moral Judgments, Descriptive Statistics

4.2.1.1 Effect of TDC on judgments

An independent t ranging over all conditions indicated in the TDC murder (m= .96, SD .196) and In Order To (m= .86, SD .344) judgments were higher than in the basic conditions (m= .94, SD .233; M= .81, SD .394), t(2672)= -2.107, p= .035, and t(2670)= -3.832, p> .001. I therefore looked at the effects between conditions.

The TDC only caused a significant difference to moral judgments in the Man Kills Woman condition. An independent t test indicated that in the man kills woman condition the results for Murder, and In Order To judgments, were significantly higher in the TDC (m= .98, sd .125, p= .001; m= .92, SD .27) than in the baseline conditions (m= .93, SD .248; m= .81, SD .394), t(635)= 3.217, p= .001, Cohen's d .254, and t(635)= 4.21, p> .001, Cohen's d= .326. Small effects.

4.2.1.2 Sex effects on judgments

4.2.1.2.1 Female Patients in the TDC

An independent t test indicated that there was an effect of sex {m,f} on moral judgments {Murder, In Order To} in the TDC. More people considered men to have murdered a women (m= .98, SD .125), than a woman to have murder a woman (m= .95, SD .222). Levene's test for equality of variances indicated that they cannot be assumed here, therefore t(550.995)= 2.621, p= .009. Cohen's d = .165, a small effect. And more people considered men to have killed the woman intentionally (m= .92., SD .270 versus m= .84, SD .363). Again, equality of variances could not be assumed, therefore t(634.284)= 3.135, p= .002. Cohen's d= .249, a small effect.

4.2.1.2.2 Male Patients TDC

An independent t test indicated that there was an effect of sex {m, f} on moral judgments {murder}. More people judged the woman to have murdered the man (m= .97,

SD .169) than the man to have murdered the man (m= .94, SD .240), t(665)= -1.953, p= .05. Cohen's d = .145.

4.2.2 Mind Perception Measures:

An independent t test indicated that there was no significant difference between mind perception ratings {Agency, Experience} of men and women, as well as none between mind perception and thematic role {agent, patient}.

Aggregating over all these factors, experience for these entities was extremely high: Experience (m= .95.16, SD .177, mdn= 100); Agency (m= .94.37, SD .186, mdn= 100).

4.2.2.1 Correlations: Mind Perception and Moral Judgments

Spearman's correlation test indicated Agency and Experience measures for both agents and patients were significantly correlated with 4 moral judgments {Kill, Murder, Wrong, Punish}, however all effects were very or negligibly small: Agent = (Kill: Experience r(2672)= .116, p> .001. Agency r(2672)= .11, p> .001; Murder: Experience r(2672)= .093, p> .001. Agency r(2672)= .068, p> .001; Wrong: Experience r(2672)= .193, p> .001. Agency r(2672)= .184, p> .001; Punish: Experience r(2672)= .169, p> .001. Agency r(2672)= .15, p> .001.)

Patient = (Kill: Experience r(2671)= .122, p> .001. Agency r(2672)= .11, p> .001; Murder: Experience r(2671)= .095, p> .001. Agency r(2672)= .078, p> .001; Wrong:

Experience r(2671)= .2, p> .001. Agency r(2672)= .186, p> .001; Punish Experience r(2671)= .172, p> .001. Agency r(2672)= .157, p> .001.)

4.3 Discussion:

I did not find support for the hypothesis that women would be morally typecast as patients. Instead, I found a small effect of opposite sex (m,f & f,m) on moral judgments, but only in the time delay conditions. In those conditions, opposite sex agents were judged a bit more harshly. However, because answers to the TF questions are quite high across the board, it is possible that they think one or the other is "more wrong" or deserving of "more punishment". I could not assess that with these methods.

While I did find that mind perception correlated with moral status judgments positively, the effect sizes are exceedingly small.

5. Battle of Generations:

Children too have a kind of cryptomind, as we saw in Chapter One. I reasoned that on the TDM, children should be moral agents/patients in proportion to how much agency/experience they're perceived to have. I used the same paradigm as in the battle of the sexes. In two conditions a 25 year old and a 5 year old play the agent or patient role respectively. In two other conditions, a 5 year old plays both agent and patient role. One of those conditions is a TDC. Subjects saw only one of the conditions.

5.1 Methods

5.1.1 Participants

Participants were drawn from the first round of data collection. See section 1.1-2

5.1.2 Stimulus

5.1.2.1 Child and Child:

Imagine the following:

A five year old child, Mike, is playing with his preschool classmate and friend Teddy after school. Mike and Teddy get in an argument and[45], when no one is looking, Mike pushes Teddy into a deep empty swimming pool, causing Teddy to fall to his death.

5.1.2.2 Adult and Child:

Imagine the following:

A five year old child, Mike, is playing with his adult (25 year old) friend Teddy after school. Mike and Teddy get in an argument and, when no one is looking, [Agent] pushes [Patient] into a deep empty swimming pool, causing Teddy to fall to his death.

5.1.2.3 Moral Judgment measures:

The moral judgments are the same as in all other paradigms, involving the five true false questions {Kill, Murder, In Order To, Wrong, Punish}.

[45] The TDC condition here involves only the deletion of [and] and substitution of that with a paragraph break + "The next day" attached to ", when no one is looking…"

5.1.2.4 Mind Perception measures:

The mind perception measures are the same in prior paradigms, involving 4 questions, 2 per entity per Agency and Experience dimension.

5.2 Results:

5.2.1 Moral Judgments:

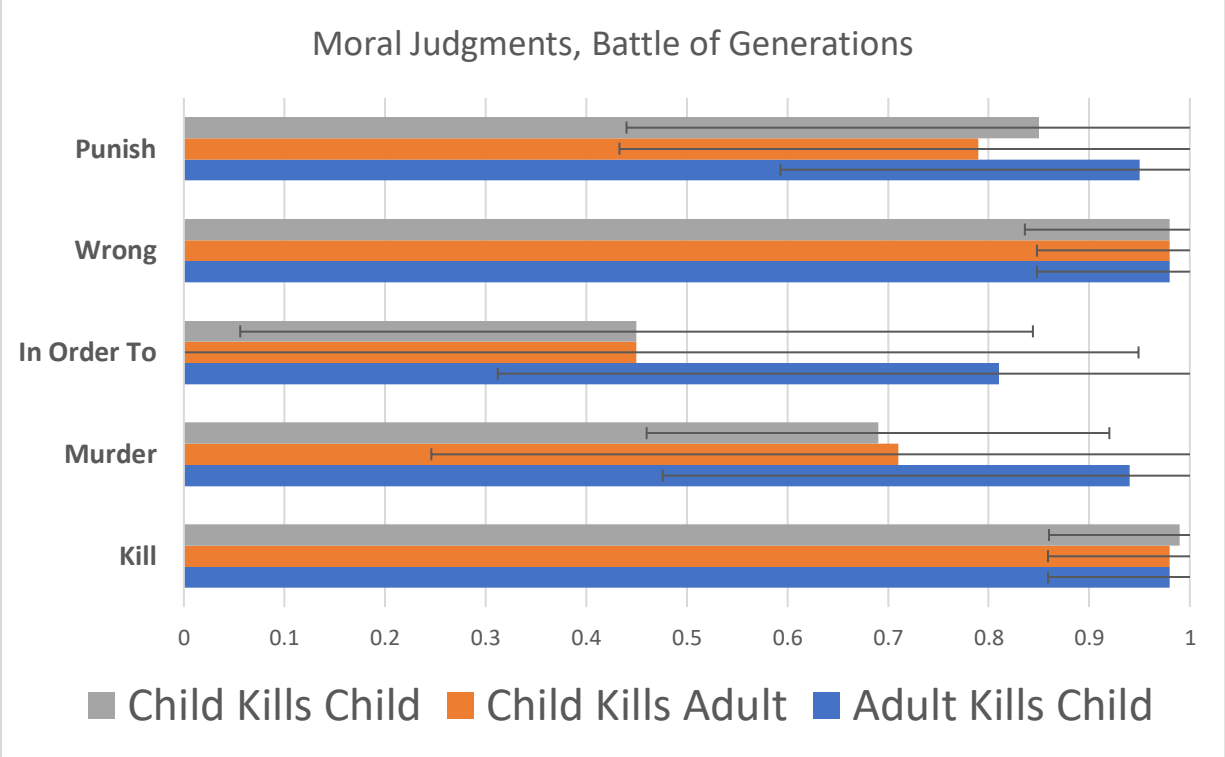


Fig.16 Battle of Generations, Moral Judgments, Descriptive Statistics

5.2.1.4 Moral Judgments and TDC

An independent t indicated that the event was more likely to be called a murder in the TDC (m= .73, SD .446) than in the basic condition (m= .65, SD .479), t(674)= 2.295, p = .022. Cohen's d= -.173, a small effect.

5.2.2 Mind Perception

The mind perception ratings for "a five-year-old child": Experience (m= 79.3, SD 24.8), Agency (m= 62.28, SD 26.55). Agency ratings were more normally distributed than Experience for the child.

The 25-year-old was rated similarly to how adult humans in general are rated. Experience (m= 94.9, SD 12.98), Agency (m= 94.07, SD .13.86).

5.2.2.1 TDC and Mind

Independent t indicated no relationship between time delay condition and mind perception as measured.

5.2.2.2 Theta and Mind

Independent t indicated no relationship between thematic role and mind perception as measured.

5.2.3 Correlations, Mind Perception and Moral Judgments

5.2.3.1 The Minds of Agents

A Spearman's correlation test found several significant relationships between the Agents Experience and Agency scores, and moral judgments: (Kill: Experience r(1333)= .071, p = .01; Murder: Experience r(1333) = .164, P> .001. Agency r(1330)= .268, p> .001; In Order To: Experience r(1333)= .093, p> .001. Agency r(1330)= .309, p> .001; Wrong: Experience r(1333)= .158, p> .001. Agency r(1330)= .062, p= .024; Punish Experience r(1333)= .171, p> .001. Agency r(1330)= .225, p> .001.)

5.2.3.2 The Minds of Patients

A Spearman's test found several significant relationships between Patients Experience and Agency scores, and moral judgments: (Kill: Experience r(1333)= .056, p=.041. Agency r(1330)= .058, p= .033; Murder: Agency r(1330)= .087, p= .001; In Order To: Agency r(1330)= .08, p= .003; Wrong: Experience r(1333)= .121, p>.001. Agency r(1330)= .055, p= .045; Punish: Experience r(1333)= .06, p=.029.)

5.3 Discussion

While mind perception has real effects on moral judgments in these cases, for the most part they are exceedingly small effects. We might expect that the Agency of agents is more relevant to moral judgments than Experience, that is not what I found. Instead, there are five effects of Experience, and four effects of Agency. Granted the largest effect is to do with Agency: subjects were more likely to say that the agent indented to kill the patient if they rated the Agency of the agent higher.

Likewise, we might expect the Experience of patients to be more relevant to moral judgments than Agency. That is not what I found. Instead, I found three effects of Experience, and four effects of Agency. All effects here were very small.

6 Replicating the cryptomind study with new mind perception measures.

While I didn't find any large effects of "Agency" and "Experience" measures, I hypothesized that this might be because those terms might have different meanings to the folk than the scientists who explicated them. I.e., "Agency" and "Experience" are labels that Gray and Wegner (2007) gave to two factors they discovered while investigating other mind perception. The factors lumped together capacities Gray and Wegner judged to be "Agency" like. For example, the capacity to plan, to remember things, to "have thoughts", etc. And "Experience" ranged over capacities to feel hunger, pain and pleasure, pride and joy, etc. Even though I told my subjects this in the stimuli material for the Agency and Experience measurements, perhaps it wasn't cognitively salient. Because of this I replicated a subset of the vignettes from section 3 using more fine grained measures for mind perception. I "un-lumped" the capacities Gray and Wenger factored together. This gave me 18 capacities to use as mind perception measures.

However, I also wanted to test to see if more sophisticated mental capacities might account for moral agency particularly. The philosopher Harry Frankfurt, for example, has argued that moral agency (at least normatively) deepens on the agent's capacity for second order desires (Frankfurt 1971). Additionally, I hypothesized that since pretending plays an important role in social learning, particularly with respect to

"theory of mind" (Friedman and Leslie 2007, Leslie 1987), that it might be salient to judgments of moral agency. Using these resources, I generated 10 more capacities to use as mind perception measures.

6.1 Methods

6.2.1 Participants

Participants were n = 558 Amazon Mechanical Turk (MTurk) workers (mean age = 39, SD = 12.9, range 18-87, 57% female).

6.2.2 Stimuli

6.2.2.1 Vignettes

The wording of the stimuli was identical to that of section 3. However, there was no TDC alternatives. And fewer entities {Man, Android, Bear, Tree).

6.2.2.2 Moral Judgments

Moral judgement stimuli were identical to that in experimental paradigms 2-5.

6.2.2.3 Mind Perception Measures

6.2.2.3.1 Gray and Wenger's 18 Capacitates

For each entity {Man, Android, Bear, Tree} subjects were asked to rate 18 different capacities from 0-6 ("not at all capable" to "extremely capable"). The text of these questions follows,

How capable is {entity} of…

Conveying thoughts or feelings to others?

Having experience and being aware of things?

Longing or hoping for things?

Experiencing embarrassment?

Understanding how others are feeling?

Feeling afraid or fearful?

Feeling hungry?

Experiencing Joy?

Remembering things?

Telling right from wrong and trying to do the right thing?

Experiencing physical or emotional pain?

Having personality traits that make it unique from others?

Making plans and working toward goals?

Experiencing physical or emotional pleasure?

Experiencing pride?

Experiencing violent or uncontrolled anger?

Exercising self-restrain over desires, emotions, or impulses?

Thinking?

6.2.2.3.2 10 New Capacitates: Theory of Mind, Second Order Beliefs and Desires, and Pretending

For each entity {Man, Android, Bear, Tree} subjects were asked to rate 10 different capacities from 0-6 (“not at all capable” to “extremely capable”). The text of these questions follows,

How capable is {entity} of …

Thinking about its own thoughts?

Thinking about the thoughts of others?

Pretending or playing make-believe?

Recognizing when others are pretending or playing make-believe?

Having desires about its desires, that is, of wanting to want things?

Having desires about others’ desires, that is, of wanting others to want things?

Having desires about its beliefs, that is, of wanting to believe things?

Having desires about others’ beliefs, that is, of wanting others to believe things?

Having beliefs about its desires, that is, of having beliefs about why it wants things?

Having beliefs about other’s desires, that is, of having beliefs about why others want things?

6.2 Results

6.2.1 Moral Judgments by Entity by Thematic Role

Bellow on coordinate graphs are the results per entity per thematic role.

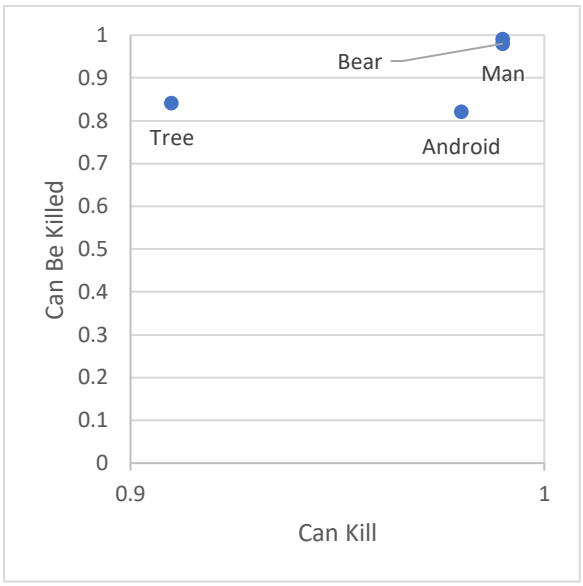


Fig. 17: Killable and Killer Cryptominds, Part Two

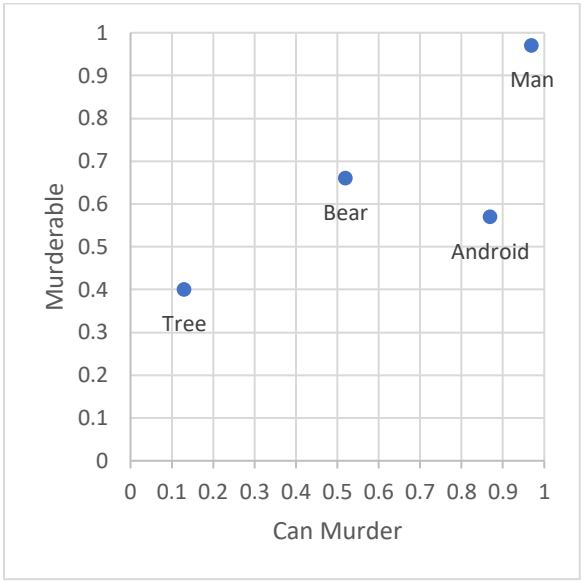


Fig. 18: Murderable and Murdering Cryptominds: Part Two

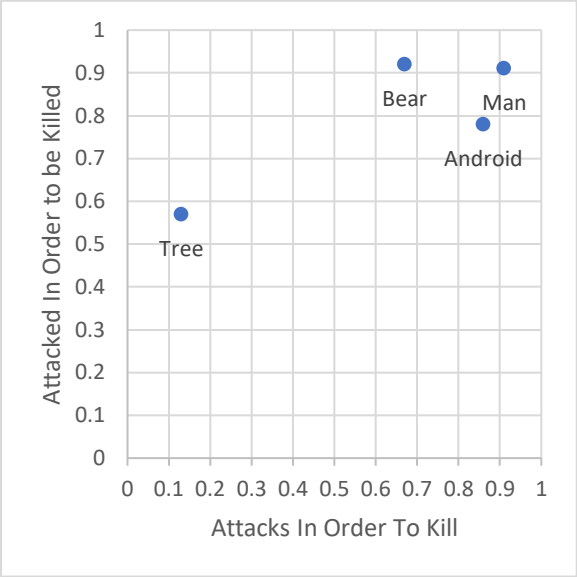


Fig. 19: Intentional and Unintentional Cryptominds, Part Two

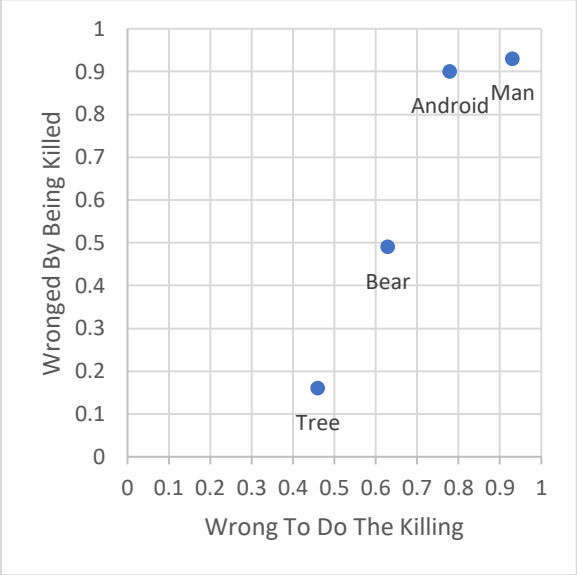


Fig. 20: Wronged and Wrongful Cryptominds, Part Two

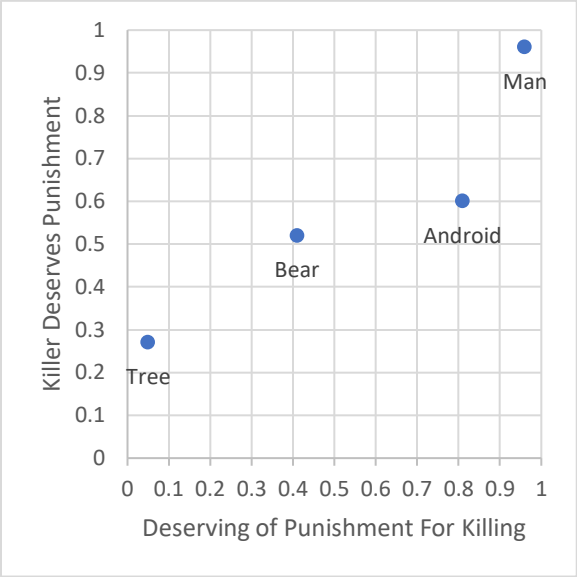


Fig. 21: Punishing and Punishable Cryptominds, Part Two

6.2.2 Mind Perception by Entity

Presented in the following for bar charts are the answers to all mind perception questions, per entity.

Fig. 22: Minds of Men

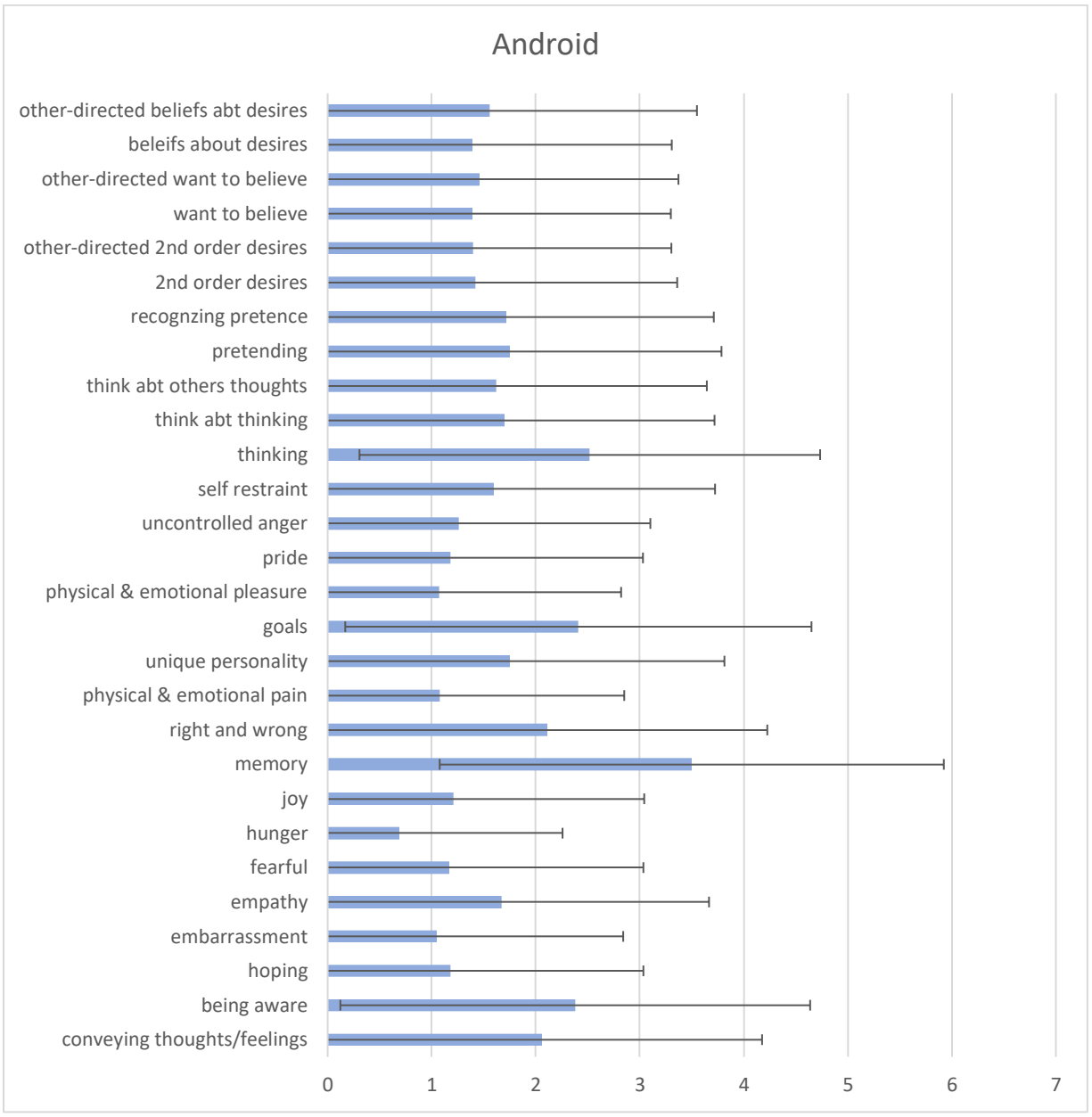


Fig. 23: Minds of Androids

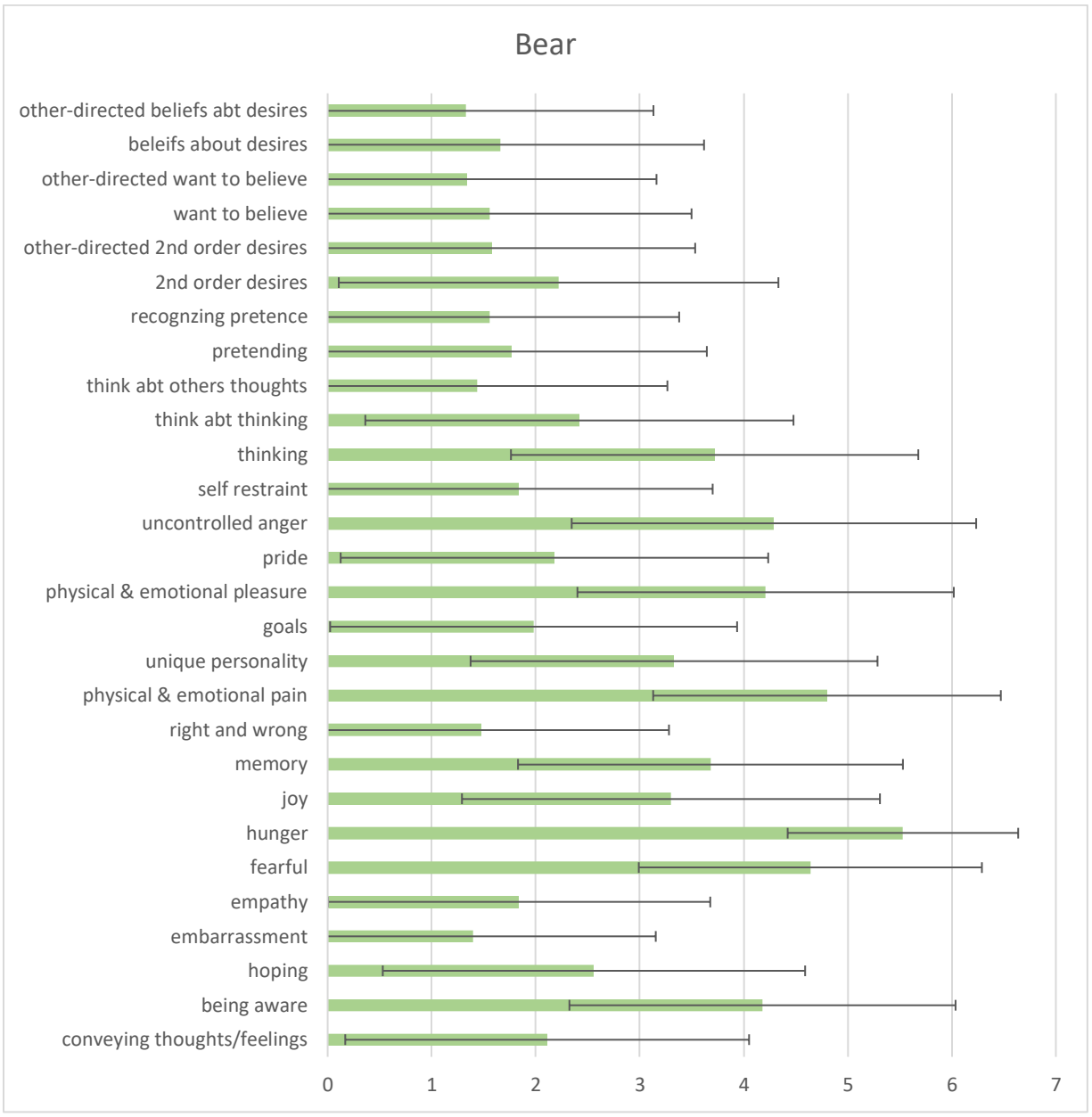


Fig. 24: Minds of Bears

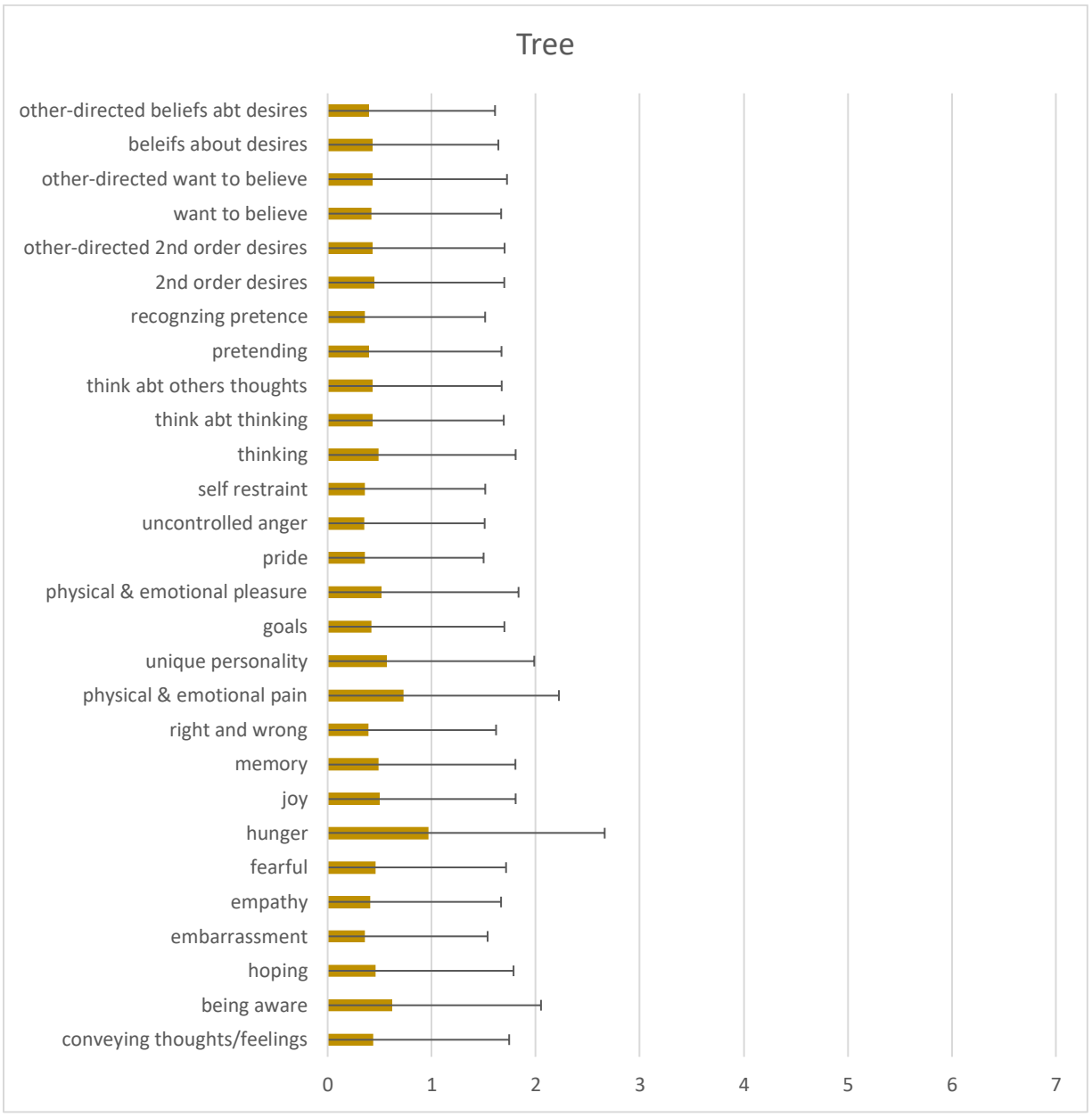


Fig. 25: Minds of Trees

6.2.2.1 Mind Perception by Entity by Thematic Role

An independent t test and Mann-Whittney test indicated that no differences in ratings of the 28 mind perception measures for the any entity depending on if they were the agent or patient of the vignette the subject was presented with.

6.2.3 Minds and Judgments

There were a number of correlations between mind perception and moral judgments. However, they all seemed to have very small effects, as before. So, I ran a binary logistic regression with all 28 mind perception measures as independent variables, against each true false question individually. Backward remove was used, excluding any variable $p < .051$. For no entity in either thematic role was any of the mind perception measures predictive of moral judgments.

6.3 Discussion

The basic descriptive results for moral judgments from section 2 were replicated here. Additionally, finer grained measurements were unable to detect any strong significant relationship between perceived mind as measured, and moral judgments as measured.

7. Review of Findings

We started our empirical investigating noting that the dominate explanation for moral status is the mind perception as moral status account championed by proponents of

the TDM. On this account moral agency is proportional to perceived Agency, and moral patience to perceived Experience (and Agency, if (Sytsma and Machery 2012) are correct). According to this account moral agent/patient are identical to thematic agent/patient (Strickland, Fisher and Knobe 2012), and so agents in morally salient events should be judged to have more Agency, and patients in a morally salient event should be judged to have more Experience.

I attempted to first check this by using a trolly like problem, weighing non-human lives {dogs, chimps} against one human life. There we saw that Agency and Experience were largely not useful predictors of moral judgments. That is, higher ratings of Agency and Experience really didn't help the non-human animals. Nor did changing the number of nonhuman-lives pitted against one human life.

Next I attempted to use the verb murder to detect moral status in a range of cryptominds {android, bear, dog, adult and baby chimp, tree, river}. I reasoned that [v, $murder_{df}$] = the wrongful, intentional, killing, of a moral patient by a moral agent. I was surprised to find that people were often willing to describe the events of one of the conditions as a murder, but as not intentional (i.e., answered true to [agent] murdered the [patient] and false to [agent] attacked the [patient] in order to kill it. In fact, across all experiments the "in order to" question received the lowest scores. This may pose a problem for my formulation of the predicate decomposition of murder (More on this in Chapter Four).

Moral agents were often viewed as deserving punishment for a wrongful action when the event wasn't described as a murder. For example, when river was the patient of

the event, Wrong and Punish answers were quite high (m= .74, SD.436; m= .63, SD .483). This case alone doesn't pose a problem for the TDM per say, though it might suggest that psychologists need a good account of the difference between intrinsic value (or moral status) and instrumental value. It may or may not provide support for the environmental philosophy which holds that ecosystems have moral patient status. Murder responses after all were low (m=.34, SD .48). One might imagine this is because rivers cannot be killed, however a startling number of respondents said they could be (m= .67, SD .47).

Importantly while I found that occasionally Agency and/or Experience was useful in a regression model, the amount of the variance explained by mind perception was always exceedingly small.

Mind perception was somewhat affected by thematic role in the conditions with cryptomind entities, however the effect all but disappeared when the entities were all humans.

My time delay conditions also had some small but real effects on moral judgments; however, they do show that the folk concept of murder doesn't include a sophisticated explication of mens rea. That is, malice aforethought seems to shift In Order To judgments and Murder, but only marginally. Most people view the baseline conditions to be appropriately called an intentional murder. So, since in all other cases we are shifting only one thing (the time delay language) the changes in judgment should be predicated on that thing.

My battle of the sexes study showed that there might be some effect of opposite sex killers on moral judgments, there were no general effects of sex/gender on mind perception or moral judgments.

My battle of the generations study showed that children are viewed very differently than adults. But we have a bit of a problem for the TDM account of moral status here. 5 year old children have higher average ratings of agency (m= 66.5) than adult chimps (m= 62), but children are judged less capable of murder (m= .56 versus m= .71 for adult chimps) while also acting less intentionally (m= .45 versus m= .69 for adult chimps). Children even had slightly lower experience ratings than the adult chimp (m = 71.6 versus m= 79.3) yet were judged to be murdered more (m = .94 versus m= .78).

The attempt to tease out effects by fissioning Agency and Experience into 28 individual measures of mental capacities largely found nothing. Overall, I was not able to support the claim that mind perception is the essence of moral status. It seems relevant, but marginally so. In the next chapter I explore the implications of these findings for the TDM in more detail and attempt to characterize the results in terms of Mikhail's account of moral cognition.

CHAPTER 4

PROBLEMS, CONCEPTUAL AND EMPIRICAL

1. Problems for and alternatives to the received view of moral status

The view that moral status is essentially related to, if not identical with, the mental capacities an entity is perceived to have is pervasive in both normative moral philosophy and descriptive moral psychology. Our data presented in Chapter Three poses some problems for this "received view of moral status", and here I want to spell out clearly what those are, and how seriously we should take them. I most directly tested the constructivist theory of dyadic morality (Schein and Gray, The Theory of Dyadic Morality: Reinventing Moral Judgment by Redefining Harm 2017), since it is a theory of moral status (as well as of moral psychology as a whole). Remember that this view fits in with the idea that moral status doesn't fit inside of a domain specialized cognitive mechanism concerning only morality. The view here is that the resourced used to make moral status judgments are really the same as those used to assess the Agency and Experience of the entities involved. Or as I said in Chapter Two, part of the broad faculty of morality FMB. Mikhail's alternative view of moral cognition requires that at least some features of moral status are part of a domain specialized mental mechanism. While I didn't attempt to disconfirm any of Mikhail's postulates, my data is certainly relevant. And more over it shows that Mikhail's simplifying assumption at the heart of his *Elements of Moral Cognition* book (Mikhail, Moral Cognition and Computational Theory 2007), that moral status questions should be held constant in experimental questions

about the operative rules of our moral psychology, while obviously useful at first, has to be discarded.

First, I argue that my data generate a few particularly acute empirical problems for the received view. This is because my experimental program focused on judgments of "murder" and how those change based on the perceived minds of the entities participating in the action of the event. On its own lights, murder represents a central or paradigmatic[46] case that the research tradition should be able to easily solve. According to proponents of the received view, all of moral psychology involves judgments about harm (Gray, Schein and Ward, The Myth of Harmless Wrongs in Moral Cognition: Automatic Dyadic Completion From Sin to Suffering 2014) (Gray, MacCormack, et al. 2022). Harms here can be nonphysical and run on a continuum from the clearest cases like murder, to distal cases like impurity. According to this view murder is a paradigmatic case of harm,

[46] In this chapter I will refer both to research paradigms, following Khun**Invalid source specified.**, and research traditions, following Laudan**Invalid source specified.**. I don't, however, want to be bogged down by their particularly different accounts of the history of science or the philosophy of history. I will follow Laudan more closely in that whenever I say "paradigm" or "research traditions" I mean researchers who can be lumped together because of their shared metaphysical and methodological commitments. But I will also follow Kuhn in my discussion about particular "paradigmatic" cases, i.e., exemplar cases, that are central problems for a particular tradition. The motivation here being that a paradigm provides guidance on exactly what empirical problems should be easily accountable for and which are liminal and so more difficult. And I suggest that central problems, empirical or conceptual, are more weighty than peripheral ones. A claim Louden would cash out with a conversation about how our auxiliary hypothesis are constantly tested by experimentalists who are adherents to a research tradition. A central case there involves one in which the auxiliary hypothesis was assumed by many researchers and so threatening it has wide ranging implications for other solved empirical problems. I find it easier at least discursively to appeal to exemplar problems or cases in the Kuhnian sense. Particularly here since "murder" hasn't been tested by lots of different researchers. Nor even the aspect of the received view that commits it to the idea that all harms are of a kind, and on a continuum.

representing the clearest crispest example of the members of the harm conceptual family. If all these claims from the received view are true, then, all else equal, moral status judgments should be strongly predictable from ratings of perceived mind, especially in cases of the most extreme harm on their scale (murder). My data do not support this conjecture.

Second, I argue that the received view and its proponents are vague about the different subsystems of the moral faculty and how they interact with moral status and mind perception. That is, researchers have often not distinguished, through their method or discourse, between moral evaluations of an action in some circumstances, and moral evaluations of justifiable punishment for some bad action in some circumstances. This problem could be acute just in case mind perception is more involved in judgements of justifiable punishment than evaluation of a moral event, as we saw might be the case, in Chapter One, when we examined the difference between juvenile justice and regular criminal court. I also here outline three other conceptual problems with the account.

With these two kinds of problems more clearly outlined, I argue that together they represent a serious, if not fatal, mark against the received view.

I then outline a problem and false prediction for my own experimental program with respect to the folk concept of murder. And I ask if it turns out that I am justified in using murder to test an entity's moral status at all. Here we see that contrary to my own expectations, subjects were often comfortable saying an entity murdered another entity, but not intentionally. Or that an entity murdered another, but that it wasn't wrong. Our project then has important implications for the understanding the folk concept of murder

as opposed to the richer legalistic concept which entails both actus reus (a harmful action) and mens re (intention regarding the harmful action).

After all this we take some stock of where we are; particularly at what pictures of the psychology of moral status we are justified in endorsing/rejecting.

2. Conceptual Problems the Received View

The received view suffers a few conceptual problems as well. I've so far highlighted the vagueness of the theory with respect to the difference in mental machinery involved in moral evaluation versus judgments of justified punishment. We will return to this issue towards the end of this chapter in its own section. But here I want to point at three more conceptual problems with the received view.

1) The problem of vagueness about the quantitative relationship between mind perception and moral status.
2) The problem of assuming perceived agency relates to moral agency in the same way that perceived experience relates to moral patiency.
3) The problem of ordinal predictions which explain only the vaguer facts.

Taken together these are some of the problems my experimental program was intended to answer. But we must clarify them a bit before answering them.

First, the received view is vague about the relative contribution of each of unit of perceived mind to moral status judgments. Let us, for this example, focus on moral agent status and perceived Agency (even though on the received view these remarks will

largely hold for moral patient status and perceived Experience as well). Remember that the received view claims that mind perception is the essence of moral status (Gray, Young and Waytz 2012): that is, the most important and crucial aspect of it. Particularly for moral agency, the perception of an entity's capacity for "Agency" in the sense outlined in Chapter Two, is what is relevant[47]. The received view has moral agency as well as perceived "Agency" on a continuum, from 0 (non) to 1 (full). But it never clearly specifies if "perceived Agency" is in a one-to-one proportion to moral agency. We can clearly imagine other relationships between perceived "agency" and judgments of moral agency other than liner, and this is never examined by the proponents of the received view – though the way they graph and present their findings, the assumption is that it is linear just because that's also how they represent perceived minds.

Second, it is also an open empirical question as to if the relationship here is different for the operation(s) that takes perceived Agency as an input and generates moral agent judgments, and the operation(s) that move from perceived Experience to moral patient judgments. This is just to say that argument for the received view has an obvious enthymeme, that mind perception effect moral agent status in the same way as it effects moral patient status. But here we are actually in a state of deep ignorance. It is quite possible that, for example, moral agency is represented not as a scale variable at all, but as a nominal one. And that moral patiency is scale or ordinal, not nominal. Or any strange combination of those. It is possible that Experience relates to moral patiency linearly,

[47] Though we saw some alternatives here, suggesting that experience might also be relevant to moral agency status, the discussion here will keep to the simpler model proposed by Kurt Gray at first.

whereas Agency relates to moral agency logarithmically. The fact that the received view is incurious about this relationship seems untenable.

This problem has a similar flavor to treating mind perception's effects on moral status the same for moral evaluation and justified punishment. It is an unmotivated assumption that hinders us from even seeing what empirical problems exist and are worth solving. If I've made progress over the course of these chapters I hope it is in this effort especially. Helping see what kinds of questions are worth asking about our existing psychological theories of moral status.

Third, the theory is additionally vague in that it gives no specific predictions based on its own model, but only relative predictions that attempt to explain the vaguer facts. Let me expand here. There are clear testable predictions implied by the received view. That was just one of the merits I mentioned about it in Chapter Two. Those involve comparative cases. For example, Gray and Wegner's "kitten (low agency) scratches man (low experience)[48]" versus "man (high agency) hits kitten (high experience)": the parenthetical remarks about agency and experience do all the work in explaining our differences in evaluative judgments (i.e., that the first is not a moral wrong, and the second is). So we could easily rephrase it with variables: "X (low agency) scratches Y(low experience) = not forbidden". And from here generalize to a 4 x 4 table[49] for every combo of low and high agency and experience.

---

[48] As mentioned before, saying a man has "low experience" is counter to Gray and Wegner's own data (not to mention my own). So, I'm not sure why they say this.

[49] As long as we keep the adjectives in place as useful and predictive parts of our account, rather than the scale model actually assumed by the received view and my own experimental paradigm

	High Agency (agent)	Low Agency (agent)
High Experience	Forbidden	Midway[50]
Low Experience	Midway	Not Forbidden

Table 9: Gray and Wegner's Moral Judgments from Mind Perception

This kind of approach enables only ordinal predictions of an entity's moral status, relative to other entities moral status. So, if some entity E1 has perceived agency N, and some other entity E2 has perceived agency > N, E2 will be more of a moral agent than E1. Again, this is comparative and ordinal, so we have no specific claims about just how much an incremental change one entities perceived Agency (or Experience) is supposed to change moral status judgments[51]. That is, we have no direct or quantitative relationship

[50] One problem arises for the interpretation of midway between forbidden and not. This is not some new deontic category, but rather would be reflected in the data as fewer people finding the circumstances acceptable; (I.e., it's a prediction about population level behavior). According to the received view, we must interpret "forbidden" as a technical word, the name we call a strong intuition that some harmful action is wrong (according to the TDM, all else equal, because of the differences in perceived minds of the entities involved). It is not in fact a proposition generated by the mental machinery involved in moral judgments[50]. It is in that sense an intuition represented by a continuum of badness. The badness of some harmful action (again holding that action-description the same) scales with the perceived Agency of the agent and the perceived Experience of the patient. So, the products of the received view of moral status are not the output of a specialized, deontic, logic.

[51] This is particularly strange since many of the experiments within this research tradition take the same entity and the same description of an action, but change the description of the entity's mental capacities, and then illicit judgments like "is it permissible to experiment on such and such an animal". In principle this gives us a way to try and measure exactly how much additional perceived agency increases moral status judgments. However, because the experimentalists don't have a measure of just how much more perceived Agency and Experience is gotten out of the different descriptions (it is just assumed that a description of what the experimenter thinks counts as is a richer

between Agency and moral agency (as outlined in the prior problem), and this means if we have only one case where all the relevant information is known, we cannot make a prediction about the kinds of moral judgments a subject will be more likely to make. So we cannot, for example, use the theory to predict subjects judgments in the following way: A has agency x, and B has experience y, therefore our model generates moral judgments = ? in all conditions where A harms B. This is what I mean by we cannot explain the specific facts with the received view.

It is however rather clear that, given their graphing of their data about mind perception and moral status, Gray and his colleges are supposing the relationship is linear and symmetrical between Agency and moral agency, Experience and moral patiency. As we highlighted in Chapter Two, ordinal predictions get us somewhere with respect to testability. However, my data does not provide support for even those vaguer ordinal predictions: just because some individual thinks a dog has just as high agency as a child, does not mean they judge them to be equally wrong or blameworthy for their actions when they both attack and kill a man. As we saw in my trolly alternative experiments, a chimp is rated significantly higher than a dog in both perceived Agency and Experience, but it is not judged to be significantly more morally valuable than a dog when compared to a human (CH3. 2-2.4).

It seems reasonable to claim that my data counts against the received view more to the extent that its own vaguer predictions (by any measure easier to meet) cannot be

---

description of the mind of an entity and that that will be interpreted exactly that way by the subject) no such scale can be developed from this work.

confirmed. Vaguer predictions are a lower bar to hurtle, and if your account doesn't get them right then there is some good reason to think the account itself is flawed. By a similar logic, not being able to solve central/paradigmatic cases for a research tradition, like judgments about murder for a harm centered moral psychology, should count more than not being able to solve peripheral cases.

All this may undermine, in some ways, the question about the quantitative relationship between mind perception and moral status. My data doesn't show a strong relationship at all. However, some kind of ad hoc hypothesis might be reasonably appealed to in order to save the received view from apparent disconfirming evidence. For example, perhaps if the relationship between perceived Agency and Moral Agency follows the least noticeable difference principle a Webarian analysis could be a way to generate/specify a threshold of Agency[52] below which an entity doesn't qualify as a moral agent at all. At least conceptually, some threshold of that sort could be the bar above which the received view's ordinality prediction holds.

Introducing this kind of additional theoretical paraphrenia after the fact in order to save our hypothesis obviously comes at some cost. If the threshold rule exists, (unless it falls out of Weber's law for free[53]), it seems to shift a great deal of focus away from mind

[52] I'm still focusing on perceived Agency here and moral agency, but all these remarks apply equally well to perceived Experience and moral patiency. Indeed, they could have different thresholds for each. Or the thresholds could be something arbitrarily set by the local cultural.

[53] This seems unlikely to me. While applying a logarithmic relationship to two continuous variables does cause thresholds to fall out "for free" as it were, the labeling of one of the many thresholds generated as the one relevant for moral status must be motivated by some other mental machinery – presumably via a learning algorithm of some kind. And again, this seems to move attention away from mind perception itself and

perception itself, as relevant to moral status, and onto whatever learning algorithm allows for the acquisition of the operation that labels whatever threshold as the relevant one. Whatever, the input conditions for that learning algorithm would be, the algorithm itself would seemingly be a large, dare I say essential, part of the acquisition of moral knowledge about moral status. So, we've replaced one problem with another, not an encouraging sign of progress. But not wholly illicit either. In any case, my experiments don't obviously show evidence that these thresholds exist when it comes to mind perception and moral status. But they were also not designed to test this.

Detailing these conceptual problems shouldn't be viewed as a wholly destructive project. I hope elucidating these problems can potentially serve researchers within the received view tradition, just as much as those outside of it, in formulating interesting questions to test experimentally. Obviously more experimental work is called for to answer these questions.

3. Empirical Problems

Our major empirical problem to be addressed is the lack of usefulness of ratings of Agency and Experience in predicting moral status judgements in cases central to the scientific paradigm (i.e., extreme harm/killing). Philosophers of science have long argued that prediction is intimately connected with ideas about scientific explanations. Whatever we think about the vast philosophical literature on scientific explanation, at the very least

---

onto whatever the inputs into the labeling operation are/the learning algorithm that allows you to acquire the labeling operation.

we should agree that if the received view is incapable of making good predictions, it doesn't matter much if they think they have a reasonable causal/mechanistic model.

3.1: Empirical Problems in the Driverless Car Study

Recall that in CH3, 2.2.1, we reported the descriptive statistics for my trolly alternative problems. There we reported that there was a significant difference in both Agency and Experience ratings for Dog and Chimp (ordinally Chimp > Dog for both Agency and Experience). That means that on Gray's proposal where moral patient status is fed only by perceived Experience (Causal Graph Fig., # 03), and on Machery's proposal that moral patient status is feed both by perceived experience and agency (Causal Graph Fig., # 04), we should expect to see a significant difference in moral patient status judgments about chimps versus dogs. To gauge this, however, we first needed a known moral agent, the person programming the car, and a known moral patient, the human pedestrian(s), to make sure our structurally similar cases are judged the same as standard trolly problems. Subjects responded to our well controlled case involving both those entities (the programmer and the pedestrians) the same way they tend to respond the basic trolly problem (with its switch-flipper and men on the tracks), with roughly 90% saying it is permissible to sacrifice 1 to save 5 by causing the train to switch tracks, and 10% saying it is impermissible (CH3, 2.2.2).

Next, we swapped out the human pedestrians for entities whose moral status was unknown a priori. So, the tradeoff there was 1 chimp against 5 chimps. And there too we saw the same 90/10 response pattern from our paradigmatic case. This shows us that

inserting an entity with a cryptomind who's moral status is not known a priori doesn't itself cause problems for the psychology responsible for churning out moral judgments. The structurally identical events were evaluated the same, irrespective of the presence of an odd entity. And this is promising for our experimental set up. (We assume that if all the patient entities were described as small rocks, subjects would not be inclined to make any moral evaluation of the difference involved in switching lanes.)

Next, we compared the tradeoff involving one human life versus 5, 10, and 20 chimps/dogs. On hypothesis, we should expect that moral patient statuses are commensurable with each other. In that way we should be able to discover how many dog or chimp lives are worth one human life. First, we should expect that if a Chimp is viewed as having 70% the capacity for Experience that a human does, then, on theory, they should be 70% the moral patient that a human is. We do not have an experimental case that asked about a tradeoff between one chimp and one human. However, unless the addition of 4 other moral patients reduces the moral value of the patient group instead of increasing it, then this is not close to being the case. Only about 20% of respondents say that it is permissible to sacrifice one human for 5 chimps/dogs. Because we don't have any hypothesis as to how to add moral patient values up across entities, we are against left with only a vague prediction that is not met. Also recall that there was a statistically significant difference between a Chimp's perceived mind ratings and a Dog's, but not between ratings of the permissibility of trading off chimp lives for a human, dog lives for a human. That is, as implied by the received view, we should predict people are more

willing to tradeoff a human life for chimp lives, than a human life for dog lives. And we cannot support claim with our data.

More problematic still is the fact that permissibility responses here didn't significantly change as the number of dogs/chimps increases from 5, to 10, to 20. We must here set aside the issue the additive value of each moral patient. Just as we will largely ignore cases of multiple agents (principles and accessories to a morally salient event). Obviously they're deeply relevant to Mikhail's projects (CH2, 3), if we take his postulates as empirical hypothesis as intended. However, these are complications that must be approached once a clearer idea of how to measure moral status emerges.

So, we have seen that ratings of Agency and Experience in our first experimental set up do not predict ordinal relationships between the moral patient status of dogs and chimps, nor do they predict if a subject will be willing to trade the life of one human for five, ten, or 20 dogs/chimps. This is clearly a problem for the received view.

3.2: Empirical Problems for the Received View, Murderable Minds Study One

Predictive Failure of Agency and Experience

In (CH3, 3.4-5) we saw that while there were detectable effects on of Agency and Experience ratings on some of the true false questions, they did not align with the predictions of the received view of moral status. Again, we were expecting to see Agency have an essential role in judgments about if an agent is a moral agent, and Experience to have a similar role on perceptions of moral patiency.

We should also expect trans entity-type consistency pegged to Agency and Experience ratings but we did not; i.e., it's supposed to be the perceived minds of the entity, not the type of the entity, that drives the judgments. See (Ch3, 3). However, Dogs and Bears, with significant overlap in mind perception, were not related identically. Androids were treated as moral agents more than either of those entities even though the mean Agency scores from Androids were on average lower than both.

3.3: Empirical Problems for the received view, Murderable Minds, Study Two

Testing mind perception more got us no closer to confirming the predictions of the received view. (CH3, 6.1-6.3). If some part of mind perception is the cue for moral status, this study was reasonably well designed to detect that. Here we have evidence that might be most interesting to philosophers. An entity's capacity to feel pain wasn't strongly tied to people judgments about if it was a moral patient (Singer 2009). Nor was an entities capacity to have second order volition relevant to its perceived moral agency (Frankfurt 1971). More over philosophers might be interested in taking the mind perception scores and showing how close or far they are from a scientific understanding of those entities actual capacities. Thereby making some progress in terms of moral education. However, this also seems to be a problematic project, since even if the perceptions tracked real world understanding of the actual capacities of the entities in question, it seems like those perceptions don't make a difference to moral judgments.

4. Moral Status and the Folk Concept of Murder.

Recall that going into my "murder" experiments, I needed some hypothesis about what the logical structure of folk concept of murder was. We hypothesized that the paradigmatic concept of murder entails (i) a killing, that was (ii) intentional, (iii) defeasibly wrong/inexcusable, and which causes the agent of the action to be (iv) deserving of punishment. In Chapter Three I have listed two more conditions, that the agent be a moral agent (partial or full) and that the patient be a moral patient (partial or full). I need something like this to be true for my experimental program have a hope of testing if the entities playing the agent patient roles are moral agent/patients. If murder is not a prima facies wrong, if it doesn't involve a death, if it doesn't involve an intentional killing, then maybe it just doesn't make sense to hypothesize that it must involve a moral agent/patient pair. That assumption allows us to test if someone is a moral agent/patient by looking at their answers to the question of if our vignette describes a murder. If all other conditions are met, if it was an intentional, wrongful killing, deserving of punishment, and appropriately described as a murder, we want to be able to say that the entities involved are moral agent/patients.

Just to be clear, while I argue that murder as intentional wrongful killing is the paradigmatic meaning of murder, I also recognize that the word is polysemes and does not always denote a crime. Sentences like "You just murdered that burger!" don't describe a homicide. Rather, it is perfectly interpretable as a kind of hyperbole/joke that depends for its humor on a mismatch between "murdered" and "ate" in argument structure[54] and emotional valence. We think it is implausible that respondents could read

[54] The object of murder must be a thematic patient, the object of eat need not be a patient.

our vignettes and think we had this hyperbolic sense of the word in mind when we asked whether "The [agent] murder the [patient]" is true or not. However, taken together our 5 conditions of murder, in the form of true/false questions, allow subjects this degree of freedom in their responses. They could, for example, endorse the idea that the vignette they read described a murder, but didn't involve a killing, didn't involve the intention to kill, wasn't wrong, and wasn't an action deserving of punishment.

In critical fairness with respect to my own experimental paradigm, these conditions were not always met, and I report on those results below. I recognize also that even when the word murder is used to denote a crime (the intentional wrongful/unjustified killing of a moral patient), subjects might say that the vignette described a murder but that the murder wasn't wrong and/or doesn't deserve punishment. Utilitarian philosophers are fond of thought experiments that make this point: say you're told you must kill your mother or else the government will torture and kill 10,000 people. In extreme examples like this, someone might reasonably think that in intentionally killing your mother a murder occurred, but that in the final analysis the murder wasn't wrong. Again, it is implausible that respondents judged our vignettes in this way, since none of them involve this kind of moral dilemma. Nevertheless, our true/false questions allow for this degree of freedom.

Of special interest will be two sets of respondents. The set of those who say the vignette the read described a murder (i.e., answered true to "the [agent] murdered the [patient]"). And a proper subset of this set, those who answered yes to all the true/false questions. We can call the first set "murder, thin" and the second "murder, thick". The

murder, thick group, comport basically with my a priori expectations about the folk concept of murder. That if a subject called the action described by my vignette as a murder, they would always say yes to all the other questions. The murder, thin group violated my expectations by answering yes to the murder question, and no to at least one of the other true/false questions.

4.1: What do we do with murder thick and thin?

Since I have argued that in order for "murder" to be a good experimental tool, it really should meet all my conditions, we may be in some kind of problem here. To solve this I could restrict my analysis to the subgroup of murder thick respondents. However, this seems to me to be elicitly excluding respondents because they failed to have the same intuitive idea of murder that I do and hypothesized they would share. Even if we do restrict ourselves to the murder thick subsets when testing the prediction that mind perception feeds moral status, we're in no better predictive shape; i.e., Agency/Experience still doesn't help us predict membership in the murder thick set any better. Instead, we've discovered a more complicated account of the folk concept of murder. One that should be explored more in future work.

5. Revisiting Results from Experiment Two

Recall the methods, subject demographics, and results of experiment two (CH3, 3-3.5) which we called Murderable Minds.

5.1: Does our vignette describe a murder?

In the condition in which a man kills another man (n = 188), percent true answers to the five true/false questions were as follows: 99.5% endorse “the man killed the other man”; 96.8% endorsed “the man murdered the other man”; 91.5% for “the man attacked the other man in order to kill him”; 96.3% for “the man was wrong to attack the other man”; and 96.8% endorsed “The man should be punished for killing the other man”.

This means that 96.8% (n=182) of respondents were in the “murder, thin” set: i.e., endorsed ‘the man murdered the man’ as true. And 88.8% (n=167) were in the “murder, thick” subset; i.e., endorsed as true all the true/false questions.

Narrowing our frequency analysis to the “murder, thin” set of cases, 99.5% of these respondents said that the agent killed the patient, 94% said that the agent indented to kill the patient; 98.4% said that the agent was wrong to kill the patient; 98.9% said that the agent should be punished. Which is a rather high degree of concordance here. Notice the least supported part of the concept of murder was the “intentionality” of the agent. The “murder, thick” subset comprised 91.8% of the “murder, thin” set.

While subjects did violate my expectations that answering True to the murder question would lead to total conformity on the other questions (all True), when I looked to see if agency or experience was driving those differences, I also didn’t find any strong relationship between mind perception and any other individual true false question that would better support the received views position.

Therefore, this comparison between two known entities with moral status (generic men) gives us some baseline for comparison here going forward, as we complicate the

picture by changing one of the entities with known moral status to another entity with unknown moral status.

5.2: Does entity type matter?

5.2.1: What can murder?

CH3, 3.2.1.3 show the results "murder" question results for all non-human agents. The "murder, thick" subset did not always comprise a majority of the "murder, thin" set. For android agents, 80.9% (83.8%, TDC) of the "murder, thin" set is comprised of the "murder, thick" subset. For bears this rate was 39.2% (42.5%, TDC); for trees it was 13% (20.8%, TDC); for rivers it was 9.1% (13.6%, TDC); for dogs it was 66.3% (67%, TDC); for the adult chimp it was 60.3%; and for the baby chimp it was 50.6%.

The small n's for the river and tree conditions make the percentage changes there look stark. In all four of these conditions the n of the "murder, thick" subgroup was 5 or less in all cases.

As we can see, the time delay variants had a moderate effect on if respondents consistently endorsed all five true/false questions as true (CH3, 3.2.2). Obviously, this is relevant to a conversation about mens re, k-generation versus i-generation for Mikhail, and not at all to the received view of moral status. Remember the received view deemphasis's act-descriptions (Dillon and Cushman 2012), and therefore has no theoretical paraphrenia to account for an agent's intended goals versus foreseen but unintended side effects of an agent's action.

5.2.2: What can be murdered?

In the condition with a bear patient (n=224), 67.9% said the sentence “the man murdered the bear” was true. Of this subset (n=152), 94.1% said that the man attacked the bear in order to kill him; 78.3% of respondents judged that “the man was wrong to attack the bear”; and 70.4% endorsed the idea that the man deserved punishment for the attack on the bear. This means that 5.9% of respondents endorsed the idea that the man murdered the bear but didn’t intend to kill it; and that a sizable portion of respondents felt that the man murdered the bear, but wasn’t wrong to do so (21.7%) and doesn’t deserve punishment for the killing (29.6%). Only 67.1% (n=102) of respondents who agreed that this vignette described a murder, also agreed to the other four true false questions.

Almost all cases where “a man” was the agent and a nonhuman entity was the patient saw this pattern of decline, wherein, out of the cases in which the respondent said the agent murdered the patient was true, respondents were more likely to say the agent acted intentionally than they were to say that the agent was wrong to kill the patient, and more likely to say the agent was wrong to kill the patient than that the agent should be punished for killing the patient. The one exception is the case in which a river is the patient of the action (n=214). In that case, of the respondents who said the man murdered the river was true (n = 73), 83.3% of them said the man filled-in the river in order to kill it; 91.7% said the man was wrong to do so; and 86.1% said he should be punished for the action.

Of respondents who said that “the man murdered the [patient]” was true, only a subset agreed with all the other 4 true false questions: when the [patient] was a bear

67.1% (102 out of the 152 who agreed that the condition described a murder) agreed with the other four true false questions; when the [patient] was an android 71% (83 out of 117); when the [patient] was a tree 43.75% (28 out of 64); when a river 71.23% (52 out of 73). In most cases, the narrative time delay conditions moderately increased the consistency between endorsing that the man murdered the [patient] was true, and true answers to the other four true false questions. In the case of the tree [patient], consistency between the true false questions increased to 57.3%; in the river condition to 81.58%; in the android case, up to 74.68%. For bear and man patients, the time delay caused basically no change in overall true false answer consistency.

5.3: Does gender matter?

In Chapter Three, section 4, I reported on exploratory findings related to perceived gender and moral status. The results here don't bear directly on the major thesis of the TDM, that mind perception feeds moral status, because men and women were rated the same in terms of Agency and Experience (CH3, 4.2.2). So we will set them aside.

5.4: Does age matter?

In Chapter Three, section 5, I reported on exploratory findings concerning age differences and moral status judgments. Here I did not find significant support the received view either. People varied on their ratings of children's minds (5.2.2) however they largely agreed that the when the child was the agent, they had committed a murder,

and deserved punishment. Though they were less likely to say that the child acted intentionally (5.2.1).

While these vignettes are a bit different in description than those from my first murderable mind experiments, they are structurally quite similar. While directly comparisons should be avoided here, it's worth noting that children are viewed as moral agents much more than chimps, even though they are rated as having similar kinds of minds.

6. Mikhail and Moral Status

Mikhail quite clearly abstracts away from problems of moral status in his work for well-motivated reasons. Which means that his account of moral grammar is largely silent about questions of moral status, except where it cannot help itself, in the postulates outlined to some extent in Chapter Two. If we had discovered a relationship between Agency and Moral Agency, Experience and Moral Patientcy, then we could have without much fuss modified his postulates as proposed in two (changing "persons" to "moral patients" and interpreting those through whatever relationship it does have to perceived Experience, etc). But without that we are left in a difficult position. While our generalizing of Mikhail's postulates in Chapter Three (3.1-2) was motivated by a desire to make them capable of handling entities other than "persons", renaming the unanalyzed "persons" with something we're struggling to analyze empirically, "moral patients", doesn't seem like useful progress.

Recall that Mikhail's hope was for a "computational analysis that renders… [moral agent and moral patient] distinct from one another and specifics their standard range of application" (Mikahil 2011, 302). If, as our evidence suggests, the received view cannot serve this function, we need to start considering alternatives. That is, if Agency and Experience aren't driving moral status judgments, what could be?

7: Relationships Matter to Moral Status

The received view has a kind of philosophical merit: if it is true, it partially justifies folk morality about moral status. That is, a good portion of the philosophical literature agrees that the kind of mind an entity has is the place to pin a justified theory of moral status. A compelling version of arguments of this kind run as follows. We care about ourselves and demand moral consideration from others. We care about something particular about ourselves, namely our subjective experiences. Our mental capacities mark out what exactly it is coherent to care about; namely the things we have the capacity to feel positively or negatively about. So basic pain and pleasure fit easily here. And this kind of an account gives philosophers who sometimes demand that all animals be treated equally (Singer, All Animals Are Equal 1974), the capacity to say that nevertheless we don't need to worry about the voting rights of pigs and clams. Those animals do not have the mental capacities necessary to feel deprived (i.e., be harmed) by the lack of voting rights.

The major alternative I'll consider does have philosophical proponents, but it also seems to embody personal biases in an objectionable way. In this section I point to some

of the evidence that we make moral status judgments based on the social relationships we have with an entity or entity type. The idea here is that the difference in treatment between a bear and a dog, an android and a man, a child and an adult, and opposite sex antagonists, is predicated on the different social roles.

Take as a toy example, the difference in obligations we might have to our pet cat, versus a stray cat that lives around our house. In fact, this is an acute example; in contemporary American culture pet animals like dogs and cats are given legal protections. You cannot abuse your pet animals, you cannot deprive them of food, water, etc. But you're under no legal obligation to feed and house the stray, just because it has just as sophisticated a mind as your own cat. Your pet cat here will count as one of the relevant social relationship types: a ward. A ward relationship involves a powerful entity engaging in an argument (implicit or explicit) to care for a weaker entity. This relationship type also encompasses our relationship to children, biological or adopted.

The work in this tradition is disparate and far ranging. It runs from Tooby and Cosmides (Tooby and Cosmides 2010) arguments that ingroup outgroup competition is the evolutionary reason for all moral psychology. And to (Earp, et al. 2021) efforts to show that their relational norms model is effective in predicting sophisticated moral judgments. Now while we might think it is normatively unjustifiable, some philosophers have argued that relationship types should govern our obligations to nonhuman animals (Donaldson and Kymlicka 2011).

Yet why should we think it is morally justifiable to prefer our own children over someone else's; to prefer our own cat to someone else's; to prefer our mother to a stranger. Moral impartiality, the veil of ignorance, the view from nowhere, moral universalizability: we give these things up to some significant extent of we choose a social-relational view of moral status. I will return to this issue in Chapter Five, where we will deal more directly with the consequences of violating the veil of ignorance in the ways we've done throughout this dissertation. For now, we can of course get on with things: this is after all centrally a descriptive and explanatory project.

8. Changing Obligations or Changing Moral Status

Harming other animals and a deontological mind

> Consider the following (too minimal) position about the treatment of animals. So that we can easily refer to it, let us label this position "utilitarianism for animals, Kantianism for people." It says: (1) maximize the total happiness of all living beings; (2) place stringent side constraints on what one may do to human beings. Human beings may not be used or sacrificed for the benefit of others; animals may be used or sacrificed for the benefit of other people or animals only if those benefits are greater than the loss inflicted. […] Following Orwell, we might summarize this view as: all animals are equal but some are more equal than others.
>
> (Nozick 1974, 38)

The quote here from Nozick above should illustrate a kind of conceptual complication we should spend some time untangling. Any moral theory involves both rules and entities the rules apply to (moral status). It is conceptually possible that different sets of moral rules apply to different moral statuses. In some ways this is trivially true because we can frame some moral rule as either binding an agent or protecting a patient: ("your right to swing your fist ends at the tip of my nose" can be paragraphed as "my right to not be harmed by the motions of your body ends at the tip of my nose").

But perhaps different kinds of rules apply to different agent/patient pairs. There's some real-world support for this, as it is the standard view of applied ethics with respect to the use of animals in research. Additionally, the example of Data from *Star Trek: The Next Generation* given in Chapter One made clear that some rules seem to apply to different kinds of moral patient status. Data is a moral agent, and because of this it is wrong to obstruct his agency. Just as keeping humans in zoos, no matter how comfortable or well provided for, is generally considered impermissible, but keeping animals in them is not. This kind of feature of our moral psychology cannot be ignored if we want to make some progress on a descriptively adequate account of the psychology of moral status.

So far, I have avoided saying "this is the deontological view of moral status" or "that is the utilitarian view of moral status." I'm not sure any one view about moral status is well characterized by either association. And it is useful, at least at first, to consider moral status as separable from philosophical or psychological accounts of moral rules.

However, Nozick's suggestion here stands out for its clarity. Let us consider three conceptual possible alternatives

1) Different deontic rules apply to different entity types/moral statuses
2) The same deontic rules apply to all entities/moral statuses
3) The same deontic rules apply to all entities, weighted by type/moral status

The first possibility here is a generalized version of Nozick's suggestion. It is compatible with "deontology for humans and utilitarianism for animals" as well as the far more unlikely "utilitarianism for humans and deontology for animals". Of course, the "types" could still be other than "humans" and "nonhuman animals". We might prefer "persons" talk to "humans", in order to not build in an assumption that our psychology is speciesist (even if that assumption is plausible, it is an open empirical question at the heart of this project).

The second possibility is certainly implausible. It predicts that our moral psychology should be triggered by rocks falling down a mountain. A softer version would have to draw a line around entities that counted, entities with moral status of some sort. Whatever that line is, entities within it would have "full moral status" as it is sometimes called in the philosophical literature. If this characterized our moral psychology, humans would treat any entity that got any moral consideration at all with that equal to what they give to other humans. Implausible again.

The third possibility, that the same deontic rules are operative for any entity that minimally counts as having moral status, but that the rules are weighted for entities that don't have "full moral status".

These three alternatives obscure the distinction between moral agent/patient. It's possible that the same deontic rules apply to anything that minimally qualifies as a moral agent. But that for partial moral patients' different moral rules are applied. Additionally, we must consider the distinction between full and partial moral status with respect to the agent/patient distinction more carefully. Does moral agency come in full and partial forms? Does patiency? Does this mean that moral agency and patiency are represented as continuous variables (you can be 3/5ths of a moral patient, for example) or are they represented as categorial variables. We might reframe the alternatives above with these considerations in mind.

4) Different deontic rules apply to entities that qualify as both moral agents and moral patients, as opposed to those that qualify as plain patients.
5) The same deontic rules apply to moral agents that are also patients, as to those entities that are plain patients
6) Moral agency is mentally represented as a categorial variable
7) Moral agency is mentally represented as a scale variable
8) Moral patiency is mentally represented as a categorial variable
9) Moral patiency is mentally represented as a scale variable

As we start framing these hypotheses about how moral status might work in the mind, the picture can become quite complicated. For example, 4 and 5 above represent the fact that we don't yet know if in order to be a moral agent, you must also be a moral patient. i.e., that there are no plain moral agents. Kant of course suggested that there are no plain moral agents or patients. That in order to be morally considerable at all you had to be a verbally articulate member of the community of rational ends. And Bentham argued instead that animals were surely plain moral patients, on account of their capacity to suffer. Exploring these views makes it evident that it is important to discover if moral agent/patient status comes in degrees or kinds. And furthermore, how different degrees/kinds of moral agent/patient status trigger different rules, or merely discount the same rules.

Yet much of my experimental work was unable to tease these issues apart. Moral status at least doesn't seem to be additive. And if it comes in degrees those degrees are not closely linked to perceived agency or experience. However, my subjects did not take anything like a strict utilitarian view for all cryptominds. The complex elements of moral cognition still seem operative when cryptominds are in play.

More over postulating two different sets of moral rules, that are pegged to two fundamentally different types of moral entities (humans, everything else) seems unparsimonious at best. Yet we must continue to be curious about these options. Again, if I've done useful work so far I think it is in illustrating how easy it is to not be curious about important implications of our psychological theories.

9. What we can be Justified in Saying about the Psychology of Moral Status.

After all this it is easy to get a bit lost in the details. It seems like we both have a lot of experimental data relevant to psychological theories of moral status, however none that supports a unified picture about what moral status is. We started out with some significant confidence that mind perception was crucially involved in judgments of moral status. However, this seems to not be descriptively accurate about folk judgments. We even presented as plausible the idea that juvenile courts were a function of this psychology. That minors were viewed as less agentive and therefore less worthy of punishment. And while my experiments do show they are rated as less agentive, and as less intentional, they also show that people largely think they should be punished just as much as an adult acting the same way (CH3, 5.2.1).

We know from my experiments and others I've referenced that mind perception does seem to have detectable effects on judgments. Though those effects are small, and somewhat contradictory.

In frustrating but perhaps classical form for a philosopher, I suggest some epistemic humility. It seems that we are quite a bit further away from a workable theory of the psychology of moral status than some argue and any of us would like. In the last chapter I briefly sketch out further work that might make some inroads. However, I cannot back away from what I take to be the most important moral implied by my data: the best extant theories are not well supported by the evidence when tested directly, by their own lights.

CHAPTER 5

SUMMARY AND CONCLUDING REMARKS

1. Summary

We began this project by asking how humans make moral status attributions: who counts morally, and for what reasons. I narrowed that down further to ask how humans intuitively make these judgments. Or, stated within the language of the computational representational research tradition in psychology, what are the mental mechanisms that generate human intuitions about moral status (both agent, and patient). I also promised that understanding how the moral mind worked, particularly with respect to moral status, would enrich our understanding of ourselves, and maybe also have some advice as to what we should expect from ourselves and others morally. Yet as we explored the extant theories from psychology in Chapter Two, and my own testing of some of their assumptions in Chapter Three and Four. We can see we're in a difficult situation with respect to clarity about yourselves with respect to this topic.

To remind us, over the course of the past four chapters, I attempted to explain what moral status is, and maybe more directly what it isn't. In Chapter One I pointed to the kinds of behaviors and judgments this concept picks out. I argued that a first look at the relevant legal and social history show a preoccupation with the minds of the entities involved. Did the statue intend to kill the man (Ch1. 3.1.3)? Are the piglets mature enough to be responsible for their action (Ch1. 3.1.4)? Is abortion murder (CH1. 3.2.1)? Do androids deserve self-determination (CH1 3.2.4)?

Real world moral disagreements often revolve and resolve around the kinds of minds/mental states the entities involved have. Can the agent knowingly intend its actions? Does the patient really feel pain when injured? Starting out our investigation, it seemed hard to deny that these must be the relevant criteria on which to make at least some moral status judgements. At least, appealing to that framework easily explains very clear and distinct judgements – like that kicking a rock and kicking a child have different moral values.

That framework (that moral status equals other mind perception, the received/orthodox view) also aligns with explicitly recorded judgments from the law and history. For example, in positive law[55] Thomas Aquinas objections to the trail of animals was on the grounds that they didn't have the correct capacities for understanding the world to be morally culpable. They were incapable of reason, and so their actions were amoral (Scott and Coester 2015). For Aquinas, and others before and since, the capacity to "reason" was thought to be human specific. And it was also thought to be good enough grounds on which to base moral status judgments.

Kant also required that an entity have the capacity to reason in order to have moral status – i.e., to have mental capacities sophisticated enough to be part of the community of rational ends. But I argued in Chapter One that "reason"/"conscience" here are a black box. Boxes out of which come rich and complexly structured moral intuitions. The conscience, after all, is the thing to be explained here. And moral reasoning, the kind

[55] It's important to remember that positive law and other codified legal codes embody reflective human judgments, and so may sometimes poorly characterize folk psychology. Much of this was explored in chapter one.

attributed to the conscience, involves quite distinct forms of logic.[56] So in this we identified the thing to be explained by a mature moral psychology. What kind(s) of logic is(are) used by the mind when processing morally salient events. And particularly, what kind of logic is used as inclusion/ exclusion criteria for moral relevance: who/what gets to be included in the sphere of our moral concern and for what reasons.

I also presented some examples of how different positive legal systems developed specifically to deal with entities of various moral status (animals, children, and sometimes objects). This was suggestive evidence that moral status judgments came in degrees, that it picked out identifiable individuals rather than classes or groups, and again that it was in some ways related to the kinds of minds of the entities involved. By Chapter Four, we saw that these assumptions are unsupported by my evidence and argument.

Again in Chapter Two, I outlined how the orthodox view in philosophy (moral status judgments should appeal to mental capacities of an entity for justification) became the received view in moral psychology (other mind perception, along the two dimensions Agency and Experience), provides the essential and maybe only inputs into moral status attributions. According to the received view (CH2 2-2.1) we value an entity in the world as a moral agent in so far as we perceive them as being capable of Agency; we value an entity in the world as a moral patient in so far as we perceive them as being capable of experience. This view I attributed largely to the psychologists Wenger and Gray. In their major work, (Wegner and Gray, The Mind Club 2016) they argue that the perception of minds in other entities isn't something we learn or are taught. Instead, it is an innate part

[56] The logic of promise keeping for example seems to violate some norms of logic.

of core cognition (Spelke and Kinzler, Core Knowledge 2007). And that the mechanism is triggered by a range of subtle cues such as, how fast an entity moves (feeding agency). Wenger and Gray report on experiments with subjects where the motion of a plant was sped up. As the speed increases, subjects are more likely to attribute agency to the plant. The faster the sunflower turns towards the sun, the more people will think it is because it wants the light[57].

Wegner and Gray argue that their account of other mind perception also explains moral status judgments. This received view of moral status was the target of most of my criticism. And my experimental work reported on in Chapter Three was largely an attempt to disconfirm this theory of moral status.

In Chapter Two I also outlined a somewhat different approach to moral psychology: moral grammar as proposed by John Mikhail (Ch2 3-3.2). Mikhail has few if any written views about how the moral grammar interacts with/utilizes the concept of moral status. However, as I showed, his moral postulates depend on some simplifying assumptions about who counts and how much. And as soon as we start to examine those

[57] This means that like judgments of Agency and Experience, moral status judgments (agent/patient) are also not idiosyncratic parts of culture. It isn't that moral status judgments shouldn't be predicted to vary culturally. It is just that they shouldn't be varying culturally any more than other mind perception does. And since other mind perception is part of core cognition, part of the very psychological equipment humans use to learn about other minds at all, then the variation here shouldn't be dramatic.

This means that like judgments of Agency and Experience, moral status judgments (agent/patient) are also not idiosyncratic parts of culture. It isn't that moral status judgments shouldn't be predicted to vary culturally. It is just that they shouldn't be varying culturally any more than other mind perception does. And since other mind perception is part of core cognition, part of the very psychological equipment humans use to learn about other minds at all, then the variation here shouldn't be dramatic.

simplifying assumptions, we can see how deep of a problem moral status is for proponents of the moral grammar hypothesis.

While Mikhail attempted to avoid complications of moral status, his modes imply that significant aspects of both moral agency and patientcy are part of the narrow faculty of morality. That is, part of a domain specialized computational mechanism. As outlined by Mikhail, his postulates represent a way to characterize trade offs between entities with different types of moral status (degrees or kinds) and different types of harms (degrees or kinds). Yet in the driverless car experiments reported on in Chapter Three, our moral psychology seems not particularly good at adding up the moral value of moral patients. Testing Mikhail's postulates further is intended in future work.

And the end of Chapter Two we had some exposure to two different accounts of moral psychology, and how they both struggle to account for and integrate moral status into their models. Left with these problems we turned to an experimental program I developed largely with the help of my experimental collaborator, Dr Barlev.

When we moved to the third chapter, concerning my experimental program, my arguments depended on the idea that we could evaluate the criteria for moral status by looking at how and when entities are exculpated for prototypically immoral actions (e.g., arbitrary interpersonal violent killing). Both Gray's constructivist moral psychology, and Mikhail's moral grammar, take interpersonal harms to be prototypical cases that any good theory of moral psychology should be able to account for (Ch2, end of section 2). Yet when I tested Gray's hypothesis, I found almost no real support for it (CH3. 2.4, 4.3, 5.3, 6.3). The difference in treatment (judgments of moral status) between humans and

nonhuman entities is certainly not explicable by a difference in mind perception (as measured in my study, and I measured several ways Ch3. 2.1.2.2; CH3. 6.2.2.3).

Mikhail's k-generation versus i-generation is deeply embedded in his moral grammar. Roughly, to say something is k-generated is to say that the agent knows that some event will happen as a consequence of their actions. To say something is i-genereated is to say that the agent intends that some event will happen as a consequence of their actions. So, for Mikhail moral agents have to be (perceived as) the kinds of entities that can know, but not intend some consequences. As well as the kind that can know and intend some consequences. If an action was undertaken with the intention of causing a harm to a moral patient (with the intention of committing a moral violation) it should be judged more harshly than if the same action and caused the same harm to the moral patient but as an unintended side effect of some other morally permissible action. This should make the "in order to" question, a measure of intent proposed by Mikhail and his colleagues, particularly good at predicting if some entity is even potentially a moral agent (an entity capable of both k-generation and i-generation). It should have also made the time delay variations of my vignettes more likely to be judged harshly. However, here too the evidence was thin (CH3. 4.3). Subjects were sometimes willing to say that an agent murder a patient, but didn't intend to kill them (CH3. 3.2-3.5). Now this could be because I didn't give subjects an alternative to murder, like manslaughter. But the time delay variations didn't dramatically (or often significantly) change mean results on the true false questions I asked to assess moral judgments.

In Chapter Four I went into some depth about the empirical and conceptual problems plaguing the received view (CH4. 1-3). We also showed evidence that the folk concept of murder is very different from the positive law concept (CH4. 5). We covered the failings of our mind perception measures, and considered some alternatives cues to moral status attributions like relationship type (CH4, 6). And we saw again and again how the received view simplifies things in a not so helpful way. For example, we left open the question of if we use different moral rule sets all together for different types of entities with moral status: (utilitarianism for animals, deontology for people, as an example) (CH4. 7).

In the background of all of this was a suggestion of mine, that moral status might be a different kind of thing with respect to judgments of rightness and wrongness of an action in some circumstances, than it is with respect to judgments of appropriate punishments for some bad action in some circumstances. While there was evidence for this here and there, sometimes it was contradictory. For example, in Chapter Three section 5, we saw that agent children who murdered a patient were viewed as doing something wrong deserving of punishment. Almost as much as an adult agent was. The dramatic difference was in the "in order to" judgments[58].

2. Descriptive Psychology and Moral Education:

[58] We also saw that issues like this really does pose a problem for views like Mikhail's. A large feature of his model of moral cognition (outlined in Chapter two) is tracking the intended effects of an agent. If an agent is viewed as not intending the bad action, the agent should be viewed as less wrong, less worthy of punishment as well.

After all this we are left with a somewhat uncomfortable situation. The major philosophical and psychological traditions, the major legal systems, seem to implicate the kinds of minds as being the grounds upon which we can rationally make moral status judgments. Yet when tested rather directly, and as I argue with methods that should be perfectly acceptable within these research traditions themselves (CH4, 1), I find that a person's judgments about the minds of an entity had very little to do with if they also judged the entity to be a possible moral agent and/or patient.

Both descriptively, and normatively, we have some difficulties here. Descriptively because of issues I've pointed to, both empirical and conceptual. But also, normatively. Let me take a moment to explain why.

One major takeaway from my experimental research is that folk use of the word murder, as a word to refer to a kind of (defensibly?) wrongful killing, is vastly different from its positive law counterpart. If the folk concept of murder parts ways substantially with the concept that is socially implemented in the law, then at least we face a problem about how to educate juries when they're making judgments about these matters. I don't mean to be offhanded here. We should continue to carefully investigate the folk concept of murder. Remember again that the ques in the world that lead to the relevant folk judgments here are often subtle proxies. Subtle proxies like this can be part of useful heuristics, and normally lead to reliable judgments. However, that also means they can predictably also lead to reliable errors.

But to reiterate, the folk view of murder is different than the positive law view. The folk view of moral status also seems to be different from normative views in

morality, positive law, and in the received view of moral status in psychology. And perhaps this is one major normative takeaway from the work. If my evidence and argument is to be believed, our folk intuitions about moral status seem quite distant from any normative account of it.

When are our judgments of an entity's moral status justifiable, and when are they not? For what reasons? If we take a step back a bit, Mikhail has long argued that since the rules of our moral psychology seem to be universal (Mikhail, Universal moral grammar: theory, evidence and the future 2007), and innate (Mikhail 2007), they provide a "shared yardstick" which could serve as the foundation for a theory of natural law (Mahlmann and Mikhail 2005). I find this view deeply appealing. In this light, moral status is exactly how we get moral disagreement. We don't disagree about the rules of morality, only who they should apply to. Which is why we should examine the operative rules of our moral psychology by simplifying out moral status problems. By lowering a kind of experimental veil of ignorance.

As Rawls points out, his methods of developing a theory of justice as fairness excludes nonhuman animals by force (Rawls 1971, 14). They are incapable of bargaining behind a veil of ignorance. Of course, Rawls didn't think morality was exhausted by a theory of justice as fairness, and insisted a fuller theory of our moral obligations would include as he said, "an account of how we are to conduct ourselves toward animals and the rest of nature." Mikhail's appeal to a kind of methodological veil of ignorance argued for largely in (Mikhail 2011), to some extent misses how central moral status judgments are to its own account of the moral calculus involved. Where Rawls can afford to ignore

attitudes to cryptominds, how we morally value them is something any theory of moral status cannot hope to avoid.

Largely philosophers argue that moral status judgments should be predicated on consideration about what we ourselves care about in our own lives (DeGrazia 2021). Unsurprisingly philosophers have largely argued that we care about our experiences, and we care about not being arbitrarily obstructed from our goals by others. If we care about these things in ourselves, the argument runs, it is hypocritical to not care about them in other entities that have similar capacities. This kind of moral realism starts with an egoistic self-appraisal, and then points out that in social life we cannot arbitrarily preference ourselves and expect to convince other (with reasons) that our arbitrary self-preference should be universalized.

In this context the question might be metaphorically expressed as, who gets to play the moral game? Well, if I get to play it anyone like me in the relevant ways, but unlike me in irrelevant ways, should be treated identically by the game. Again, a black box of relevantly alike and irrelevantly different. Philosophers have argued that the relevant similarities with respect to moral status judgments are mental capacities (Rachels, 2004, but see also Clarke, Zohny, & Savulescu, 2021). As we now know very well, if my evidence is to be believed, that doesn't seem to be a strong predictor of how people do make moral status judgments. No problem for the normative theorist on some accounts. They're attempting to outline a justifiable account, not a descriptive one. However, if moral philosophy hopes to help people get from where they are morally to some place more justifiable, moral philosophers need to have a good descriptive

understanding of the relevant psychology. If one job of a philosopher is to debug bad ideas, to help people reflect more deeply on their intuitive judgments, then we need to understand the source of those ideas and judgments. This makes at some significant portion of moral philosophy morally relevant education.

One thing highlighted by our exploration of Mikhail's postulates is that moral status judgments interact with moral rules in surprising ways. Mikhail's system rightly attempts to be able to explain comparative judgments. Which action is worse in some set of circumstances, a harm to one or a harm to five. As we saw this forced us to ask about the additive value of each moral patient. Does the mind add up a partial moral status score for each entity involved? Some of our methods hoped to shine light on this issue. The data from my driverless car experiments clearly show that there's no significant effect of number of patients in the tradeoffs to moral judgments. That is, people did not think trading off 20 chimp lives for one human life significantly worse than trading five chimp lives for one human.

Normatively we're at a bit of a loss here. We can certainly ask "why should a moral patient be less valuable in a group than they are as an individual"? but there's nothing about logic that would require an additive model is the only normatively justifiable one. It doesn't involve inherent contradiction, for example, if instead the moral value of a group is composed logarithmically out of the moral statuses of the members of the group. But while not inherently contradictory, it is seemingly arbitrary to not treat two moral patients as worth the composition of each one alone. And arbitrary treatment of moral status judgments is something a normative account just cannot tolerate.

This is significant and worthy of further study. But how the mind composes the potentially partial moral status scores of multiple entities presumes we have some idea about how the mind composes the partial moral status score of one entity. My argument here has been we don't have that. Mind perception seems less important than just being labeled as a human. But being labeled as a human isn't the only thing that matters. It matters if you're an adult or a child (whatever your species).

How do we identify a moral mistake here? Should we view dogs and chimps as having moral status equal to some proportion of human moral status in the degree that they share the mental capacities we've identified as the relevant ones? Should we value the anatomy interests of a pregnant woman more than the existent interests of a developing human? These are acute questions. We cannot even begin to do moral philosophy without having some good answer for them. But as I've shown, folk judgments don't track what philosophers have called the only game in town; the mental capacities approach to moral status. So, efforts to educate people on the actual capacities of the entities in question, the mental lives of animals (Panksepp 2011), for example, might do little to change people's judgments about the treatment of those animals. Which poses some issue for at least the rhetorical and educational hopes of moral philosophy. As I pointed out in Chapter One, concerning the moral status of developing embryos, if education of true facts is morally persuasive, we might consider that premia facies justifiable. If an embryo at some stage is not an identifiable individual, for example, and telling people this changes their minds about the permissibility of abortion, then we've made some moral progress through morally relevant education. However, my evidence

suggests that educating people about the actual mental capacities of an embryo will have little effect on their moral status judgments concerning it. Though more research is needed here.

3. Theory and Incuriosity

As I've said before, I think a major contribution of this dissertation is to point out how simplifying assumptions can sometimes obscure interesting empirical questions. I tested some implications of the received view. And explored some data that has implications for any theory of moral cognition. One question my data bares on is if moral status comes in commensurable degrees, or incommensurable kinds. While entities with cryptominds were treated very differently than humans, they were not discounted totally. People do have complex and convergent views about the moral status of nonhuman animals and the like. And in many cases it seems that the rules that govern tradeoffs between harms to cryptominds are the same as those that govern the tradeoffs between harms to humans. However much more work needs to be done to establish any kind of scale of moral agency or patiency.

Again, I think the existing theories were structured in such a way as to make researchers somewhat incurious about if moral agency and patiency are established by radically different kinds of inputs. Again, this is an open area for empirical research.

Researchers have also not paid enough attention to if moral status is treated differently when making judgments about rightness and wrongness, justified punishment. We saw some evidence that punishment judgments were more effected by mind

perception than wrongness judgments, section 3.5, Chapter Three. But we also saw that deeply contradicted by our experiments involving child agents and patients.

4. Summary remarks:

Moral status is central to any account of moral psychology and moral philosophy. It is often neglected in both. To their great credit, psychologists and philosophers have attempted to remedy this over the last two decades considerably. My hope here was to take the existing theories seriously. To see what they implied descriptively but also at times normatively. And to put those implications to the test. In the course of this I think we've learned a few lessons. First, moral status probably isn't what we think it is. It is probably a complicated judgment, not predicated on some small set of simple inputs. Filled with in group biases, social relational considerations, and here and there a care for the agency and experience of the entities involved. We have to do a much better job interrogating our descriptive theories of moral status, which depend on their intuitive plausibility, and the results of experiments often designed to provide evidential support for the hypothesis.

Second, when we take moral status seriously, and when we take existing psychological theories seriously, we can productively generate empirical hypothesis. I tested a few of my own, and have waved a hand in the direction of others. I hope also to have shown that removing and interdigitating useful simplifying assumptions here is also productive.

Third, epistemic humility is called for here. It is possible to find lots of things that have small effects on moral status judgments. But whatever the essence of moral status is psychologically, that is not known. We do not know the basics. What the input conditions for moral status are. Is moral agency calculated in the same way as moral patiency? Do different moral rules apply to different kinds of agents and patients? Is our moral grammar innumerate or does it fastidiously calculate the moral value of individuals and groups? It might be unsurprising at the end to have a call for epistemic humility coming from a philosopher. Often philosophy is useful when highlighting how little is actually known about a rich and important topic.

REFERENCES

Alfano, Mark. 2016. moral psychology an introduction. Polity Press.

Allen, Colin. 2014. "Models, Mechanisms, and Animal Minds." The Southern Jounral of Philosophy 52: 75-97.

Beth Levin, Malka Rappaport Hovav. 2005. Argument Realization. Cambridge University Press.

Bettencourt, Ann, Nancy Dorr, K Charlton, and David L Hume. 2001. "Status differences and in-group bias: a meta-analytic examination of the effects of status stability, status legitimacy, and group permeability." Psychological Bulletin 520-542.

Bloom, Paul. 2016. Against Empathy. Ecco.

Bryner, Nicholas. 2018. "Colombian Supreme Court Recognizes Rights of the Amazon River Ecosystem." International Union for Conservation of Nature. April 20. https://www.iucn.org/news/world-commission-environmental-law/201804/colombian-supreme-court-recognizes-rights-amazon-river-ecosystem.

Chambers, Robert. 1869. The Book of Days: a Miscellany of Popular Antiquities in Connection with the Calendar, Including Anecdote, Biography, and History, Curriosities of Literature and Oddities of Human Life and Character. London: W.&R. Chambers.

Chomsky, Noam. 1975. Reflections On Language. New York, NY: Pantheon Books.

Churchland, Patricia. 2011. Braintrust: what neuroscience tells us about morality. Princeton, NJ: Princeton University Press.

Cohen, David. 2012. Free Will Hunting . Directed by Raymie Muzquiz. Futurama.

Cohen, Jacob. 1988. Statistical Power Analyisis for the Behavioral Sciences. Routledge.

Cudworth, Ralph. 1731. A Treatise Concerning Eternal and Immutable Morality. James and John Knapton.

Cushman, Fiery, and Joshua Greene. 2011. "Finding faults: How moral dilemmas illuminate cognitive structure." Social Neuroscience 1-11.

Dahl, Fredrik Andreas, and Gry Oftedal. 2018. "Trolley Dilemmas Fail to Predict Ethical Judgment in a Hypothetical Vaccination Contexgt." Journal of Empirical Reseach on Human Reseach Ethics.

Daly, Martin, and Margo Wilson. 1988. Homicide. Aldine de Gruyter, Inc.

Darley, John, and Daniel Batson. 1973. ""From Jerusalem to Jericho": A study of situational and dispositional variables in helping behavior." Journal of Personality and Social Psychology 100-108.

DeGrazia, David. 2021. "An Interest-Based Model of Moral Status." In Rethinking Moral Status, by Steve Clarke, Hazem Zohny and Julian Savulescu, 40-56. Oxford University Press.

Dennett, Daniel. 2013. Thinking, Intuition Pumps and Other Tools for. W.W. Norton & Company.

Dillon, Kyle D., and Fiery Cushman. 2012. "Agent, Patient … ACTION! What the Dyadic Model Misses." Psychological Inquiry 23: 150-154.

DiSilvestro, Russell. 2010. Human Capacities and Moral Status. Springer.

Donaldson, Sue, and Will Kymlicka. 2011. Zoopolis, A Political Theory of Animal Rights. New York, NY: Oxford University Press.

Dwyer, Susan. 1999. “Moral Competence.” In Philosophy and Linguistics, by Kumiko Murasugi and Robert Stainton, 169-190. Boulder, CO: Westview Press.

Dwyer, Susan, Bryce Huebner, and and Marc Hauser. 2010. "The Linguistic Analogy: Motivations, Results, and Speculations ." Topics in Cognitive Science 2: 486-510.

Dwyer, Susan, Bryce Huebner, and Marc Hauser. 2010. "The Linguistic Analogy: Motivations, Results, and Spectulations." TOPICS in Cognitive Science 486-510.

Eagleman, David. 2011. "The Brain on Trail." The Atlantic, July/August.

Earp, Brain, Killian McLoughlin, Joshua Monrad, Margaret Clark, and Molly Crockett. 2021. "How social relationships shape mral wrongness judgments." Nature Communications.

Encyclopedia Britannica. 1911. "Deodand." Edited by Hugh Chisholm. Cambridge University Press.

Evans, Edward Payson. 1906. The Criminal Prosecution and Capital Punishment of Animals. (Public Domaine), retrieved from The Internet Archive.

Evensky, Jerry. 2005. Adam Smith's Moral Philosophy. Cambridge University Press.

Fillmore, Charles. 1970. "The grammar of hitting and breaking ." In Readings in English tranformational grammar, by R.A. Jacobs and P.A. Rosenbaum, 120-133. Waltham.

Finnis, John. 1980. Natural Law & Natural Rights. New York, NY: Oxford University Press.

Flanagan, Owen, Hagop Sarkissian, and David Wong. 2008. Naturalizing Ethics. Vol. 1, in Moral Osychology, by Walter Sinnott-Armstrong, 1-27. Cambridge, MA: MIT Press.

Fodor, Jerry. 1983. The Mondularity of Mind. Cambridge, MA: MIT Press.

Francione, Gary. 1995. Animals, Property, and the law. Temple University Press.

Francione, Gary, and Robert Garner. 2010. The Animal Rights Debate: Abolition or Regulation? Columbia University Press.

Frankfurt, Harry G. 1971. Freedom of the Will and the Concept of a Person. Vol. 68. 1 vols. New York, NY: Journal of Philosophy, Inc.

Friedman, Ori, and Alan Leslie. 2007. "The conceptual underpinnings of pretense: Pretending is not 'behaving-as-if'." Cognition.

Gallistell, C. Randy, and Adam Philip King. 2009. Memory and the Computational Brain: Why Cognitive Science Will Transform Neuroscience. Chichester, West Sussex: John Wiley & Sons, Ltd.

Gelman, Rochel, and C.R. Gallistel. 1986. The Child's Understanding of Number. Harvard University Press.

Gray, Heather, Kurt Gray, and Daniel Wegner. 2007. "Dimensions of Mind Perception." Science 619.

Gray, Kurt, Adam Waytz, and Liane Young. 2012. "The moral dyad: A fundamental template unifying moral judgment." Psychological Inquiry 23: 2006-215.

Gray, Kurt, and Daniel Wegner. 2009. "Moral typecasting: divergent perceptions of moral agents and moral patients." Journal of Personality and Social Psychology 93: 505-520.

Gray, Kurt, Chelsea Schein, and Adrian Ward. 2014. "The Myth of Harmless Wrongs in Moral Cognition: Automatic Dyadic Completion From Sin to Suffering." Journal of Experimental Psychology: General 1600–1615.

Gray, Kurt, Jennifer K MacCormack, Teague Henry, Emmie Banks, Chelsea Schein, Emma Armstrong-Carter, Samantha Abrams, and Keely Muscatell. 2022. "The affective harm account (AHA) of moral judgment: Reconciling cognition and affect, dyadic morality and disgust, harm and purity." Journal of Personality and Social Psychology.

Gray, Kurt, Liane Young, and Adam Waytz. 2012. "Mind Perception Is the Essence of Morality." Psychological Inquiry 101-124.

Guarecuco, Lyanne. 2017. "Lawmaker: Criminalizing Abortion Would Force Women to be 'More Personally Responsible'." TexasObserver.org. January 23. Accessed July 2020. https://www.texasobserver.org/texas-lawmaker-no-abortion-access-would-force-women-to-be-more-personally-responsible-with-sex/.

Guéguen, Nicolas. 2012. "The Sweet Smell of … Implicit Helping: Effects of Pleasant Ambient Fragrance on Spontaneous Help in Shopping Malls." The Jounral of Social Psychology.

Hamlin, J. Kiley. 2013. "Moral Judgement and Action in Preverbal Infants and Toddlers: Evidence for an Innate Moral Core." Current Directions in Psychological Science 22 (3): 186-193.

Hamlin, Kiley, Karen Wynn, and Paul Bloom. 2007. "Social evaluation by preverbal infants." Nature 557-559.

Hansen, Ronadl, and Dennis Wagner. 2012. "Burned bodies: notes found; police convinced case is a murder-suicide." The Arizona Republic, Jun 7.

Hardin, Russell. 2007. David Hume: moral and political theorist. Oxford University Presss.

Hatfield, Gary. 1998. "The Cognitive Faculties." In The Cambridge History of Seventeenth-Century Philsophy, by Daniel Garber and Michael Ayers. Cambridge University Press.

Hauser, Marc. 2006. Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong. Harper Collins.

Hauser, Marc, Noam Chomsky, and W. Tecumseh Fitch. 2002. "The Faculty of Language: What Is It, Who Has It, and How Did It Evolve?" Science 298: 1569-1579.

Hempel, Carl, and Paul Oppenheim. 1948. "Studies in the Logic of Explanation." Philosophy of Science 135-175.

Hrynowski, Zach. 2019. "What Percentage of Americans Are Vegetarian?" Gallup. September 27. https://news.gallup.com/poll/267074/percentage-americans-vegetarian.aspx.

Hudson, W.D. 1967. Ethical Intuitionism. New York, NY: St Martin's Press.

Hutcheson, Francis. 1728/2013. Illustrations on the Moral Sense. Harvard University Press.

Jackendoff, Ray. 1992. Languages of the mind, essays on mental representation. The MIT Press.

—. 1972. Semantic Interpretation in Generative Grammar. MIT Press.

—. 1997. The Architecture of the Lnaguage Faculty. Cambridge, Massachusetts: The MIT Press.

Jackendoff, Ray, and Steven Pinker. 2005. "The nature of the language faculty and its implications for evolution of language." Cognition 97: 211-225.

Jones, John T, Brett W Pelham, Mauricio Carvallo, and Matthew C Mirenberg. 2004. " How Do I Love Thee? Let Me Count the Js: Implicit Egotism and Interpersonal Attraction." Journal of Personality and Social Psychology 665-683.

Joseph Henrich, Steven Heine, Ara Norenzayan. 2010. "The weirdest people in the world?" Behavioral and Brain Sciences.

Knobe, Joshua. 2011. "Finding the Mind in the Body ." In Future Science: Essays from the Cutting Edge, by Max Brockman, 211-224. Vintage Books.

Korsgaard, Christine. 2004. "Fellow Creatures: Kantian Ethics and Our Duties to Animals." Tanner Lectures on Human Values 24: 77-110.

Lacey, Marc. 2011. "Identical Twins, One Charged in a Fatal Shooting, Create Confusion for the Police." The New York Times, August 8.

Langlois, Judith, Lori Roggman, Rita Casey, Jean Ritter, and Loretta Rieser-Danner. 1987. "Infant preferences for attractive faces: Rudiments of a stereotype?" Developmental Psychology 363-369.

Leslie, Alan. 1987. "Pretense and Representation: The Origins of "Theory of Mind"." Psychological Review 412-426.

Levin, Beth. 1993. English Verb Classes and Alternations. Chicago, IL: University of Chicago Press.

Levin, Beth, and Malka Rappaport Hovav. 2005. Argument Realization. Cambridge University Press.

Levine, Sydney, Alan Leslie, and John Mikhail. 2018. The mental representaiton of human action . Cognitive Science.

Machery, Edouard. 2017. Philosophy Within Its Proper Bounds. Oxford University Press.

Mahlmann, Matthias, and John Mikhail. 2005. "Cognitive Science, Ethics and Law." In Epistomology and Ontology: IVR-Symposium Lund, 95-102. Germany: Franz Steiner Verlag Wiesbaden GbmH.

Maienschein, Jane. 2014. Embryos Under The Microscope, the diverging meanings of life. Harvard University Press.

Maienschein, Jane. 2016. "Embryos, microscopes, and society." Studies in History and Philosophy of Biological and Biomedical Sciences 129-136.

Marcus, Gary. 2008. Kludge: The Haphazard Construction of the Human Mind. Boston: Houghton Mifflin Company.

—. 2015. "Face It, Your Brain Is a Computer." The New York Times, June 27.

Maricopa County Attorney's Office. 2011. MaricopaCountyAttorney.org. August 19. https://www.maricopacountyattorney.org/CivicAlerts.aspx?AID=84.

McFee, Graham. 2014. Free Will. Durham.

McKie, Robin. 2013. "Jared Diamond: what we can learn from tribal life." The Guardian, Jan 5.

McShane, Katie. 2007. "Why Environmental Ethics Shouldn’t Give Up on Intrinsic Value." Environmental Ethics.

Mele, Alfred. 2019. "Free Will and Moral Responsibility: Manipulation, Luck, and Agents' Histories." Midwest Studies in Philosophy.

Mikhail, John. 2011. Elements of Moral Cognition. New York, NY: Cambridge University Press.

Mikhail, John. 2010. "Is the Prohibition of Homicide Universal? Evidence from Comparative Criminal Law." Georgetown Public Law and Legal Theory Research Paper No. 10-21.

Mikhail, John. 2007. "Moral Cognition and Computational Theory." In Moral Psychology , by Walter Sinnott-Armstrong. Cambridge, Mass: The MIT Press.

—. 2000. Rawls' linguistic analogy: A study of the 'generative grammar' model of moral theory described by John Rawls in 'A Theory of Justice'. Ithaca, NY: PhD Dissertation, Cornell University.

Mikhail, John. 2007. "Universal moral grammar: theory, evidence and the future." Trends in cognitive sciences 143-152.

Nadja Richter, Bernard Tiddeman, Daniel Haun. 2016. "Social Preference in Preschoolers: Effects of Morphological Self-Similarity and Familiarity." PLoS ONE 1-11.

Nagel, Thomas. 1997. The Last Word. Oxford: Oxford University Press.

—. 1986. The View From Nowhere. Oxford: Oxford University Press.

Nevins-Saunders, Elizabeth. 2012. "Not Guilty as Charged: The Myth of Mens Rea for Defendants with Mental Retardation." UC Davis Law Review.

Nozick, Robert. 1974. Anarchy, State, and Utopia. Basic Books, Inc.

Panksepp, Jaak. 2011. " The basic emotional circuits of mammalian brains: do animals have affective lives?" Neuroscience Biobehavioral Review 1791-804.

Piccoli, Sean. 2020. "Family of Slain Barnard Student Criticizes Sentence for 14-Year-Old." The New York Times, June 15.

Pinker, Steven. 1997. How the Mind Works. New York, NY: Penguin Group.

—. 1994. The Language Instinct. William Morrow and Company.

Planet Money. 2020. "Your Life Is Worth $10 Million, According To The Government." National Public Radio, July 17.

Prinz, Jesse. 2008. "Is Morality Innate?" In Moral Psychology, Vol 1: The Evolution of Morality: Adaptations and Innateness, by Walter Sinnott-Armstrong, 367-406. The MIT Press.

Prinz, Jesse. 2006. "The Emotional Basis of Moral Judgments." Philosophical Explorations.

Pylyshyn, Zenon. 1984. Computation and Cogniton. Boston: The MIT Press.

Quote Investigator. 2011. "Your Liberty To Swing Your Fist Ends Just Where My Nose Begins." Quote Investigator. October 15. https://quoteinvestigator.com/2011/10/15/liberty-fist-nose/.

Rachels, James. 2004. "Drawing Lines." In Animal Rights, Current Debates and New Directions, by Cass Sunstein and Martha Nussbaum, 162-174. New York, NY: Oxford University Press.

Ravitz, Jessica. 2016. "The sacred land at the center of the Dakjota pipeline dispute." CNN.COM. Nov 1. Accessed June 20, 2020. https://www.cnn.com/2016/11/01/us/standing-rock-sioux-sacred-land-dakota-pipeline/index.html.

Rawls, John. 1971. A Theory of Justice. Cambridge, Mass: Harvard University Press.

Regan, Tom. 1989. "The Case for Animal Rights." In Animal Rights and Human Obligations, by Tom Regan and Peter Singer, 19-25. Englewood Cliffs, NJ: Prentice Hall.

Reinhart, Tanya. 2002. "The Theta System - An Overview." Theoretical Linguistics 28 (3): 229-290.

Reynolds, Tania, Chuck Howard, Hallgeir Sjåstad, Luke Zhu, Tyler G.Okimoto, Roy F. Baumeister, Karl Aquino, and JongHan Kim. 2020. Man up and take it: Gender bias in moral typecasting. Organizational Behavior and Human Decision Processes.

Rhode, Deborah. 2010. The beauty biase: the injustice of appearance in life and law.

Roberts, LaVonne. 2020. "No 14-year-old should be tried as an adult." Columbia Daily Spectator, February 23.

Russell, Bertrand. 1910. Elements of Ethics (1910). Public Domain. http://fair-use.org/bertrand-russell/the-elements-of-ethics.

Schein, Chelsea, and Kurt Gray. 2017. "Chelsea Schein1 and Kurt Gray." Personality and Social Psychology Review 1-39.

Schein, Chelsea, and Kurt Gray. 2017. "The Theory of Dyadic Morality: Reinventing Moral Judgment by Redefining Harm." Personality and Social Psychology Review.

Scott, Callum David, and Yolandi Marié Coester. 2015. "Rewriting Aquinas' animal ethics: the primacy of reason in the determination of moral status." South African Journal of Philosophy.

Sealey, Raphael. 1983. "The Homicide Courts of Ancient Athens." Classical Philology 275-296.

Singer, Peter. 1974. "All Animals Are Equal." Philosophic Exchange.

—. 1975. Animal Liberation. Harper Collins.

Singer, Peter. 2009. "Speciesism and Moral Status." Metaphilosophy.

Snodgrass, Melinda M. 1989. "The Measure of a Man." Star Trek, The Next Generation. February 13.

Spelke, Elizabeth. 1990. "Principles of Object Perception." Cogntive Science.

Spelke, Elizabeth, and Katherine Kinzler. 2007. "Core Knowledge." Developmental Science 10 (1): 89-96.

Speltini, Giuseppina, and Stefano Passini. 2014. "Cleanliness/dirtiness, purity/impurity as social and psychological issues." Culture & Psychology.

Sripada, Chandra Sekhar. 2008. "Nativism and Moral Psychology: Three Models of the Innate Structure That Shapes the Contents of Moral Norms." In Moral Psychology, Vol 1: The Evolution of Morality: Adaptations and Innateness, by Walter Sinnott-Armstrong, 319-345. The MIT Press.

Strickland, Brent, Matthew Fisher, and Joshua Knobe. 2012. "Moral structure falls out of general event structure." Psychological Inquiry 23 (2): 198-205.

Sykes, Katie. 2011. "Human Drama, Animal Trials: What the Medieval Animal Trials Can Teach Us about Justice for Animals." Animal Law 273.312.

Sytsma, Justin, and Edouard Machery. 2012. "Two sources of moral standing." Review of Philosophy and Psychology 303-324.

Taylor, Donald, and Janet Doria. 1981. "Self-serving and group-serving bias in attribution." The Journal of Social Psychology 201-211.

The Editorial Board. 2020. "Even 14-Year-Olds Who Kill Are Not Adults." The New York Times, February 22.

The Office of Juvenile Justice and Delinquency Prevention. 1999. Juvenile Justice: A Century of Change. https://www.ncjrs.gov/html/ojjdp/9912_2/contents.html, U.S. Department of Justice.

Tinderholt, Tony. 2017. "Abolition of Abortion in Texas Act H.B. No. 948." Texas: capitol.state.tx.us.

Tooby, John, and Leda Cosmides. 2010. "Groups in Mind: The Coalitional Roots of War and Morality." In Human Morality and Sociality: Evolutionary and Comparative Perspectives, by Henrik Høgh-Olesen, 91-234. Palgrave-MacMillan.

Van Gelderen, Elly. 2018. The Diachrony of Verb Meaning: Aspect and Argument Structure. Routledge.

Vihvelin, Kadri. 2018. "Arguments for Incompatibilism." The Stanford Encyclopedia of Philosophy . Edited by Edward N. Zalta.

Warren, Marry. 1997. Moral Status: Obligations to Persons and Other Living Things. Oxford: Oxford University Press.

Wegner, Daniel, and Kurt Gray. 2016. The Mind Club: Who Thinks, What Feels, and Why It Matters. Penguin Books.

Whitman, Walt. 1977. Leaves of Grass. Norwalk, CT: The Easton Press.

Woodruff, Michael. 2020. "The face of the. fish." aeon.co. July 3. Accessed July 3, 2020. https://aeon.co/essays/fish-are-nothing-like-us-except-that-they-are-sentient-beings.

Wynne, Clive. 2019. "Dog is Love: why and how your dog loves you." Houghton Mifflin Harcourt.

APPENDIX A

DRIVERLESS CAR TROLLY PROBLEM, INFERENTIAL STATISTICS

1. Human Pedestrian, Baseline

A Spearman's Correlation test indicated that there was a significant effect of Experience on change lane judgments, r(1313)= .07, p= .011.

2. Chimps Baseline and Exploratory

A Spearman's Correlation test indicated that there was a significant effect of Experience on change lane judgments, r(438)= .11, p= .02, in the 10 chimps condition.

3. Dogs Baseline and Exploratory

A Spearman's correlation test indicated that there was a significant positive effect of Agency as measured and switch lane judgments, r(409)= .122, p =.013, in the 5 dogs condition.

Likewise, a Spearman's test indicated that there was a significant positive correlation between both Experience, r(447)= .132, p= .005, and Agency, r(447)= .148, p= .002, in the 20 dogs condition.

APPENDIX B

MURDERING AND MURDERABLE CRYPTOMINDS, MEANS COMPARISONS.

TIME DELAY CONDITION AND MORAL JUDGMENTS

Android Patient:

A independent t test indicated that more subjects judged the Man's killing of the Android to be wrong in the TDC (m= .81, SD .390) than in the basic condition (m= .70, SD .457), t(656)= -3.294, p= .001. Cohen's d = .259, a small effect.

Android Agent:

An independent t test indicated that fewer subjects judged the Android's killing the Man to be a murder in the TDC (m= .81, SD .392) than the basic condition (m= .88, SD .33), t(677)= 2.336, p= .02. Cohen's d, -.194, a small effect.

Bear Patient

An independent t test indicated that more subjects judged the Man's killing of the Bear to be intentional in the TDC (m= .96, SD .199) than in the basic condition (m= .91, SD .288), t(679) = -2.642, p= .008. Cohen's d= .202, a small effect.

An independent t test indicated that more subjects judged the Man's killing of the Bear to be wrong in the TDC (m= .71, SD .453) than in the basic condition (m= .61, SD .487), t(679) = -2.715, p= .007. Cohen's d= .213, a small effect.

An independent t test indicated that more subjects judged the Man's killing of the Bear to be deserving of punishment in the TDC (m= .62, SD .486) than in the basic condition (m= .53, SD .5), t(679) = -2.442, p= .015. Cohen's d= .183, a small effect.

Dog Agent

An independent t test indicate that fewer subjects judged the Dog to have killed the man in the TDC (m= .97, SD .172) than the basic condition (m= 1, SD .055), t(659)= 2.774, p= .006. Cohen's d = -.236, a small effect.

Tree Patient

An independent t test indicated that more subjects judged the Man to have killed the Tree intentionally in the TDC (m= .62, SD .486) than the basic condition (m= .5, SD .501), t(649)= -3.053, p= .002. Cohen's d= .243, a small effect.

An independent t test indicated that more subjects judged the Man's killing of the Tree to have been wrong in the TDC (m= .55, SD .499) than the basic condition (m= .4, SD .49), t(649)= -3.847, p= .000. Cohen's d= .303, a small effect.

An independent t test indicated that more subjects judged the Man's killing of the Tree to have been deserving of punishment in the TDC (m= .39, SD .489) than the basic condition (m= ..24, SD .428), t(649)= -3.847, p= .000. Cohen's d = .243, a small effect.

Tree Agent

An independent t test indicated no significant differences in true false moral judgments between the TDC and basic conditions for the tree as an agent.

River Patient

An independent t test indicated that more subjects judged the River's being filled-in as deserving of punishment in the TDC (m= .67, Sd .47) than the basic condition (m= .59, SD .493), t(660) = -2.228 , p= .026 . Cohen's d =.166, a small effect.

APPENDIX C

MURDERING AND MURDERABLE CRYPTOMINDS, MEANS COMPARISONS, THEMATIC ROLE AND MORAL JUDGMENTS

Man:

A Mann-Whitney test indicated that Experience measures were smaller for thematic agents (Mdn =100) than for thematic patients (Mdn = 100), U=7842802.5, p = .022. Eta (Z= -3.321, n= 1337) ≅ .008. A negligibly small effect.

Android:

An independent t test indicated that Agency measures were greater for thematic agents (M = 44.43, SD 32.57) than patients (m = 40.11, SD 33.24), t(1335)= -2.408, p = .016. Cohen's d = .132, a very small effect.

There was no significant effect of thematic role on Experience measures

Bear:

An independent t test indicated that Agency measures were smaller for the thematic agent (m= 47.82, SD 29.57) than thematic patient (m=57.27, SD 29.57), t(1336)=5.849, p> .001. Cohen's d = -.320, a small effect.

An independent t test indicated that Experience measures were smaller for the agent (m= 66.38, SD 27.72) than patient (m= 71.1, SD 26.7), t(1336)= -3.19, p= .002. Cohen's d = .-175., a small effect.

Dog:

An independent t test indicated that Agency measures were smaller for the thematic agent (m= 52.4, SD 28.85) than for the patient (m= 56.71, SD 29.61), t(1663)=2.53, p= .007. Cohen's d = -.147. A small effect.

An independent t test indicated that Experience measures were Smaller for the thematic agent (m= 66.19, SD 28.18) than patients (m= 69.54, SD 28.25), t(1332)= 2.164, P= .031. Cohen's d = -.119. A small effect.

Adult, Baby Chimp

There was no significant difference between mind perception of either chimp in either agent or patient thematic roles.

Tree:

A Mann-Whitney test indicated that Agency measures were smaller for thematic agents (Mdn =0) than for thematic patients (Mdn = 1), U=202545, p = .002. Eta (Z= -3.133, n= 1337) ≅ .007. A negligibly small effect.

A Mann-Whitney test indicated that Experience measures were smaller for thematic agents (Mdn =1) than for thematic patients (Mdn = 1), U=200411, p = .001. Eta (Z= -3.321, n= 1337) ≅ .007. A negligibly small effect.

River:

A Mann-Whitney test indicated that Agency measures were smaller for thematic agents (Mdn =0) than for thematic patients (Mdn = 0), U=199022.5, p> .001. Eta (Z= -3.847, n= 1336) ≅ .011. A very small effect.

A Mann-Whitney test indicated that Experience measures were smaller for thematic agents (Mdn =0) than for thematic patients (Mdn = 1), U=203003.5, p = .001. Eta (Z= -3.214, n= 1336) ≅ .007. A negligibly small effect.

APPENDIX D

MURDERING AND MURDERABLE CRYPTOMINDS, TIME DELAY CONDITION AND MIND PERCEPTION

Man

A Mann-Whitney test indicated no significant differences between mind perception in the basic and the time delay conditions.

Android

An independent t test indicated that Experience measures were greater in the TDC (m= 27.37, SD 31.27) than is the standard condition (m= 23.55, SD 29.11), t(1334)= -2.347, p= .019. Cohen's d = .128, a very small effect.

Bear:

An independent t test indicated no significant difference between mind perception in the basic condition and TDC.

Dog

An independent t test indicated that Experience measures were greater in the TDC (m= 69.84, SD 27.75) than the standard condition (m= 65.96, SD 28.63), t(1663)= -2.515, p= .012. Cohen's d = .138, a very small effect.

Tree and River:

A Mann-Whittney test indicated no significant difference between mind perception for Tree or River in either thematic role.

Adult/Baby Chimp:

There was no TDC for either of these entities

APPENDIX E

MURDERING AND MURDERABLE CRYPTOMINDS. INFERENTIAL STATISTICS. CORRELATIONS MIND PERCEPTION ANDTHEMATIC ROLES

Android:

A two tailed Spearman's correlation test indicated here were several significant positive correlations between mind perception measures (Agency, Experience) for the patient {xx} and moral judgments {Kill, Murder, In Order To, Wrong, and Punish): (Kill: Spearman's rho for Experience r(656)= .248, p> .001. Agency r(656)= .214, p> .001; Murder: Spearman's Experience r(656)= .289, p> .001. Agency r(656)= .233, p> .001; In Order To: Experience r(656)= .200, p> .001. Agency r(656)= .159, p> .001; Wrong: Experience r(656)= .132, p= .001. Agency r(656)= .175, p> .001; Punish: Experience r(656)= .132, p= .001. Agency r(656)= .175, p> .001.)

Bear:

A two tailed Spearman's correlation test indicated there were several significant positive correlations between mind perception measures (Agency, Experience) for the patient {Bear} and moral judgments {Kill, Murder, In Order To, Wrong, and Punish): (Murder: Spearman's rho for Experience r(679)= .166, p> .001. Agency r(679)= .216, p> .001; In Order To: Experience r(679)= .093, p= .015; Wrong: Experience r(679)= .170, p> .001. Agency r(679)= .191, p> .001; Punish: Experience r(679)= .218, p> .001. Agency r(679)= .238, p> .001.)

Dog:

A two tailed Spearman's correlation test indicated there were four significant positive correlations between mind perception measures (Agency, Experience) for the patient {Dog} and moral judgments {Kill, Murder, In Order To, Wrong, and Punish): (Murder: Spearman's rho for Experience r(669)= .114 p> .001. Agency r(669)= .138, p= .003; In Order To: Experience r(669)= ..018, p= .035 . Agency r(669)= .099, p= .01.)

Adult Chimp:

A two tailed Spearman's correlation test indicated there were three significant positive correlations between mind perception measures (Agency, Experience) for the patient {Adult Chimp} and moral judgments {Kill, Murder, In Order To, Wrong, and Punish): (Kill: Spearman's rho for Experience r(333)= .125, p> .000. Agency r(333)= .129, p= .018; Punish: Agency r(333)= .132, p= .016.)

Baby Chimp:

A two tailed Spearman's correlation test indicated there were three significant positive correlations between mind perception measures (Agency, Experience) for the patient {Baby Chimp} and moral judgments {Kill, Murder, In Order To, Wrong, and Punish): (Murder: Spearman's rho for Agency r(334)= .138, p= .011; Punish: Experience r(334)= .189, p= .001. Agency r(334)= .146, p= .007.)

Tree:

A two tailed Spearman's correlation test indicated there were several significant positive correlations between mind perception measures (Agency, Experience) for the patient {Tree} and moral judgments {Kill, Murder, In Order To, Wrong, and Punish): (Murder: Spearman's rho for Experience r(648)= .204, p> .001. Agency r(649)= .215, p> .001; In Order To: Experience r(648)= .157, p> .001. Agency r(649)= .166, p> .001; Wrong: Experience r(648)= .201, p> .001. Agency r(649)= .187, p> .001; Punish: Experience r(648)= .206, p> .001. Agency r(649)= .231, p> .001.)

River:

A two tailed Spearman's correlation test indicated there were several significant positive correlations between mind perception measures (Agency, Experience) for the patient {River} and moral judgments {Kill, Murder, In Order To, Wrong, and Punish): (Kill: Spearman's rho for Experience r(661)= .120, p= .002. Agency r(661)= .106, p= .006; Murder: Experience r(661)= .224, p> .001. Agency r(661)= ..211, p> .001; In Order To: Experience r(661)= .119, p= .002. Agency r(661)= .111, p= .004.)

3.5.2 Agent Cryptominds

Android:

A two tailed Spearman's correlation test indicated there were several significant positive correlations between mind perception measures (Agency, Experience) for the Agent {Android} and moral judgments {Kill, Murder, In Order To, Wrong, and Punish): (Murder: Spearman's rho for Experience r(676)= .164, p> .001. Agency r(677)= .195, p>

.001; In Order To: Agency r(676)= .126, p= .001; Wrong: Experience r(675)= .128, p= .001. Agency r(676)= .137, p> .001; Punish: Experience r(676)= .229, p> .001. Agency r(677)= .289, p> .001.)

Bear

A two tailed Spearman's correlation test indicated there were no significant positive correlations between mind perception measures (Agency, Experience) for the agent {Bear} and moral judgments {Kill, Murder, In Order To, Wrong, and Punish):

Dog

A two tailed Spearman's correlation test indicated there were several significant positive correlations between mind perception measures (Agency, Experience) for the agent {Dog} and moral judgments {Kill, Murder, In Order To, Wrong, and Punish): (Murder: Spearman's rho for Experience r(658)= .084, p= .03. Agency r(658)= .167, p> .001; In Order To: Agency r(657)= .138, p> .001; Wrong: Agency r(659)= .133, p= .001; Punish: Experience r(658)= .105, p= .007. Agency r(658)= .141, p > .001,.)

Adult Chimp

A two tailed Spearman's correlation test indicated there were xx significant positive correlations between mind perception measures (Agency, Experience) for the agent {Adult Chimp} and moral judgments {Kill, Murder, In Order To, Wrong, and Punish): (Murder: Spearman's rho for Agency r(355)= .133, p= .012; In Order To: Agency

r(354)= .123, p= .02.; Punish: Experience r(354)= .149, p= .005. Agency r(354)= .192, p> .001.)

Baby Chimp

A two tailed Spearman's correlation test indicated there were three significant positive correlations between mind perception measures (Agency, Experience) for the agent {Baby Chimp} and moral judgments {Kill, Murder, In Order To, Wrong, and Punish): (Murder: Spearman's rho Agency r(307)= .121, p= .033; In Order To: Experience r(307)= .115, p= .044; Punish: Agency r(307)= .123, p= .031.)

Tree:

A two tailed Spearman's correlation test indicated there were several significant positive correlations between mind perception measures (Agency, Experience) for the agent {Tree} and moral judgments {Kill, Murder, In Order To, Wrong, and Punish): (Murder: Spearman's rho for Experience r(684)= .144, p> .001. Agency r(656)= .193 p> .001; In Order To: Experience r(684)= .131, p= .001. Agency r(684)= .212, p> .001.; Wrong: Experience r(684)= .207, p> .001. Agency r(684)= .291, p> .001.; Punish: Experience r(684)= .237, p> .001. Agency r(684)= .291, p> .001.)

River

A two tailed Spearman's correlation test indicated there were xx significant positive correlations between mind perception measures (Agency, Experience) for the agent

{River} and moral judgments {Kill, Murder, In Order To, Wrong, and Punish): (Murder: Spearman's rho for Experience r(669)= .232, p> .001. Agency r(669)= .233, p> .001; In Order To: Experience r(670)= .251, p> .001. Agency r(670)= .279, p> .001; Wrong: Experience r(671)= .237, p> .001. Agency r(671)= .235, p> .001; Punish: Experience r(670)= .294, P> .001. Agency r(670)= .310, P> .001.)

APPENDIX F

MURDERING AND MURDERABLE CRYPTOMINDS. REGRESSION ANALYSIS. MINDS AND MORAL JUDGMENTS, PER ENTITY PER THEMATIC ROLE

Android Patient

Kill: Binary logistic regression indicated that android Agency and Experience both were significantly related to if subjects thought an Android could be killed. ($\chi^2(2)= 36.92$, $p> .001$). The Cox & Snell R^2 (= .055) can be interpreted as indicating that 5.5% of the variance in Kill judgments can be predicted by a model that includes both these independent variables.

Murder: Binary logistic regression indicated that android Agency was significantly related to if subjects thought an Android could be murdered. ($\chi^2(1)= 58.37$, $p> .001$). The Cox & Snell R^2 (= .085) can be interpreted as indicating that 8.5% of the variance in Murder judgments can be predicted by a model that includes Agency as an independent variable.

In Order To: Binary logistic regression indicated that android Experience was significantly related to if subjects thought the human attacked the Android in order to kill it. ($\chi^2(1)= 20.34$, $p> .001$). The Cox & Snell R^2 (= .03) can be interpreted as indicating that 3% of the variance in In Order To judgments can be predicted by a model that includes Experience as an independent variable.

Wrong: Binary logistic regression indicated that android Agency was significantly related to if subjects thought it was wrong to kill an android. ($\chi^2(1)= 17.99$, $p> .001$). The Cox & Snell R^2 (= .027) can be interpreted as indicating that 2.7% of the variance in wrong judgments can be predicted by a model that includes Agency as an independent variable.

Punish: Binary logistic regression indicated that android Agency and Experience both were significantly related to if subjects thought the man should be punished for killing the android. ($\chi^2(2)= 32.15$, $p> .001$). The Cox & Snell R^2 (= .048) can be interpreted as indicating that 4.8% of the variance in Punish judgments can be predicted by a model that includes both these independent variables.

Android Agent

Kill:0

Murder: Binary logistic regression indicated that android Agency was significantly related to if subjects thought an Android could murder a man. ($\chi^2(1)= 26.2$, $p> .001$). The Cox & Snell R^2 (= .039) can be interpreted as indicating that 3.9% of the variance in Murder judgments can be predicted by a model that includes Agency as an independent variable.

In Order To: Binary logistic regression indicated that android Agency was significantly related to if subjects thought the android attacked the human in order to kill it. ($\chi^2(1)= 11.02$, $p= .001$). The Cox & Snell R^2 (= .016) can be interpreted as indicating that 1.6% of the variance in In Order To judgments can be predicted by a model that includes Agency as an independent variable.

Wrong: Binary logistic regression indicated that android Experience was significantly related to if subjects thought it was wrong to kill the human. ($\chi^2(1)= 10.73$, $p= .001$). The Cox & Snell R^2 (= .016) can be interpreted as indicating that

1.6% of the variance in Wrong judgments can be predicted by a model that includes Agency as an independent variable.

Punish: Binary logistic regression indicated that android Experience was significantly related to if subjects thought the android should be punished for killing the man. ($\chi^2(1)= 57.98.15$, $p> .001$). The Cox & Snell R^2 (= .0.82) can be interpreted as indicating that 8.2% of the variance in Punish judgments can be predicted by a model that includes both these independent variables.

Bear Patient:

Kill: 0

Murder: Binary logistic regression indicated that bear Agency was significantly related to if subjects thought a bear could be murdered. ($\chi^2(1)= 32.73$, $p> .001$). The Cox & Snell R^2 (= .047) can be interpreted as indicating that 4.7% of the variance in Murder judgments can be predicted by a model that includes Agency as an independent variable.

In Order To: Binary logistic regression indicated that bear Experience was significantly related to if subjects thought the human attacked the bear in order to kill it. ($\chi^2(1)= 5.42$, $p= .02$). The Cox & Snell R^2 (= .008) can be interpreted as indicating that 0.8% of the variance in In Order To judgments can be predicted by a model that includes Experience as an independent variable.

Wrong: Binary logistic regression indicated that bear Experience and Agency both were significantly related to if subjects thought it was wrong to kill

the bear. ($\chi^2(2)= 28.47$, p> .001). The Cox & Snell R^2 (= .041) can be interpreted as indicating that 4.1% of the variance in Wrong judgments can be predicted by a model that includes Agency and Experience as independent variables.

Punish: Binary logistic regression indicated that bear Agency and Experience both were significantly related to if subjects thought the man should be punished for killing the bear. ($\chi^2(2)= 45.21$, p> .001). The Cox & Snell R^2 (= .064) can be interpreted as indicating that 6.4% of the variance in Punish judgments can be predicted by a model that includes both these independent variables.

Bear Agent

Kill: =0

Murder: 0

In Order To: 0

Wrong: 0

Punish: 0

Dog Patient

Kill 0

Murder: Binary logistic regression indicated that dog Agency was significantly related to if subjects thought a dog could be murdered. ($\chi^2(1)= 12.6$, p> .001). The Cox & Snell R^2 (= .019) can be interpreted as indicating that 1.9%

of the variance in Murder judgments can be predicted by a model that includes Agency as an independent variable.

In Order To: Binary logistic regression indicated that the dog Experience was significantly related to if subjects thought the human attacked the dog in order to kill it. ($\chi^2(1)= 5.42$, $p= .02$). The Cox & Snell R^2 (= .008) can be interpreted as indicating that 0.8% of the variance in In Order To judgments can be predicted by a model that includes Experience as an independent variable.

Wrong: 0

Punish: Binary logistic regression indicated that dog Experience was significantly related to if subjects thought the man should be punished for killing the dog. ($\chi^2(2)= 5.48$, $p= .019$). The Cox & Snell R^2 (= .008) can be interpreted as indicating that 0.8% of the variance in Punish judgments can be predicted by a model that includes Experience as an independent variable.

Dog Agent:

Kill: 0

Murder: Binary logistic regression indicated that dog Agency was significantly related to if subjects thought a dog could murder a man. ($\chi^2(1)= 18.64$, $p> .001$). The Cox & Snell R^2 (= .028) can be interpreted as indicating that 2.8% of the variance in Murder judgments can be predicted by a model that includes Agency as an independent variable.

In Order To: Binary logistic regression indicated that dog Agency was significantly related to if subjects thought the dog attacked the human in order to kill it. ($\chi^2(1)= 13.2$, $p> .001$). The Cox & Snell R^2 (= .02) can be interpreted as indicating that 2% of the variance in In Order To judgments can be predicted by a model that includes Agency as an independent variable.

Wrong: Binary logistic regression indicated that dog Agency was significantly related to if subjects thought it was wrong to kill the human. ($\chi^2(1)= 11.78$, $p= .001$). The Cox & Snell R^2 (= .018) can be interpreted as indicating that 1.8% of the variance in Wrong judgments can be predicted by a model that includes Agency as an independent variable.

Punish: Binary logistic regression indicated that dog Agency was significantly related to if subjects thought the dog should be punished for killing the man. ($\chi^2(1)= 13.13$, $p> .001$). The Cox & Snell R^2 (= .02) can be interpreted as indicating that 2% of the variance in Punish judgments can be predicted by a model that includes both these independent variables.

Adult Chimp Patient

Kill: Binary logistic regression indicated that adult chimp Agency was significantly related to if subjects thought it could be killed. ($\chi^2(1)= 6.21$, $p= .013$). The Cox & Snell R^2 (= .018) can be interpreted as indicating that 1.8% of the variance in Kill judgments can be predicted by a model that includes Agency as an independent variable.

Murder: 0

In Order To: 0

Wrong: 0

Punish: Binary logistic regression indicated that adult chimp Agency was significantly related to if subjects thought the man should be punished for killing the adult chimp. ($\chi^2(1)= 5.48$, $p= .019$). The Cox & Snell R^2 (= .016) can be interpreted as indicating that 1.6% of the variance in Punish judgments can be predicted by a model that includes Agency as an independent variable.

Adult Chimp Agent:

Kill: 0

Murder: Binary logistic regression indicated that adult chimp Agency was significantly related to if subjects thought a adult chimp could murder a man. ($\chi^2(1)= 5.91$, $p= .015$). The Cox & Snell R^2 (= .016) can be interpreted as indicating that 1.6% of the variance in Murder judgments can be predicted by a model that includes Agency as an independent variable.

In Order To: Binary logistic regression indicated that adult chimp Agency was significantly related to if subjects thought the adult chimp attacked the human in order to kill it. ($\chi^2(1)= 4.97$, $p= .026$). The Cox & Snell R^2 (= .014) can be interpreted as indicating that 1.4% of the variance in In Order To judgments can be predicted by a model that includes Agency as an independent variable.

Wrong: 0

Punish: Binary logistic regression indicated that adult chimp Agency was significantly related to if subjects thought the adult chimp should be punished for killing the man. ($\chi^2(1)= 13.39$, $p> .001$). The Cox & Snell R^2 (= .037) can be interpreted as indicating that 3.7% of the variance in Punish judgments can be predicted by a model that includes Agency as an independent variable.

Baby Chimp Patient:

Kill:0

Murder: Binary logistic regression indicated that baby chimp Agency was significantly related to if subjects thought a baby chimp could be murdered. ($\chi^2(1)= 6.34$, $p= .012$). The Cox & Snell R^2 (= .019) can be interpreted as indicating that 1.9% of the variance in Murder judgments can be predicted by a model that includes Agency as an independent variable.

In Order To: 0

Wrong: 0

Punish: Binary logistic regression indicated that baby chimp Experience was significantly related to if subjects thought the man should be punished for killing the baby chimp. ($\chi^2(1)= 12.95$, $p> .001$). The Cox & Snell R^2 (= .038) can be interpreted as indicating that 3.8% of the variance in Punish judgments can be predicted by a model that includes Experience as an independent variable.

Baby Chimp Agent

Kill:0

Murder: Binary logistic regression indicated that baby chimp Agency was significantly related to if subjects thought a baby chimp could murder a man. ($\chi^2(1)= 4.55$, $p= .033$). The Cox & Snell R^2 (= .015) can be interpreted as indicating that 1.5% of the variance in Murder judgments can be predicted by a model that includes Agency as an independent variable.

In Order To: 0

Wrong: 0

Punish: Binary logistic regression indicated that baby chimp Agency was significantly related to if subjects thought the baby chimp should be punished for killing the man. ($\chi^2(1)= 4.58$ $p= .032$). The Cox & Snell R^2 (= .015) can be interpreted as indicating that 1.5% of the variance in Punish judgments can be predicted by a model that includes Agency as an independent variable.

Tree Patient

Kill:0

Murder: Binary logistic regression indicated that tree Agency was significantly related to if subjects thought a tree could be murdered. ($\chi^2(1)= 29.58$, $p> .001$). The Cox & Snell R^2 (= .044) can be interpreted as indicating that 4.4% of the variance in Murder judgments can be predicted by a model that includes Agency as an independent variable.

In Order To: Binary logistic regression indicated that the tree Agency was significantly related to if subjects thought the human attacked the tree in order to kill it. ($\chi^2(1)= 22.5$, $p> .001$). The Cox & Snell R^2 (= .034) can be interpreted as indicating that 3.4% of the variance in In Order To judgments can be predicted by a model that includes Agency as an independent variable.

Wrong: Binary logistic regression indicated that the tree Agency was significantly related to if subjects thought the human was wrong to kill the tree. ($\chi^2(1)=30.59$, $p> .001$). The Cox & Snell R^2 (= .046) can be interpreted as indicating that 4.6% of the variance in Wrong judgments can be predicted by a model that includes Agency as an independent variable.

Punish: Binary logistic regression indicated that the tree Agency was significantly related to if subjects thought the human should be punished for killing the tree. ($\chi^2(1)= 52.56$, $p> .001$). The Cox & Snell R^2 (= .078) can be interpreted as indicating that 7.8% of the variance in Punish judgments can be predicted by a model that includes Agency as an independent variable.

Tree Agent

Kill:0

Murder: binary logistic regression indicated that tree Agency was significantly related to if subjects thought a tree could murder a man. ($\chi^2(1)= 49.79$, $p> .001$). The Cox & Snell R^2 (= .070) can be interpreted as indicating that

7% of the variance in Murder judgments can be predicted by a model that includes Agency as an independent variable.

In Order To: binary logistic regression indicated that tree Agency was significantly related to if subjects thought the tree fell on the man in order to kill him. ($\chi^2(1)= 40.38$, $p> .001$). The Cox & Snell R^2 (= .057) can be interpreted as indicating that 5.7% of the variance in In Order To judgments can be predicted by a model that includes Agency as an independent variable.

Wrong: binary logistic regression indicated that tree Agency was significantly related to if subjects thought the tree fell was wrong to kill the man. ($\chi^2(1)= 56.54$, $p> .001$). The Cox & Snell R^2 (= .079) can be interpreted as indicating that 7.9% of the variance in Wrong judgments can be predicted by a model that includes Agency as an independent variable.

Punish: binary logistic regression indicated that tree Agency was significantly related to if subjects thought the tree should be punished for killing the man. ($\chi^2(1)= 60.45$, $p> .001$). The Cox & Snell R^2 (= .084) can be interpreted as indicating that 8.4% of the variance in In Punish can be predicted by a model that includes Agency as an independent variable.

River Patient

Kill:0

Murder: binary logistic regression indicated that river Experience was significantly related to if subjects thought a river could be murdered by a man.

($\chi^2(1)= 36.48$, p> .001). The Cox & Snell R^2 (= .054) can be interpreted as indicating that 5.4% of the variance in Murder judgments can be predicted by a model that includes Experience as an independent variable.

In Order To: Binary logistic regression indicated that river Experience was significantly related to if subjects thought the human attacked the river in order to kill it. ($\chi^2(1)= 10.35$, p= .001). The Cox & Snell R^2 (= .016) can be interpreted as indicating that 1.6% of the variance in In Order To judgments can be predicted by a model that includes Experience as an independent variable.

Wrong: 0

Punish: Binary logistic regression indicated that river Agency was significantly related to if subjects thought the human should be punished for filling in the river. ($\chi^2(1)= 4.56$, p= .033). The Cox & Snell R^2 (= .007) can be interpreted as indicating that .7% of the variance in Punish judgments can be predicted by a model that includes Agency as an independent variable.

River Agent

Kill:0

Murder: binary logistic regression indicated that river Agency and Experience were significantly related to if subjects thought a river could murder a man. ($\chi^2(1)= 43.91$, p> .001). The Cox & Snell R^2 (= .063) can be interpreted as indicating that 6.3% of the variance in Murder judgments can be predicted by a model that includes Agency and Experience as independent variables.

In Order To: binary logistic regression indicated that tree Agency was significantly related to if subjects thought the river flood in order to kill the man. ($\chi^2(1)= 54.5$, $p> .001$). The Cox & Snell R^2 (= .078) can be interpreted as indicating that 7.8% of the variance in In Order To judgments can be predicted by a model that includes Agency as an independent variable

Wrong: binary logistic regression indicated that river Agency and Experience both were significantly related to if subjects thought the river flood in order to kill the man. ($\chi^2(2)= 42.09$, $p> .001$). The Cox & Snell R^2 (= .061) can be interpreted as indicating that 6.1% of the variance in Wrong judgments can be predicted by a model that includes Agency and Experience as independent variables.

Punish: binary logistic regression indicated that river Agency and Experience both were significantly related to if subjects thought the river flood in order to kill the man. ($\chi^2(2)= 78.87$, $p> .001$). The Cox & Snell R^2 (= .111) can be interpreted as indicating that 11.1% of the variance in Wrong judgments can be predicted by a model that includes Agency and Experience as independent variables