COMPARING CAPACITIES

By

Brittany Elizabeth Shupe

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Philosophy

Written under the direction of

Elisabeth Camp

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May 2020

ABSTRACT OF THE DISSERTATION

COMPARING CAPACITIES

by BRITTANY ELIZABETH SHUPE

Dissertation Director:

Elisabeth Camp

This dissertation is comprised of three chapters pertaining to issues related to the comparison of capacities (perceptual, cognitive, and normative) across various species.

Acknowledgement

Much of this work was completed while receiving generous support from the Social Sciences and Humanities Research Council of Canada. Chapter One previously appeared in print under the title “Perceiving Deviance” in the journal Synthese, 1-13 (2019).

I am grateful beyond measure to my dissertation advisor Elisabeth Camp, as well as to committee members Kristin Andrews, Frances Egan, and Susanna Schellenberg. Special thanks are also owed to Doug Husak and Larry Temkin. For their moral support, I am indebted to Cameron Domenico Kirk-Giannini, Paul Rizza, and to my father William Shupe.

Finally, I could not have completed my dissertation were it not for the work of two men: Johnathan Mudge, the inventor of the inhaler, and Don Pedro Sánchez de Tagle y Pérez Bustamante, 2nd Marquis of Altamira, who was a trailblazer of another kind altogether.

Table of Contents

List of Illustrations

Introduction

This is a dissertation about capacities. Some capacities belong only to us: all saxophonists are human. Some capacities belong only to others: no human can regrow a limb, but an axolotl can. Finally, there are some capacities we share with other species: the capacity to recognize a pattern; the capacity to recognize a friend. When *homo sapiens* shares a capacity with another species, what does that say about us? When another species shares a capacity with *homo sapiens*, what does that say about them? Moreover, what does that say about the capacity?

This is a dissertation in philosophy, but philosophy looking outward. My central questions are philosophical, but my methodology is interdisciplinary. I ask: How should we delineate capacities and what counts as good evidence for the possession of a particular capacity? How might the possession of certain particular capacities by a non-human species constrain our broader philosophical views (in the philosophy of mind and elsewhere)? Some of my answers to these questions are philosophical, spun from pure argument; others, however, are informed by work in the sciences, most particularly psychology and neuroscience. I believe that these questions cannot be productively answered without help from the sciences, and that the best answers to them are those that will help the sciences in return. Science gives us evidence for our theories and we compensate them with conceptual clarity.

Three self-standing chapters follow this introduction, each of them about a different capacity. Chapter One is about a perceptual capacity: the capacity to represent objects

events in perceptual experience as deviating from an expectation. Chapter Two is about a rational capacity: the capacity for what psychologists call 'insightful' problem-solving. And Chapter Three is about a normative capacity: the capacity to respond to a moral reason.

In Chapter One, the chapter about a perceptual capacity, I am concerned with how we go about establishing the possession of a particular capacity. I show how one way to do so is to collaborate with the sciences, which I demonstrate by using studies in neuroscience to ground my arguments that we perceptually represent stimuli as being unexpected.

In Chapter Two, the chapter about a rational capacity, I am concerned with how we can best go about using what we know about animal capacities in order to learn about human capacities, and vice versa. I am also concerned with methodological questions in the philosophy of science: what is the ultimate project of comparative psychology and how can it best be achieved? I answer that question by proposing a framework for what I call the comparative ideal, on the basis of which I argue for some revisions to current practices.

In Chapter Three, the chapter about a normative capacity, I am concerned with the theoretical upshots of attributing a particular capacity to a non-human animal. In particular, I look at the consequences in the philosophy of law and punishment for the view that some non-human animals respond to moral reasons.

More detailed descriptions of each of these chapters follows below.

Chapter One: Perceiving Deviance

It is widely accepted that we perceptually represent basic, low-level properties like color and shape. Some philosophers have argued that these are the only properties we perceptually represent (Byrne 2009; Dretske 1995; Tye 1995). But many others argue that

we also perceptually represent certain high-level properties, such as causal relations (Siegel 2009), kind properties like being a pine tree (van Gulick 1994; Siegel 2006), affordances like being edible (Nanay 2011), and even normative properties like being wrong and being unjust (Cowan 2015; Werner forthcoming). We can call the basic claim that high-level properties are represented in perception the Rich Content View.

The Rich Content View is an existential claim about the representational contents of perception. All we need to do in order to show that the Rich Content View is true is to show that at least one high-level property is represented in perceptual experience. A clear way forward, then, is to identify high-level properties that are good candidates for being represented in perception, and to then scrutinize whether they in fact are so represented. To this end, in this chapter, I defend – on largely empirical grounds – the claim that a particular high-level property is among those that are represented in perceptual experience. My thesis, call it DEVIANCE, is this:

> DEVIANCE: We have the capacity to perceptually represent objects and events in experience as deviating from an expectation, or, for short, as deviant.

Here I mean 'expectation' in a very thin sense. An expectation needn't be consciously occurrent—it may exist simply as a background assumption or implicit prediction. The way in which we form and maintain some of these expectations might even be entirely sub-personal.

DEVIANCE stands in contrast to the idea that we only ever infer that an object or event deviates from our expectations; or, alternately, that we do sometimes recognize

deviance non-inferentially, but that the methods by which we do so are exclusively non-perceptual—perhaps the process is facilitated by affective states, for example.

My project in this chapter serves as a unique contribution to a well-worn debate. Not only do I argue in favor of the perceptual representation of a novel high-level property, but my arguments are supported by data from neuroscience that have not previously been brought to bear on the debate about perceptual contents. If my arguments are successful, then my methodology might generalize to defenses of the presence of other high-level perceptual contents.

Chapter Two: No Work for a Comparative Concept of Insight

We are said to experience insight when we suddenly and unexpectedly become aware of the solution to a problem that we previously took ourselves to be unable to solve. These 'aha' moments figure prominently in cultural narratives of creativity and scientific discovery, and there is a large psychological literature dating back to the early twentieth century devoted to understanding precisely what insight is and how it occurs. In the field of comparative cognition, there is rising interest in the question of whether non-human animals are capable of insightful problem-solving. Scientists working in this research program claim there is evidence of insight in elephants, various great apes, and several species of bird.

These claims have attracted two types of criticism from within comparative cognition. First, claims of insight in non-human animals have been criticized for conflating insight with other cognitive capacities (e.g. causal cognition, or mental trial and error). Second, it is not always clear that the relevant performances reflect anything other than

associative learning—and on the received understanding of insight within comparative cognition, insight necessarily involves non-associative processes. There are nonetheless reports of animal insight that withstand these criticisms. As a result, insight in non-human animals has gained a certain level of acceptance within the field of comparative cognition and is regarded by many as a promising object of study.

I argue, however, that one of the primary motivations for studying animal insight proves to be illusory, even if we grant that there are instances of animal problem solving that withstand the two criticisms mentioned above. As a result, these cases of purported animal insight cannot shed light on the nature of insightful problem-solving in humans. For the phenomenon studied by cognitive psychologists under the heading of insight is fundamentally different than that studied in comparative cognition.

My project is positive, since I offer the comparative psychologists a productive way forward, namely by reinterpreting the extant research on animal insight in terms of other high-level cognitive capacities which can form the basis for successfully comparative research programs. The most promising of these capacities, I suggest, is means-end reasoning.

Chapter Three: Punishing Moral Animals

There are different theories of what it takes to be a moral agent, but what most of them have historically agreed on is that, in the actual world, only humans are moral agents. Recently, though, a growing number of philosophers and scientists have broken away from that orthodox view. They claim that there are some non-human animals who should also be regarded as genuine moral agents.

I argue that if moral animals do exist, they are burdened with several cognitive shortcomings such that they are likely to be severely limited in their ability to morally evaluate one another, if they are even capable of moral evaluation at all. I contend that the severity of these deficits is such that impartial human observers would generally be more competent to see to matters of animal desert than the moral animals themselves, even accounting for the pragmatic challenges in our doing so.

Building on those arguments, I make the case that if there are moral animals, then, contrary to intuition, both retributivists and deterrence theorists about punishment ought to recognize a strong reason to punish animal wrongdoers and perhaps even a duty to do so. For the believer in animal morality, this is a worry; for the skeptic, a potential reductio. Although I present some ways this conclusion can be avoided by the proponent of animal moral agency, each comes at a significant theoretical cost.

Chapter One: Perceiving Deviance

Introduction

It is widely accepted that we perceptually represent basic, low-level properties like color and shape. Some philosophers have argued that these are the only properties we perceptually represent (Byrne 2009; Dretske 1995; Tye 1995). But many others argue that we also perceptually represent certain high-level properties, such as causal relations (Siegel 2009), kind properties like being a pine tree (van Gulick 1994; Siegel 2006), affordances like being edible (Nanay 2011), and even normative properties like being wrong and being unjust (Cowan 2015; Werner forthcoming). We can call the basic claim that high-level properties are represented in perception the Rich Content View.

The Rich Content View is an existential claim about the representational contents of perception. All we need to do in order to show that the Rich Content View is true is to show that at least one high-level property is represented in perceptual experience. A clear way forward, then, is to identify high-level properties that are good candidates for being represented in perception, and to then scrutinize whether they in fact are so represented. To this end, in this paper, I defend – on largely empirical grounds – the claim that a particular high-level property is among those that are represented in perceptual experience. My thesis, call it DEVIANCE, is this:

> DEVIANCE: We have the capacity to perceptually represent objects and events in experience as deviating from an expectation, or, for short, as deviant.

Here I mean 'expectation' in a very thin sense. An expectation needn't be consciously occurrent—it may exist simply as a background assumption or implicit prediction. The

way in which we form and maintain some of these expectations might even be entirely sub-personal.

DEVIANCE stands in contrast to the idea that we only ever infer that an object or event deviates from our expectations; or, alternately, that we do sometimes recognize deviance non-inferentially, but that the methods by which we do so are exclusively non-perceptual—perhaps the process is facilitated by affective states, for example.

My hope is that my project in this paper serves as a unique contribution to a well-worn debate. Not only do I argue in favor of the perceptual representation of a novel high-level property, but my arguments are supported by data from neuroscience that have not previously been brought to bear on the debate about perceptual contents. If my arguments are successful, then my methodology might generalize to defenses of the presence of other high-level perceptual contents.

In Section 1, I explain the DEVIANCE thesis at greater length and introduce a pair of contrast cases, Mistake and Control, the former of which involves the phenomenal experience of DEVIANCE and the latter of which does not. In Section 2, I show how a phenomenal contrast argument in defense of DEVIANCE using Mistake and Control might proceed. In particular, and continuing into Section 3, I discuss how evidence from neuroscience might come to bear on our inferences to the best explanation about the phenomenal difference between Mistake and Control. Section 4 discusses one candidate EEG study (Pieszek, Schröger, and Widmann, 2014) that suggests the distinctive phenomenology of Mistake is best explained by DEVIANCE. Sections 5 and 6 defend a pair of claims about the responses elicited in the study; if both of these claims are true, then the study and others like it form a solid evidence base in favor of DEVIANCE.

Section 1: Mistake and Control

Experiences of unexpected objects and events commonly share certain phenomenal features. Often such experiences elicit the sensation of surprise. Sometimes the elicitation is weaker: an unexpected stimulus might fail to palpably shock a perceiver but still evoke feelings of unease or a more general impression of ill-fit. This weaker elicitation seems to be a common factor. Although not all experiences of violated expectations are accompanied by the sensation of surprise, the more minimal sensation of ill-fit does seem ubiquitous across examples (both when surprise phenomenology is present and when it is absent). Because I take it to be more fundamental in this way, it is the sensation of ill-fit that I seek to isolate and explain as the characteristic feature of experience of deviance.[1]

More specifically, I will focus on the experience of having one's auditory expectations violated. Since DEVIANCE being true of one modality is enough to make DEVIANCE true simpliciter, we can defensibly restrict our focus in this way. My exemplar case, Mistake, is one in which the final note of a musical sequence sounds 'wrong' to the listener. I contrast Mistake with Control, a case in which the same sequence of notes is heard but, lacking context, the listener does not experience the final note in the sequence

[1] A number of ethicists have proposed fittingness as a basic normative property (Mandelbaum 1955; Chappell 2012; Audi 2013; Audi 2015; McHugh and Way 2016). Some of these philosophers have further conceived of it as an apprehended property, and have argued that felt moral demands have their basis in the phenomenology of apprehending the fittingness or ill-fittingness of an act (Mandelbaum 1955; Audi 2013; Audi 2015). There are commonalities between these latter philosophers and myself. Although I am concerned with more basic, non-moral violations (such as misplayed notes), I, too, think that the detection of such violations is characterized by a phenomenology of ill-fit.

as sounding wrong. After introducing these cases, I will explore the argument that what elicits the phenomenology of ill-fit in one case but not the other is a difference in what is perceptually represented. Here are the cases:

> Control: Jenna is listening to the sound of a neighbor's piano lesson. She hears a musical triplet being played (A-B-C). After a pause, she hears another slightly different triplet (A- B-D).
>
> Mistake: Jenna is listening to the sound of a neighbor's piano lesson. This time, before the music starts, she hears the teacher tell her neighbor to repeat the notes that she plays. Then, Jenna hears a musical triplet being played (A-B-C). After a pause, she hears another slightly different triplet (A-B-D).

Jenna perceives all of the same low-level auditory content in Control and Mistake. However, as will be familiar to the reader, in cases like Mistake, but not Control, we tend to experience the misplayed note (D) as sounding wrong. Or, in other words, we experience the sensation of ill-fit described previously. The phenomenology is the same when we listen to a singer whose voice cracks. That the note sounds wrong in one but not both of the cases above is taken to be an intuitive explanandum. DEVIANCE can be evaluated partly on the basis of whether it is a serviceable explanans, i.e. whether it is the thesis that best explains the phenomenal difference in the pair of contrast cases.

Section 2: The Phenomenal Contrast Argument

By far the most popular strategy for defending positive theses about the contents of perception has been the phenomenal contrast method (Siegel 2007). Once a candidate

thesis like DEVIANCE is proposed, the strategy begins with the identification of two cases like Mistake and Control, i.e. two experiences that involve the representation of all of the same low-level perceptual content, but which have different overall phenomenologies. Then, the cases are used to perform an inference to the best explanation: we ask, does the addition of some particular high-level perceptual content best explain the distinctive phenomenology of Mistake?

To illustrate, a phenomenal contrast argument for DEVIANCE that proceeds from Mistake and Control would go roughly as follows.

1. Control and Mistake differ in overall phenomenology.

2. This change in overall phenomenology is best explained as a change in the phenomenology of the auditory experience that is part of the overall experience.

3. This change in the phenomenology of the auditory experience is best explained as a change in the representational content of the auditory experience.

4. On the best explanation of (3), the change in representational content is that in Mistake, but not Control, Jenna auditorily represents that the D-note is deviant.

To expand on this, the explanation on offer is that, when you have no beliefs or other expectations about which notes an incoming triplet will contain, as in Control, then when you hear the triplet, auditory experience only represents the notes the triplet actually contains (i.e., the stimulus' low- level properties). But once you form a prior belief or expectation about which notes an incoming triplet will contain, as in Mistake, auditory experience then not only represents the notes the triplet actually contains, but also

represents whether those notes were the notes that you expected to hear (a high-level property).

Remember, however, that on DEVIANCE's rival views one only ever infers that a note is deviant, or recognizes deviance by some other non-perceptual means. Proponents of these rival views will reject the argument above and claim that Jenna perceptually representing deviance is not the best explanation of the difference in her overall phenomenal state. Instead, these theorists will argue that Jenna's auditory experience is the same in both cases and will point to changes in one or more of Jenna's non-perceptual states as making for the overall phenomenal difference. Are any of these alternate explanations better than the explanation I provide?

This question is often adjudicated from the armchair. The best explanation is said to be the one that is the most explanatorily parsimonious, for example, or the least vulnerable to conceptual counterexamples. Instead of examining these well-worn considerations, I would like to take a somewhat different approach. Setting aside the purely theoretical advantages that one hypothesis about the contents of perception might have over another, my question will be: What role might empirical considerations play in deciding which explanation is best?

At the very least, I believe that empirical evidence should help to delineate the criteria of admissibility for candidate explanations. The best explanation of Jenna's phenomenal contrast will be the best explanation among those that are also consistent with our best science of the mind. This is, I think, a relatively unambitious thesis. We might press it further and argue that, when comparing two theories that are both minimally

compatible with the findings of science, we should look to see if either is better supported by those findings, and show more favor to theories that enjoy greater support.

Many philosophers who employ the phenomenal contrast method to settle questions about the content of perceptual experience ignore or underemphasize the supplementary role that cognitive science can play in weighing candidate explanations. Siegel (2011), for instance, discusses only how the Rich Content View (the view that we perceive high-level properties like being a pine tree), if it is better than rival explanations on theoretical grounds, might subsequently guide the inquiries of cognitive science:

> Aside from its bearing on philosophical issues, the Rich Content View is relevant to several other areas of research in psychology and neuroscience. [...] Where should we look for neural correlates of conscious visual experience? Whether the Rich Content View is true will influence what we will count as a neural correlate of visual experience. If the view is false, then we might expect to find these neural correlates in brain areas devoted to "early" visual processing, such as visual areas V1 and V5. If the view is true, then we should expect neural correlates to involve "later" areas, such as the fusiform face area (FFA) and the inferotemporal cortex (IT). (Siegel 2011)

But what of the other direction of influence? Might not the inquiries of cognitive science be independently brought to bear on whether we accept the Rich Content View? Scientists are already searching for the correlates of visual experience, and not always with strong prior conceptions about where to look for them. Siegel envisions that discovering the Rich Content View is true would guide the direction of cognitive research. But if the cognitive researcher independently discovered a correlate to the representation of higher-level properties, wouldn't the location of that correlate be precisely the kind of thing that could settle the question of whether such representation is perceptual or not, precisely the kind of thing, in other words, that could help philosophers decide whether the Rich Content View is true?

For instance, imagine an EEG study with two subjects, one with a species-typical brain and the other afflicted by a kind of blindsight. Say that both subjects look at an object and their EEG results show the same increased levels of activity in areas of the brain devoted to "early" visual processing (V1 and V5), but only the non-blindsighter shows increased levels of activity in parts of the brain devoted to "later" perceptual processing (FFA and IT). Furthermore, say that only the non-blindsighter reports visually experiencing the object; the blindsighter says that she sees nothing.

The fact that only the subject who showed "later" stage brain activity was conscious of whether they saw the object would be some evidence for the fact that the neural correlate of such experience is to be found in these "later" perceptual processing areas of the brain. And this is precisely where the Rich Content View, but not many its rivals, suggests such a correlate is to be found. So surely Siegel's observation might run in reverse. To the extent that whether the Rich Content View is true or false can tell us where to look for neural correlates, then where neural correlates are found might help us decide whether the Rich Content View is true or false. The experiment sketched above, for example, would at least lend credence to the Rich Content View, and detract from the plausibility of rival views that associate perceptual experience with only the earlier stages of perceptual processing (and thus the earlier areas of the visual system). In other words, empirical results have a role to play in the inference to the best explanation that decides on a favored theory about the contents of perceptual experience.

Unfortunately for the Rich Content View, blindsight does not work as described above. Rather, it involves damage to the V1 area of the visual system (Barbur et al., 1993). Thus, a blindsighter will not enjoy the same levels of early V1 activation as a typically

sighted person while differing only in suppressed downstream levels of activation. Rather, blindsighters exhibit diminished V1 activation, and consequently also diminished downstream activation. So a real life experiment along the lines of the one sketched above wouldn't lend support to either the Rich Content View or its rivals. But this is just a toy example—the point was simply that other, feasible experiments might.

Section 3: Empirical Considerations

Returning to DEVIANCE, which is a species of the Rich Content View, we may ask: is DEVIANCE consistent with our best science of the mind? Moreover, are there empirical findings that are like the fantasy blindsight case above, ones which lend support to DEVIANCE over and above rival explanations? I will argue that there are some empirical results that can contribute to our debate over DEVIANCE, namely recent experimental work on error perception in the auditory system.

First, what sort of empirical work would lend support to DEVIANCE? DEVIANCE hypothesizes that the property of being deviant is represented in perceptual experience. One way for DEVIANCE to be true is for the property of being a deviant sound to be represented in auditory experience. If deviance really is represented in audition in this way, then we should see some evidence that, when we experience sounds that upset our expectations, and which sound wrong, there is some signature brain event that occurs in centers of auditory processing, or those parts of the brain that otherwise give rise to or affect auditory experiences. Furthermore, this activity should be distinct from activity that occurs when we hear sounds that meet our expectations, and which do not sound wrong. If

such activity is found, and is indeed localized to parts of the brain that determine what is experienced as heard, then that is consistent with a story on which deviance is being represented in auditory experience, like the other auditory phenomena processed in those areas. In other words, there should be some sort of unique brain state that is seen only when deviant sounds are heard, and which is localized to parts of the brain where differences in activity correlate with differences in auditory experience.

If there were no sensory brain event of the kind described above, then critics of DEVIANCE would likely be correct: recognizing deviance would just be a matter of entering some purely cognitive or otherwise completely non-perceptual mental state.

Section 4: Pieszek, Schröger, and Widmann (2014)

Let us turn to one study that may be able to do the work described above, namely an EEG study conducted by Pieszek, Schröger, and Widmann (2014). I will argue that this study potentially confirms DEVIANCE, in that it appears to show (1) that we do represent whether a heard sound deviates from our expectations, and (2) that such representation is perceptual.

I believe that this study brings us an important distance towards meeting one of the challenges described above, in that it delivers a candidate for the neural state that makes for the phenomenal difference when we hear a note that sounds wrong. But whether or not this neural activity will satisfy our explanatory burden will depend on whether or not we consider it sensory. If the activity detected is cognitive rather than perceptual, then it actually lends support to the opponent of DEVIANCE, who argues that, although deviance

might be represented in experience, it is not represented in perceptual experience. This question will be addressed following the discussion of the study.

In the study, Pieszek, Schröger, and Widmann trained 25 subjects to associate particular symbols (white-on-top, white-on-bottom) with particular notes (high-pitched and low-pitched, respectively). Once they were taught these associations, they were then shown a row of five symbols. Then, after a duration of two seconds, subjects were played a five-tone melody (while still looking at the row). Following the melody, subjects were given a two second response window within which to evaluate the congruency of the trial, i.e. whether the notes they expected to hear on the basis of the symbols were the notes they actually heard. Subjects reported their evaluation by pressing one of two buttons, one for success, when the sounds corresponded to the symbols, and one for failure, when the sounds did not correspond to the symbols. Subjects performed this task an average of two hundred times each. Half of the rows were fully congruent with the melodies, and half of the rows contained one incongruent note, making the rate of incongruent symbol-pitch pairs 10%. Figure 1 (below, from Pieszek, Schröger, and Widmann, 2014) helps to illustrate the experimental design.

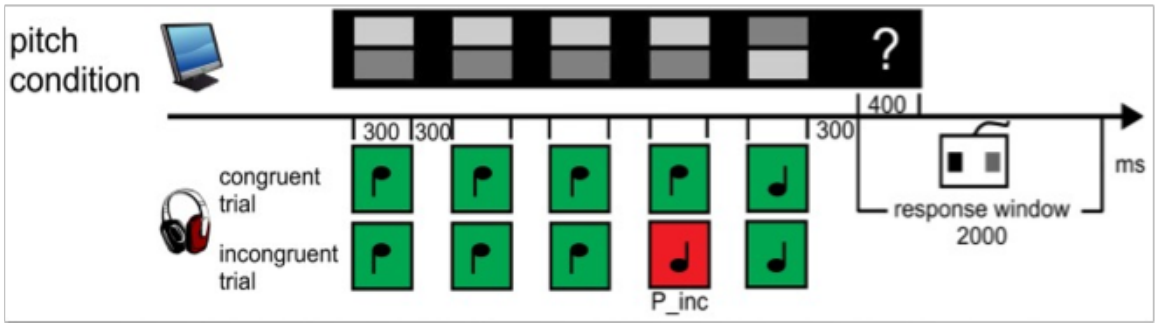

Figure 1

Subjects were typically very skilled at identifying whether or not the rows were congruent with the melodies. They had an accuracy rate of about 99% in identifying both congruent and incongruent sequences, and there was no difference in response times between the two categories (i.e., subjects did not take longer to identify incongruent sequences than congruent sequences, or vice versa). Indeed, the task is an easy one. Once the association between the visual symbols and the pitches of the notes has been learned, notes that fail to match the given symbols stand out. They are experienced as surprising, and as sounding wrong.

The experimenters analyzed EEG data from subjects as they listened to each note within the melodies. There were two distinct electrical responses, or event-related potentials (ERPs), that were only detected in the brains of subjects currently listening to notes that were incongruent with the accompanying symbols: the IR (Incongruency Response) and the N2b signal. In other words, incongruent note-symbol pairs elicited both responses and congruent note-symbol pairs failed to elicit either. Thus, Pieszek, Schröger, and Widmann identified patterns of neural activity that corresponded to perceivers' experiences of hearing notes that sounded wrong. Let's look more closely at the signals the experimenters detected.

4.1: IR (Incongruency Response): A sensory prediction error

The IR is a modality-specific response to deviant auditory stimuli. It is localized to regions of the auditory cortex that are associated with the relatively early stages of auditory processing. The detection of the IR is consistent with a previous experimental finding of several of the authors, which detected signals resembling the IR during pitch violations (Pieszek et al., 2013). Furthermore, the IR can be thought of as an analogue to the more general 'mismatch negativity signal' (MMN), which has been extensively described in the neuroscience literature on sensory perception. The MMN signal is thought to be generated by an automatic response within the brain's sensory system to infrequent discriminable changes in auditory input (Näätänen, Jacobsen, and Winkler, 2005).

On one of the most commonly accepted models, the predictive coding model, signals like the IR and MMN are generated by a mismatch between sensory input and automatically pre-activated auditory sensory representations (Bar 2007; Friston and Kiebel 2009). On such a model, once one has learned to associate certain symbols with certain sounds, the visual stimulus of a row of symbols actually causes one to form a kind of auditory map, which is represented in audition itself. Before each note has even played, an auditory representation of either the high note or the low note pre-activates in the early auditory cortex, depending on which symbol is seen.

This process serves to maximize the efficiency of the perceptual system, making it easier for the auditory cortex to process incoming stimuli in cases where it guesses correctly about what those stimuli will be like. In cases where the auditory cortex guesses incorrectly, it detects those mistakes at an early stage, and, instead of completing and

forwarding its pre-activated representations, it newly forms and forwards correct auditory representations instead (using the feedback it acquires to hone its future predictions).[2]

It is important to note that the IR is almost certainly pre-attentive (and thus likely not conscious). This is indicated by the authors of the study when they describe the IR signal (and MMN signals) as corresponding to the detection of violations "at sensory levels of processing", which they distinguish from the "attentive detection" that occurs at "cognitive- attentive level" of processing (Pieszek, Schröger, and Widmann, 2014).

That early-stage sensory signals like the IR and MMN are typically pre-attentive and pre-conscious has also been shown in other studies, such as Moreau's, Jolicœur's, and Peretz's 2013 study of amusia, a form of auditory agnosia. Amusia is generally associated with deficits in processing pitch, and those afflicted often struggle to recognize and reproduce musical tones (Pearce 2005). In Moreau's, Jolicœur's, and Peretz's study, amusic subjects were played a tone that suddenly changed in pitch, a phenomenon they were unable to detect. Nonetheless, EEG results showed that the subjects' brains did register the difference pre-attentively "at early unconscious levels of processing", and that this pre-attentive registration was reflected by the presence of an MMN signal (Moreau, Jolicœur, and Peretz, 2013). This is the same response that is elicited in non-amusic subjects. If both amusic and non-amusic subjects exhibit MMN responses, as Moreau, Jolicœur, and Peretz have shown, then such signals are not sufficient for eliciting the experience of deviance; and, because the IR signal is the same kind of signal as the MMN, it cannot then be a candidate correlate brain state to the experience of deviance.

---

[2] This account is also in line with the functional model of symbol-to-sound match processing described in Widmann et al. (2007).

What kind of neural activity does a person with amusia lack, then, when they fail to notice an unexpected change in the tone they are hearing? As it happens, Moreau, Jolicœur, and Peretz (2013) found that what their amusic subjects lacked was the exact same response that was elicited in subjects who successfully detected incongruent notes in the Pieszek, Schröger, and Widmann (2014) study: the N2b response.

4.2: N2b: A later prediction error

The N2b response occurs slightly after the IR, and takes place at what the experimenters call the "cognitive-attentive" level of processing (whether the signal is truly 'cognitive' in the philosopher's sense of the word will be discussed shortly) (Pieszek, Schröger, and Widmann, 2014). At its most general level of description, the N2b response is a signal that reflects the conscious detection of the violation of a prediction about perceptual inputs. In other words, the deviance of the sound is selected for attention, draws our attention, and becomes something that we notice (Ibid, 2014; see also Näätänen and Gaillard, 1983).

As mentioned previously, the Moreau, Jolicœur, and Peretz (2013) study of subjects with amusia also tracked this response and that work showed that neurotypical but not amusic subjects exhibited the N2b during unexpected changes in stimulus pitch.

So, to take stock: we have a perceptual representation of deviance that is not conscious (corresponding to the MMN or IR), and we have a conscious representation of deviance that may or may not be perceptual (corresponding to the N2b response).

It is important to separate the question of what we perceive from the question of what we perceptually experience. It seems safe to say that, at the very least, the results of

Pieszek, Schröger, and Widmann (2014) and others suggest that we perceive deviance, but this is only to say that our sensory system tracks it by way of mismatch negativity signals like the MMN and IR. Whether deviance is represented in perceptual experience, as the DEVIANCE thesis claims, is a further question. In order to answer that question in the affirmative, I will need to argue for two distinct claims. The first claim is that MMN responses like the IR do indeed single out stimuli as deviant, as I have suggested, as opposed to their being possessed of some concatenation of low-level properties. The second claim is that the N2b signal corresponding to the shift of attention toward the relevant stimulus is best understood as a perceptual rather than cognitive event. If I can successfully argue for each of these claims, then I have made a good case for DEVIANCE. I shall address both of these claims in turn.

Section 5: Do MMNs like the IR track deviance?

It might be argued that the IR and other MMNs do not necessarily single out stimuli as deviant. Perhaps, one might argue, they merely single stimuli out as being possessed of some concatenation of low-level properties, namely the low-level properties in virtue of which we take them to be deviant.

The first thing to note here is that the IR and the auditory MMN in general token in response to a wide range of auditory stimuli and in a variety of circumstances, where low-level perceptual features vary greatly. As a rule, the more general the circumstances under which a response is elicited, the more likely it is that a response is tracking a general, abstract property, rather than any particular surface-level feature. Furthermore, the auditory

MMN may be caused not only by deviations from regularity in pitch, as in the studies discussed above, but also by deviations in intensity or duration, and it is more pronounced and occurs more rapidly relative to how deviant the stimulus is from what a subject might reasonably expect (Paavilainen, 2013).

Secondly, the auditory MMN does not only occur in response to passively experienced violations of regularity; interestingly, it can also be triggered by violations of action intention. In a study by Korka, Schröger, and Widmann (2019), MMNs were elicited in subjects whose button presses were intended to produce a specific tone, even in a paradigm where the button presses only produced the intended tone 50% of the time (and so the effect cannot be attributable to a failure of regularity). This experimental condition maps nicely onto certain real-world cases of experienced auditory deviance, particularly cases where the subsequent phenomenology of ill-fit is especially pronounced. The singer who botches her opening note, for example, is in an analogous situation: she is thwarted in producing her intended sound, and because of that the note she sings sounds wrong to her.

Finally, the predictive coding model of sensory processing, mentioned previously, is one of the most standard models of sensory processing, and it is premised on the idea that the sensory system maximizes efficiency by not constructing sensory representations from the bottom-up, but rather by making educated guesses about incoming sensory inputs and then subsequently picking out—at the earliest stages of perception—those stimuli which deviate from those predictions, precisely on the basis of their deviance. As the sensory system learns more and more about the regularities of its environment, it benefits more from catching out falsehoods, so to speak, than it does from freshly constructing truths. The success of the predictive coding model thus itself forms the basis for a strong

argument that we should recognize deviance not only as a property that is tracked in perception, but as one that is tracked in very early perception.

Section 6: Is the N2b Perceptual?

If the N2b tracks a perceptual phenomenon, then the results of Pieszek, Schröger, and Widmann (2014) provide evidence that deviance is represented in perceptual experience. But perhaps the N2b does not track a perceptual phenomenon, but rather simply indexes the phenomenal experience of stimuli being cognitively classified as deviant, the way one might look at a painting and experience cognitively classifying the visual stimuli one perceives as indicative of a Rothko or a Pollock. Therefore, in order to rule out this latter cognitive explanation, and to thus make the N2b at least a strong candidate for grounding the representation of deviance in perceptual experience, we must motivate the idea that the N2b is not merely an experiential phenomenon but a perceptual phenomenon as well. Is it?

On the face of it, it may seem as if the answer is no; after all, Pieszek, Schröger, and Widmann (2014) discuss N2b as "cognitive". However, it is important to understand that the authors' use of "cognitive" is largely contrastive; they mean to differentiate the later N2b signal from the earlier IR, which is pre-attentive, unconscious and occurs at a much earlier stage of auditory processing. Other authors, whose motives are less contrastive, are more neutral in their descriptions and do not take pains to characterize the N2b as cognitive. So this terminological evidence alone is insufficient to establish that the N2b is not perceptual, at least in the sense relevant to philosophers. Indeed, I think we can argue that we should indeed understand the N2b as perceptual in the relevant sense.

First, the deviant stimulus to which attention is directed during an N2b response has, as I explained previously, already been classified as deviant by the IR or MMN early on in perceptual processing. Since the sound has already been classified as deviant by the perceptual system, it is reasonable to assume that it is presented as such in perceptual experience; the alternative is that the sound is stripped of that pre-assigned perceptual value and instead merely presented simpliciter in perceptual experience, whereupon it is subsequently re-categorized as deviant at the level of cognitive processing. Though not impossible, this seems unlikely.

Given the early perceptual processing of deviance, then, the argument that deviance is not represented in perceptual experience proceeds with much more difficultly than similar arguments concerning other high-level properties like that of being a pine tree. The analogous situation would be if we took that dispute, and then added to it the miraculous discovery of the existence of tiny, pine-tree-specific receptors in the early visual system. Were that to happen, I think it would be fair to say that Susanna Siegel and others who would forest our perceptual experience with such properties would have gained a significant edge in the debate.[3]

The most plausible explanation for the empirical and experiential data is that the deviant sound is not merely noticed as deviant and then redundantly classified once more as such via an inference or judgment, the way one might notice a painting and subsequently

---

[3] This argument as well as other things I have said previously might reasonably invite the question of whether deviance is actually a previously undiscussed low-level perceptual property, rather than a candidate high-level perceptual property. I think that to interpret deviance in this fashion would be mistaken, not to mention highly revisionary with respect to the philosophical literature's current understanding of high- and low-level perceptual properties. However, I will save my arguments to that effect for a future paper.

identify it as a Pollock; rather, it is most probable that the sound is heard as deviant from the very beginning.

There is one final argument for claim that the N2b corresponding to the perceptual experience of deviance. This argument asks us to consider the nature of attention, to which a shift in which toward the deviant stimulus the N2b corresponds. Although the shift in attention towards deviant sounds tracked by the N2b can occur partly as a result of top-down, cognitive influences (e.g. beliefs and expectations about what sounds one is likely to hear or produce), it can be argued that the mere presence of these influences does not merit describing the attentional shift itself as a cognitive phenomenon. This is because this attentional shift is a shift in auditory attention, and any shift in auditory attention is a perceptual phenomenon, whatever its causes. What it is to auditorily attend to a stimulus just is for the auditory system to be preferentially devoted to processing that stimulus. As Christopher Mole writes, "The processes responsible for the allocation of attention [are] inextricable from the processes that are responsible for the perception of the things to which we attend" (Mole, 2015). Thus, to the extent that the N2b tracks the direction of attention within a sensory modality, then, it may be considered a perceptual phenomenon, regardless of the contributions of top-down influences.

Section 7: Taking Stock

I hope to have made a good case that (1) ERPs like the MMN and IR track the representation of deviance in perception; and (2) later ERPs associated with conscious attention like the N2b track the representation of deviance in perceptual experience. If I am

right, then studies like the one conducted by Pieszek, Schröger, and Widmann (2014) provide empirical support for DEVIANCE, the thesis that we have the capacity to perceptually represent objects and events in experience as deviating from our expectations.

Chapter 2: No Work for a Comparative Concept of Insight

1: Introduction

We are said to experience insight, or insightful problem-solving, when we suddenly and unexpectedly become aware of the solution to a problem that we previously took ourselves to be unable to solve. These 'aha!' moments figure prominently in our cultural narratives of creativity and scientific discovery. Accordingly, there is a large literature in cognitive psychology dating back to the early twentieth century that is devoted to understanding precisely what insight is and how it occurs.

In recent years, comparative psychologists, who study the mental processes of non-human animals, have turned to the question of whether species other than humans are capable of experiencing insight. Scientists working in this research program claim to have found evidence of insight in elephants, various great apes, and several species of bird. As their results have drawn attention, however, so too have they drawn criticism. Comparative psychologists who are sceptical of animal insight worry that many of the studies do not successfully rule out competing deflationary explanations of the behaviors in question. They argue that many of the alleged cases of insightful problem-solving in animals are no more than the products of associative learning—which is a problem, as we shall see, given that on the received understanding of insight in comparative psychology, it must necessarily involve non-associative mental processes.

Despite these criticisms, the phenomenon of insight in non-human animals has gained a certain level of acceptance within the field of comparative cognition and many comparative psychologists regard it as a promising object of study. I will argue that they

are wrong to do so. Not only that: I will contend that these researchers are wrong even if the above criticism is unsound and some cases of animal insight genuinely are the results of non-associative processes. For, as I will show, there are deeper problems with the prospect of a comparative, cross-species study of insight.

2: The Comparative Ideal

Some comparative psychologists work to understand the thought and behaviour of just one particular animal family, genus, or even species. Others seek out truths that are common across various particular species, hoping to better piece together their intertwined phylogenetic histories. Perhaps the gold standard of comparative research – the comparative ideal, so to speak – is when we are able to use what we come to learn about animal minds in order to increase our understanding of human minds, thus contributing to and working in tandem with the project of cognitive psychology.

How might the study of insight in animals advance this comparative ideal by increasing our understanding of the phenomenon in humans? First, if insight is not a uniquely human phenomenon, by discovering it in animals we might learn something about its phylogenetic origins. Perhaps we can inform our guesses about when, in our evolutionary history, we developed the capacity for insightful problem-solving. Furthermore, if we discover it in species that bear little close relation to *homo sapiens*, this evidence of convergent evolution can teach about the kinds of evolutionary pressures that are conducive to the development of the capacity.

Second, discoveries about animal insight have the potential to constrain our theories about the cognitive underpinnings of insight. We might come to discover which processes

or capacities are necessary or sufficient for the occurrence of insightful problem-solving, and which are not. Consider an example. Say that we agree that creatures that φ demonstrate insight, and we observe both humans and domestic cats φing. By our lights, then, both humans and cats would be capable of insight. On the basis of this evidence, we could rule out theories on which the cognitive processes underlying insight necessarily rely on the presence of sophisticated, language-like representations, as it is unlikely that domestic cats possess such mental representations.

But can the study of insight in non-human animals in fact allow us to better understanding its counterpart in human cognition? To be sure, a number of comparative psychologists have thought so. Most studies of animal insight, however, do not make more than a token effort to connect with the literature on insight in humans, and so it remains a live question whether such a project could succeed. To answer this question, let us consider some of what would be required for a successful comparative study of insight in humans and non-human animals.

At first pass, we can specify three indispensable requirements.

1. We must be able to identify cases of insight in humans ($\text{insight}_{\text{human}}$).
2. We must be able to identify cases of insight in non-human animals ($\text{insight}_{\text{animal}}$).
3. $\text{Insight}_{\text{human}}$ and $\text{insight}_{\text{animal}}$ must plausibly be exercises of the same capacity ($\text{insight}_{\text{general}}$) and whatever criteria make it so they plausibly count as exercises of $\text{insight}_{\text{general}}$ must not be arbitrarily disjunctive.

While perhaps not sufficient, at the very least these conditions are necessary prerequisites: A comparative study of insight in humans and non-human animals must not fail to satisfy any the three conditions above if it is to be a productive scientific enterprise. (1) and (2) are relatively straightforward to explain. (1) requires that the cognitive psychologists have some agreed upon performance measure for determining when, by their lights, a human has solved a problem insightfully. Likewise, (2) requires that the comparative psychologists have some agreed upon performance measure for determining when, by their lights, an animal has solved a problem insightfully. (3) is the requirement that when we are talking about insight in humans and insight in animals, we are, more or less, talking about the same thing. As for the prohibition on arbitrary disjunction, I shall return to that later, for I think it is on precisely that count that the unification of the human and animal literatures on insight is doomed to fail.

First, let us consider (1) and (2), by looking at what researchers working on insight in animals and in human mean by insight, and how they operationalize that notion in order to identify its instances. Sara Shettleworth (2012) has the right of it when she writes that "[r]ecent comparative research on insightful behaviour has not been well integrated with contemporary research on human insight, largely as a result of confusions about definition." So we must get clearer about the relevant definitions before we can assess the possibility of integrating them.

3: Two Accounts of Insight

3.1: Insight in Cognitive Psychology

What do cognitive psychologists mean by insight? Sometimes, when you are making little progress on a problem, the means of solving it will come to you 'like a bolt from the blue': suddenly and surprisingly. Sometimes called a eureka experience or an 'aha!' moment, this is the phenomenon studied by cognitive psychologists under the heading of insight.

When it comes to the nature of insight, there is much debate, including debate as to whether insightful problem-solving is fundamentally different in kind from conventional analytic problem-solving (Sternberg and Davidson 1995). But cognitive psychologists who study insight typically agree on the following:

1. Insight has a distinctive and indispensable 'aha!' phenomenology.
2. Insight involves an experience as of breaking through a psychological impasse—i.e., typically one takes oneself to be unable to solve a problem before insightfully realizing its solution.
3. The onset of insight is sudden, and tends to be all-or-nothing; rather than the problem-solver making stepwise progress toward an insightful solution, they will typically become consciously aware of the solution (or means of reaching it) all at once.
4. Insight involves mental restructuring, wherein the problem at hand is approached in a new way.

Experimental work on insight in humans often uses puzzles in order to study the phenomenon in a controlled environment, relying especially on participants' verbal reports of 'aha!' phenomenology and other aspects of their problem-solving experiences. Certain kinds of puzzles are particularly good at eliciting the experience of insight in problem-

solvers, typically those whose solutions require ‘lateral thinking’ or some other radical reinterpretation of the constraints of the problem. A popular example is the Nine Dot Problem (used, for instance, in Kershaw and Ohlsson 2004, Chronicle et al 2001, and MacGregor et al 2001, along with innumerably many other studies) (Figure 2 below). Given the dot array on the left, the solver is told she must connect all nine dots using no more than four straight lines, and without lifting her pencil from the paper once she has started. The solution, shown at right, requires her to realize she can extend her lines beyond the square ‘box’ formed by the dots.

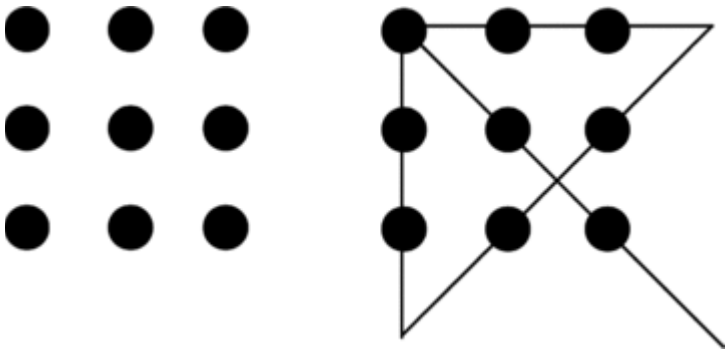

Figure 2

3.2 Insight in Comparative Psychology

Despite for the most part not having substantively engaged with contemporary debates about insight in the cognitive psychology literature, most comparative psychologists who study insight in animals do take themselves to be studying the same phenomenon as the cognitive psychologists. Often, for example, they introduce the idea of insight to their readers by way of reference to the cognitive psychologists’ oft-cited ‘aha!’ moment, and other aspects of the phenomenology that we are familiar with from human experience (as

in Renner et al 2017 and Neilands et al 2012). Many of them see promise in connecting their work on insight in animals to the work being done on insight in humans, and it is clear that for the most part they take themselves to be studying the same phenomenon.

However, whereas cognitive psychologists rely on their human participants' verbal reports of that 'aha!' phenomenology to identify moments of insight, those who work with animals must advert to strictly behavioral means of identifying insightful problem-solving in their non-verbal subjects. With that in mind, by far the most commonly cited characterization of insight in the comparative cognition literature comes from the ethologist William Homan Thorpe (1956), who defines it without reference to phenomenology as "the sudden production of a new adaptive response not arrived at by trial behavior … or the solution of a problem by the sudden adaptive reorganization of experience."

The comparative psychologist Nathan J. Emery interprets the Thorpean criteria in a way that is representative of his peers. He writes:

> The important terms to consider here are sudden, new, adaptive and reorganization of experience. For an action to be considered the result of insight, it must be spontaneous (i.e., not the result of explicit training or trial and error), novel (i.e., not performed before), functional (i.e., solve the problem and be goal-directed) and built from previous, untrained, similar behavior (i.e., not produced from copying earlier learned responses, but adapting previous behavior into new actions). (Emery 2013)

By introducing stipulations such as that the behavior must not be the result of explicit training or trial and error, or that it should not be produced by copying earlier learned responses, Thorpe and his intellectual descendants are, first and foremost, trying to rule out the case where the behavior is the product of mere association. A behavior that is an associative response is one that is stimulus-bound, and performed autonomically as the result of prior conditioning. Indeed, this is by far the most important part of the definition

in practice for comparative psychologists, who treat insightful problem solving as the counterpart to problem-solving that proceeds by way of automatic mechanisms, which are taken to be less 'cognitive' or less 'rational'. Indeed, elsewhere, in an earlier piece, Emery and co-author Bird gloss the Thorpean definition of insight simply as "a concept developed to explain sophisticated behavior that could not be the result trial-and-error learning" (Bird and Emery 2009). This account of insight, though simplistic, is the predominant one in comparative psychology; for example, Heinrich (1995) blithely proclaims that "[I]nsight can be shown indirectly to play a role in a behavior where learning and/or responses present from birth can be eliminated."

Given that most work on insight in cognitive psychology relies on a more nuanced concept of insight and benefits from participants who can be verbally cued and queried, it is unsurprising that experiments on insight non-human animals differ greatly from experiments investigating the capacity in humans. Comparative psychologists rely on tasks that are primarily designed to test whether animals can produce novel behaviors in order to solve novel problems without adverting to trial-and-error or behaving in ways that can otherwise be explained solely in reference to associative learning.

One of the earliest experiments into animal insight – and indeed, into insight in general – was performed by primatologist Wolfgang Koehler. Koehler hung bananas out of reach of his chimpanzee subjects and then provided them with various objects, including sticks and boxes. He claimed that his chimpanzees were behaving insightfully when, for example, after a period of inaction, they suddenly stacked the boxes atop one another in order to climb them to access the hanging bananas. (Koehler 2013, orig. 1921).

For any given instance of putative animal insight, there has been considerable debate in comparative psychology over whether it actually satisfies Thorpe's criteria, and in so doing counts as genuine. Consider, for example, Koehler's box-stacking experiment. A classic study by Epstein et al (1984) showed that pigeons can be induced to solve a similar box-stacking problem using only the tools of associative learning, though their performances of the solutions appear very like those of Koehler's chimpanzees, such that observers of the end result would have difficulty distinguishing their associative performances from genuinely insightful non-associative ones.[4] More recently, Taylor et al (2012) used New Caledonian crows to perform a similar debunking of much contemporary work in avian insight. This work demonstrates the extent to which associative mechanisms can produce behaviors that appear insightful, and it is easy to imagine that it might lead some researchers to worry that all putative cases animal insight may likewise just turn out to be, at best, the products of stimulus-response associations (albeit sometimes quite complexly layered networks of such associations).

However, that level of pessimism is often unwarranted. There are cases of animal insight that do appear to be sufficiently non-associative so as to satisfy Thorpe's criteria, especially when we know a great deal about the previous learning histories of the animals involved in the relevant experiments.

[4] Although I will not go into the details of Epstein's experiment here, Emery (Sanz et al 2013) provides some commentary: "What about Epstein's pigeons? Their behavior was not spontaneous (it was based on training), not functional (no reward), not novel (pecking and pushing – a form of peck – are both in a pigeon's repertoire) and was not adapted from previous similar behavior (rather than previous same behavior), so could not be considered as insight in any sense." A follow up experiment with pigeons by Cook and Fowler (2014) has also used appropriate controls to conclusively show that the driving mechanisms here are purely associative.

One example is the Aesop's fable paradigm (and its close cousins), where subjects typically obtain a floating reward in a deep, narrow container only by adding water to the volume or displacing the existing water with stones (rather than via direct approach such as reaching for the reward with fingers, beak, or claw).[5] Animals who have successfully solved this problem in controlled experimental settings include chimpanzees (Hanus et al 2011), orangutans (Mendes et al 2007), and rooks (Bird and Emery 2009). Successfully navigating this challenge has, in the case of many of these experiments, been seen by the researchers and even some outside commentators (e.g., Shettleworth 2012) as good evidence for insight in animals. This is attributable to the novelty of the presented problem, the novelty of the application of the 'insightful' behavior to the relevant class of stimuli, the suddenness of the emergence of the 'insightful' solution after failed direct attempts but often within the first session of exposure to the problem. In short, it is unlikely that previous learning or trial-and-error contributed to the animals' successes.[6]

Although I will not discuss them, other experimental paradigms have also provided robust evidence for Thorpean insight in animals, particularly those involving the intelligent use and even shaping of novel tools (Emery 2013; Bird and Emery 2009; Shettleworth 2009).

So it seems that the failure of some putative cases of animal insight to satisfy Thorpe's criteria is not in itself necessarily an ill omen for the possibility for the successful comparative study of insight, as there remain a class of cases that *do* plausibly satisfy the

[5] The paradigm is named for Aesop's fable of the Crow and the Pitcher, where a parched crow drops stones into a pitcher of water in order to drink from it.

[6] Note, however, that there are, as always, researchers who remain critical; see Jelbert et al (2015) for a skeptical explanation of these performances in rooks as merely prompted by rewarding perceptual feedback.

criteria. At the very least, let us grant proponents of animal insight access to these cases, and concede for the sake of argument that researchers were wrong to have been pessimistic *on that score*. I will argue that even these successes, bracing as they might be, are insufficient for the study of insight in animals to achieve what Section 2 called the comparative ideal. Although the response to cases that seem to succeed in satisfying Thorpe's criteria has often been to treat them as a proof of concept of the legitimacy of the comparative study of insight, I believe that this is inappropriate, as I will proceed to show.

4: Arbitrary disjunction

With our accounts of insight in place, let us return to the consideration of what is necessary in order to achieve the comparative ideal of collaborative work on insight between the fields of comparative and cognitive psychology.

Recall our three indispensable requirements from Section 2.

1. We must be able to identify cases of insight in humans ($insight_{human}$).
2. We must be able to identify cases of insight in non-human animals ($insight_{animal}$).
3. $Insight_{human}$ and $insight_{animal}$ must plausibly be exercises of the same capacity ($insight_{general}$) and whatever criteria make it so they plausibly count as exercises of $insight_{general}$ must not be arbitrarily disjunctive.

Section 3 demonstrated the requirements in (1) and (2) are nominally satisfied, discussing how each cognitive and comparative psychology defines and operationalizes the concept of insight. One thing we saw in that section was how very *different* the performance measures are that are deployed to measure insight by each of these fields. To illustrate, let

us directly compare a human participant who demonstrates insight by solving the nine-dots problem with an animal participant who demonstrates insight by solving an Aesop's fable water displacement problem.

In both cases, the insightful behaviors functionally solve the problems under investigation; furthermore, they are novel, sudden, and not reflexive first-pass responses. So there is much commonality in response despite the different means of elicitation. How, then, do the relevant behaviors and the researchers' interpretive focuses differ? In the case of the human who solves the nine-dots problem, her 'aha!' phenomenology, available to the researchers via her self-report, would be considered an indispensable diagnostic feature of insight. In contrast, the Thorpean criteria that are applied to an animal solving an Aesop's fable paradigm by design eschew questions of phenomenology. Additionally, the way in which a human's nine-dots insight involves a representational restructuring of the problem (in her case, a restructuring of the 'boundaries' of the nine-dots grid) would be of considerable significance to the cognitive psychologists, and taken by them to be a key part of what makes her solution insightful. In contrast, this kind of representational shift is not itself something that comparative psychologists look for in instances of insightful problem-solving in animals, nor is it something (looked for or not) that they typically discuss as if they assume it to be present in such instances (the way they often discuss insight's 'aha!' phenomenology). So comparative and cognitive psychologists who study insight use somewhat different working notions of insight and, accordingly and in accordance with species-relevant experimental constraints, deploy very different tests for insight.

This divergence of means of identifying insight is not in itself a strong worry for the project of attaining the comparative ideal. It is not unusual for a means of identifying

the exercise of a capacity in animals to differ radically from how we identify it in humans. We cannot explain a task to an animal nor can we ask it questions about its judgments or mental processes. In many cases, there is not even a test for the relevant capacity in humans, because it is transparent that humans possess it, and instances of its exercise are obvious (consider the literature on moral reasoning in humans and non-human animals, for example, where the presence of the capacity in the former is straightforwardly taken for granted). As long as the underlying capacity in question is well understood, and both the comparative and cognitive performance measures make sense as picking out the self-same capacity for $\text{insight}_{\text{general}}$, the comparative ideal might yet be achieved.

This is what is meant by (3)'s requirement that what makes something count as an instance of $\text{insight}_{\text{general}}$ must not be arbitrarily disjunctive. The desiderata is that $\text{insight}_{\text{general}}$ should *best* be understood as a singular capacity (or possibly a unified suite of capacities) that is exercised by both humans and animals, and not as the fusion of two fundamentally different capacities, one that holds for the human case and one that holds for the animal. For example, consider the capacity for vision (or rather, $\text{vision}_{\text{general}}$) as is instantiated in humans and in serpents (respectively, $\text{vision}_{\text{human}}$ and $\text{vision}_{\text{snake}}$). There are indeed many notable physiological differences between the human and reptilian visual systems and empirical measures for assessing both do vary; yet nevertheless, it is fairly easily accepted that $\text{vision}_{\text{human}}$ and $\text{vision}_{\text{snake}}$ are instantiations of the same functional capacity, $\text{vision}_{\text{general}}$, and thus the fusion of $\text{vision}_{\text{human}}$ is $\text{vision}_{\text{snake}}$ is not an arbitrarily disjunctive set but rather one that might form the basis for meaningful comparative work (in, say, how radically different evolutionary histories might shape the different expressions of the same functional capacity).

Contrast the 'good' case of $\text{vision}_{\text{general}}$ above as it is understood to operate over the particulars $\text{vision}_{\text{human}}$ and $\text{vision}_{\text{snake}}$ with the following concocted general 'capacity': visi-sodic $\text{memory}_{\text{general}}$. Let something count as an exercise of visi-sodic $\text{memory}_{\text{general}}$ if it is *either* an exercise of $\text{vision}_{\text{human}}$ *or* an exercise of episodic $\text{memory}_{\text{snake}}$. This is an extreme example of a general capacity that is arbitrarily disjunctive, and it should be self-apparent that the study of visi-sodic $\text{memory}_{\text{general}}$ on the basis of its component species manifestations is not worth pursuing for the sake of any legitimate comparative end (at least not one that is not far better characterized as in pursuit of understanding some other general capacity to which both vision in humans and episodic memory in serpents are relevant). If we hope to attain the comparative ideal, then our worry when conducting research in humans and non-humans animals is that we are really investigating something less like $\text{vision}_{\text{general}}$ and more like visi-sodic $\text{memory}_{\text{general}}$, where the animal performances we pick out are *not* in fact performances of the same general functional capacity as the human performances to which we are comparing them.

A reliable 'tell' for capacities that are arbitrarily disjunctive in the sense we are worried about is that they are difficult to articulate in species-neutral terms. $\text{vision}_{\text{general}}$, which is not arbitrarily disjunctive, is easily describable in species-neutral terms: it is the capacity to process a certain class of perceptual stimuli. The arbitrarily disjunctive visi-sodic $\text{memory}_{\text{general}}$, however, can most easily be described as the capacity to process visual stimuli if one is a human and to have certain kinds of remembrances if one is a serpent. A more species-neutral characterization is hard to devise. Humans who exercise visi-sodic $\text{memory}_{\text{general}}$ may be incapable of episodic memory; and serpents who exercise it might very well be blind.

I contend that $\text{insight}_{\text{general}}$ is arbitrarily disjunctive to an unacceptable degree. It is not possible to describe $\text{insight}_{\text{general}}$ as a species-neutral capacity such that the tests used in (1) and (2) are both still reliable indicators of that capacity's exercise. Not only are the comparative and cognitive psychologists currently identifying exercises of irreconcilably different capacities in their respective studies of insight, the *reasons* that comparative and cognitive psychologists are interested in the study of insight are so different that it is hard to imagine either party being willing to change their conception of insight to accommodate the other.

There is a general presumption that, as a matter of procedural order, attaining the comparative ideal involves comparing our work with non-human animals *against* our extant and ongoing work with humans in order to better understand the latter species. So in my arguments that follow, I shall mainly focus on the insufficiency of the Thorpean criteria as it is used by comparative psychologists to form the basis for a cross-species study of insight that attains the comparative ideal, but it should also become clear that the human-centric understanding of insight is equally unsuitable for application to non-human animals, should our end be to better understand our non-human counterparts.

In the following sections, I will raise three increasingly serious problems for any comparative study of insight, and particularly one that adverts to Thorpe's criteria.

The first problem with Thorpe's criteria for animal insight is that their stipulation that the behavior must be non-associative is out of step with the current literature on insight in humans. The second problem with Thorpe's criteria for animal insight is that they are too permissive: there is a broad set of clearly non-insightful behaviors they deem to be insightful. The third problem is a problem with the search for animal insight in general:

namely, that even if we *could* test for the distinctive 'aha!' phenomenology in animals, there is little reason to think it would be present in most of the putative cases of animal insight. Rather, the best explanations for 'insightful' animal behaviors will focus on the use of particular high-level cognitive capacities, rather than on the elusive capability for insightful problem-solving.

5: Issue One: The Non-Associative Constraint

As I discussed previously, on the comparative understanding of insight, insightful problem-solving has always been understood as necessarily involving something over and above associative processing. Indeed, no quality is seen to be more fundamental: to give an associative explanation for a putative case of animal insight is to disqualify it. Yet, as I will explain, cognitive psychology no longer considers this an uncontroversial assumption. Some cognitive psychologists are open to associative explanations of insight, some believe that all cases of insight are the result of associative processes, and others, interested in other aspects of insightful problem-solving altogether, simply want to remain neutral on the matter. By taking it to be foundational that insight is not associative, comparative psychologists will have trouble relating their research program to the work of any of these parties.

Let us further explore the cognitive psychologist's reluctance to claim that insightful problem-solving is non-associative and the comparative psychologist's universal endorsement of the same. Despite the consensus of comparative psychologists, it certainly cannot be said to be self-evident that the processes responsible for insight are not associative, given that part of what makes insight unique is that we lack conscious access to its processes, experiencing only the surprising and sudden awareness of the insightful

solution. So, at the very least, more must be said about why the comparative psychologists believe that a process that is inaccessible to us must be non-associative, especially given that a mental process not being consciously accessible is typically seen as a marker of fast, 'system 1' type reasoning, which, if anything, that makes such a process *more* likely to be associative.

One standard line of thinking among comparative psychologists is perhaps that, although insight's processes are inaccessible, its products are typically more sophisticated (and more impressive) than even standard non-associative and consciously-accessible forms of mental processing, and thus are unlikely to themselves be associative given that those processes are not.

It is important background that during the first half of the twentieth century, it was the standard assumption in the study of insight in *both* humans and animals that it was a non-associative phenomenon. At this early stage of intellectual progress, then, the currency of the view among comparative psychologists was not a particularly obstacle to collaboration with cognitive psychology. However, this status quo began to change beginning in the latter half of the twentieth century, as an increasing number of cognitive psychologists studying insight in humans entertained the idea that conventional (and associative) learning mechanisms could fully account for insight, and even creativity more generally (Sternberg and Davidson 1995). A number of other cognitive psychologists went further, not only expressing doubts that insight was as cognitive as previously thought, but explicitly arguing that some or all of its exercises are fundamentally associative. The associative theory of creativity, for example, suggests that insightful ideas are formed via a series of associative processes in semantic memory (Bowden, Jung-Beeman, Fleck, &

Kounios, 2005; Mednick, 1962). There is also a rich literature attempting to characterize insight in terms of dual process theory, debating its place in the System 1/System 2 framework, some researchers seeing it as within the province of the former and others as at least an interesting intermediary process between the two systems (Sowden et al 2015; Lin and Lien 2013).

During many of these developments in cognitive psychology, the study of insight in non-human animals remained in a state of relative dormancy, and so comparative psychologists weren't really in a position to respond to these sea changes. Yet when comparative psychologists resumed their study of animal insight starting in the mid-nineties, they picked up more or less where they left off, without having adjusted to their notion insight to reflect the developments that had taken place within cognitive psychology in the interim. Thus, by taking for granted that insight is non-associative, comparative psychologists beg the question on something that is currently a matter of heated debate within cognitive psychology.

To be clear, I do not think this worry alone constitutes an insurmountable obstacle. The comparative psychologist might respond thus: Though there is no longer consensus, some cognitive psychologists still *do* think insight is a distinct and non-associative process. Thus, although the way we comparative psychologists understand insight is in tension with the way *some* cognitive psychologists understand it, perhaps a successful collaborative research program can still emerge between our work and the work of those cognitive psychologists who do agree with us that insight is non-associative.

A comparative psychologist takes this line, however, has still more to contend with, as the following sections reveal.

6: Issue Two: Thorpe's Criteria are Too Permissive

A second worry for the reconcilability of the cognitive and comparative concepts of insight is that there are many human behaviors that all cognitive psychologists would agree are non-insightful, but which nonetheless satisfy Thorpe's criteria and are thus counted as insightful by the lights of comparative psychology.

Consider the following case:

> Although I have always been exceptionally tall, a recent back injury has left me with a pronounced stoop. While cooking breakfast, I go to retrieve a box of pancake mix from the pantry. The box is on the highest shelf. Though I was able to obtain things from that shelf before my injury, when I attempt to do so now I discover that the box is beyond your reach, even when I stand on my toes. "If I want to reach the pancake mix," I think, "I will need something to stand on." Accordingly, I fetch a chair from the dining room, and with it I am able to reach the pancake mix and commence with breakfast.

This case satisfies all of Thorpe's criteria for insightful problem solving. I fetch the chair only after my habitual methods of reaching the pancake mix have failed, and my use of the chair is genuinely novel, as I have never before needed to use a chair or similar object to extend my reach in this manner. We might even imagine that this is not even something that I have previously observed people doing. My solution was the product of at least some cognitive processing, as I reached it through occurrent deliberation. And once I committed to that solution, I executed it in its entirety, smoothly and rapidly. By the lights of the comparative psychologist, then, I have demonstrated insight.

To the cognitive psychologist, on the other hand, my solution is not insightful. For, firstly, according to cognitive psychologists, insight necessarily involves an *experience* as of breaking through a psychological impasse--i.e. taking oneself to be unable to solve some problem, before suddenly realizing its solution. I did not experience such an impasse when confronted with the out-of-reach pancake mix; although I failed to reach it when I tried, it never seemed to me as if there were no possible means of obtaining it. Nor did I experience the 'aha!' phenomenology characteristic of the moment when the impasse is broken thanks to a flash of insight. And the 'aha!' phenomenology is absolutely central to insight, as understood in cognitive psychology (which explains why cognitive psychologists rely so heavily on phenomenological self-report in their investigations of insight).

Secondly, there is considerable agreement among cognitive that insight involves a mental restructuring of the problem at hand. On this view, insight involves realizing that the problem can be seen in a new way, just as much as it involves realizing the solution afforded by that new viewpoint. In my case involving pancake mix, it is true that the chair was not a part of my initial representation of the problem. Nonetheless, the realization that I needed the chair didn't require me to 'think outside the box'. My initial representation of the problem space did not explicitly include the chair among its elements, but neither was the chair excluded by some assumption. Indeed, the moment reaching the box actually became a problem that I needed to solve, it was immediately obvious that I needed the chair or something like it to solve it.

So, it is clear that merely applying Thorpe's criteria for insight to human performances will lead to what cognitive psychologists would consider 'false positives'. What, then, are the prospects for reconciling the cognitive psychologist's notion of insight

with that of the comparative psychologist? The fact that the two fields use different criteria for identifying instances of insight, and even disagree about some cases, is not, all by itself, decisive. For if there were enough overlap between the instances classified as insight by cognitive psychology and those classified as insight by comparative psychology, there would be some hope that both fields are, in their own ways, latching on to the same phenomenon. If there were to exist some identifiable subset of both human and animal behaviors that both parties were willing to call insightful, progress would still be possible.

Unfortunately, there exists no such subset. Even if the comparative psychologists were willing to accept all of the human performances of insight ratified by the cognitive psychologists, there is likely no extant case of animal 'insight' (let alone subset of cases) that the cognitive psychologists would agree to accept. This is because, for cognitive psychologists, the distinctive phenomenological profile of insight is not just a central diagnostic criterion but an indispensable one. The pancake mix case in the previous section did not count as insight partly because insight's distinctive phenomenology was absent; extant putative cases of animal insight are similarly inadmissible, because satisfying Thorpe's criteria does not entail anything about one's phenomenology. The likely stance a cognitive psychologist will take towards the comparative psychologists' proposed comparison class is simply to suspend judgment about whether or not insight has occurred in the animals under investigation.

This problem runs deeper than the insufficiency of Thorpe's particular criteria to track what the cognitive psychologists care about. We cannot reconcile the comparative and cognitive notions of insight simply by augmenting Thorpe's criteria or replacing it altogether with a phenomenological criterion for animal insight, because given our lack of

access to animal phenomenology, such a criterion--as Thorpe well knew--cannot presently be operationalized.

Quite simply, we lack the means of identifying the occurrence of 'aha!' phenomenology in non-human animals. It is true that some work in comparative psychology is aimed at tapping into animals' affective states (e.g., Bekoff 2004). This work is still immature but even on a very sanguine interpretation of its current and future prospects, things don't look promising for the study of insight. Identifying the phenomenology that is characteristic of insight requires a lot of specific knowledge-- the distinction between 'aha!' phenomenology and other more basic affective states is quite sophisticated and subtle. It would be quite hard to tell apart happiness at solving a problem (and thus gaining foreseeable access to a food reward, say) and the surprise of a genuinely insightful eureka moment, especially since the latter often also causes the former.

But let us imagine that we *could* tap into non-human affective states with the specificity necessary to identify 'aha!' phenomenology. The next problem I will raise is that even if this were possible, successful animal performances under the current paradigms of the insight literature are unlikely to involve any such phenomenology on the part of the problem-solvers.

7: Issue Three: The Absence of 'Aha'

The third problem I raise is not a problem with Thorpe's criteria for animal insight in particular, given that for pragmatic reasons they sidestep the question of animal phenomenology. Rather, the problem is that even if we *could* test for the distinctive 'aha!' phenomenology of insight in animals, there is little reason to think it would be present in

most of the putative cases of animal insight. Previously, I have spoken as if the putative cases of animal insight might be genuine though unable to meet the cognitive psychologists' affective burden of proof. In this section, I posit that the behaviors in question may not be instances of insight at all. Rather, the best explanations for 'insightful' animal behaviors will focus on the use of particular high-level cognitive capacities, rather than on the elusive capability for insightful problem-solving.

What, then, do I suggest that the clever animals in insight experiments are achieving instead of insightful breakthroughs? Quite simply, I believe that many alleged instances of insightful problem-solving in animals simply reflect an understanding of causal relationships. Successes on the Aesop's fable paradigm and on many other tool use paradigms are certainly best interpreted in this way.

When initial, automated efforts to solve a problem fail, an animal, if they are able, might shift their attention to the causal relationships within the problem space. Though this shift in attention may make the solution to the problem newly accessible, I do not think it constitutes an instance of insight, as cognitive psychologists understand the phenomenon. Rather, it is simply the methodical application of a higher-level mental capacity after first-pass lower-level approaches have failed. In other words, I take what is going on in *most* animal insight cases to be more akin to the pancake mix case described in the previous section than to anything the cognitive psychologist would recognize as insight.

To drive this point home, on the standard 'insightful' interpretations of the Aesop's fable task and many tool use tasks, the insights that are reached are already understood by comparative psychologists as insights about causal relationships--and so the capacity for causal understanding is already baked into the explanation. In that case, why not more

parsimoniously let the capacity for causal understanding do all of the explanatory work? The only thing insight adds is a phenomenology that we cannot plausibly detect.

Other higher-level cognitive capacities, too, may be behind 'insightful' animal behaviors. An animal may resort to *mental* trial and error, for example, when confronted by a difficult problem, whether that involves causal reasoning, imaginative first-person simulation, or just the visualization of how objects might fit together in the scene (Gruber et al 2019; Redish 2016; Finke and Slayton 1988). Again, these are processes that no cognitive psychologists would think are distinctive to insightful problem-solving.

If there is one thing that all of the putative cases of genuine animal ‘insight' might have in common, it is that they are cases means-end reasoning. A creature has the capacity for means-end (or instrumental) reasoning if they are able to identify and execute intermediary courses of action in service to their goals. Critically, they must grasp those intermediary courses of action as to-be-realized *because* they will centrally contribute to actualizing the overarching goal. A combination of causal understanding and goal-directedness, the capacity for means-end reasoning is an important cognitive milestone. In humans, this capacity begins to emerge in infancy between 6 and 7 months of age; competence then grows, until the infant is engaging in spontaneous and fully intentional means-end actions between 9 and 12 months. (Piaget and Cook 1952; Willatts 1999; Sommerville and Woodward 2005)

In addition to the capacity having been well-studied in humans by cognitive psychologists, there is a rich and literature on means-end reasoning in comparative psychology (see Krasheninnikova 2019 for a general overview). In fact, a great deal of the work on animal insight is by its own lights a part of that literature, so much so that some

ethologists have complained about others outright conflating the capacity for insight with elements of means-end reasoning such as causal understanding, which, as ethologist Sara Shettleworth points out, “need not be implicated in insightful solutions by people” (Shettleworth 2012).

I take the three problems raised in this and the previous two sections to present a compelling case that the comparative understanding of insight can do little productive comparative work. Certainly it falls short of what is needed for work that attains the comparative ideal. It is irresolvably incompatible with what cognitive psychologists (and their experiments) take insight to be. Yet the fact that we can readily reinterpret ‘insightful' animal behaviors as exercises of other higher-level cognitive capacities, and interpret them within the broader framework of means-end reasoning should be a consolation to the comparative psychologist. Rather than being wasted work, the bulk of animal insight experiments can be retrofitted into the existing comparative research programs that are devoted to the development and exercise of those capacities in human and non-human animals. Indeed, the comparative psychologists who study insight in animals are already sensitive to the contributions of means-end reasoning and its ancillary capacities to successful ‘insightful’ performances (e.g., Cook and Fowler 2014; Huber and Gajdon 2006; Foerder et al 2004).

Indeed, it is already the case that some comparative psychologists either straightforwardly conflate insight with these capacities. Those that are not guilty of this conflation are typically at least sensitive to the ways in which cases of animal insight involve these capacities. So, the move away from insight and towards these other capacities is less revisionary than one might expect.

A final point in favor of moving away from comparative psychology's current preoccupation with insight is that it is clear that the most important theoretical role that animal insight plays in comparative psychology is largely that of a foil to simpler, associative explanations of animal behavior. As we saw, insight *does* not play this role in the cognitive psychology literature, where it is of theoretical interest largely because of its role in creative thought as well as its distinctive phenomenological profile. Furthermore, as we saw, cognitive psychologists hotly disagree about whether the processes culminating in insight are best understood as associative or non-associative process.

In the previous section, I discussed some particular higher-level capacities that many posited cases of animal 'insight' plausibly exploit. I also discussed the way in which all such cases can be understood as exercises of a more general capacity, means-end reasoning. I recommended that comparative psychologists take up the means-end reasoning framework because it allows them to capture what their diverse cases of animal 'insight' have in common. Here, I want to recommend it on a further basis: comparative psychologists typically agree that one of the distinctive features of means-end reasoning is that it is not stimulus-bound, or, in other words, merely associative. Otherwise, it would not be a form of *reasoning* at all. Thus, in the comparative cognition literature, means-end reasoning can fulfil insight's theoretical role at least as well as insight can, and can moreover better ground projects that pursue the comparative ideal, given that the non-associative characterization of means-end reasoning will not put them at odds with cognitive psychology.

8: Conclusion

In conclusion: You cannot bring the comparative and cognitive concepts of insight into harmony under a species-neutral characterization of insight. Insight's distinctive phenomenology is diagnostically indispensable to the cognitive psychologist and we do not currently have access to strong evidence about the character of animals' phenomenal experiences, nor are we likely to have such evidence in the foreseeable future such that we could detect a phenomenal state as finely-grained as insight. Furthermore, even if we could detect insight's distinctive phenomenology in animals, we do not have strong reason to think that animals who pass Thorpe's criteria would possess it; and indeed, we have some reason to think that they would not.

However, we can reinterpret successful performances of animal 'insight' literature as exercises of other high-level cognitive capacities, many of which comparative researchers already take to be involved alongside insight. When these capacities are understood as facets of means-end reasoning, we can capture what diverse performances of animal ‘insight' have in common. Furthermore, means-end reasoning is well-suited to filling the theoretical role insight has played in the comparative literature, more so even than insight, in that it both comparative and cognitive psychologists agree that it is distinctively non-associative. So, although a comparative research program on insight that aims at the comparative ideal is doomed to fail, a comparative research program on means-end reasoning that aims at the comparative ideal may well succeed. Indeed, there are already such research programs underway, and folding the extant research on animal insight into those research programs is a strategy that holds promise.

Chapter Three: Punishing Moral Animals

Arguments to the effect that humans bear moral duties towards non-human animals are commonplace. Figures as diverse as Peter Singer, Martha Nussbaum, and Jane Goodall have offered reasons to refrain from hunting, meat-eating, and (some) animal experimentation. But what if some non-human animals are not just moral patients but agents as well, capable of moral actions in their own right? In recent years, some have speculated that they might be, having been moved by examples of courageous dolphins and sensitive apes. Animal altruism is in the air, a rising chorus of academic monographs, scientific research, and TED talks.[7]

Yet consider the following: If there are indeed animals with the capacity to act morally, we might also think that many of them are morally responsible for their conduct.[8] But wheresoever moral responsibility goes, desert is sure to follow. I will argue that if animals are moral actors, then humans have strong reason to punish animal wrongdoers. For the believer in animal morality, this is a worry; for the skeptic, a potential *reductio.*

[7] Some examples of philosophical work on the question of whether animals are moral agents are Andrews and Gruen (2014); Bekoff & Pierce (2009); Clement (2013); Clark (1984); DeGrazia (1996; 1997); Dixon (2008); Gruen (2002); Korsgaard (2006); Monsó (2015); Musschenga (2013); Peterson (2011); Pluhar (1995); Sapontzis (1987); Fitzpatrick (2017); Rowlands (2011; 2012; 2013); and Shapiro (2006).

[8] The kind of responsibility I have in mind is that which makes one blame- or praiseworthy for one's actions, in the moral sense.

This move from agency to responsibility does not seem to have been anticipated by the bulk of philosophers who have considered the ramifications of animal morality.[9] For one, they have typically focused on the emergence of the negative rather than the positive duties we might owe animals. If animals are moral actors, they say, we have further reason not eat them, experiment on them, or otherwise indenture them to our needs and desires. Furthermore, when positive duties *are* proposed, they are typically only the counterparts of the aforementioned negative duties. If we have a duty to refrain from animal experimentation, for example, then perhaps we also have a duty to liberate those animals currently being experimented upon. One notable exception that explores a positive duty in its own right is work that touches on the question of whether we ought to confer the moral or even legal status of personhood on animals who demonstrate moral capacities (see, for example, DeGrazia 1997), but even that is typically intended to serve the ultimate end of protecting animals from consumption, experimentation, and the like.

Unlike, say, the duty to liberate, the positive duty to punish moral animals is not the correlate of any negative duty we might bear towards them. And yet it clearly merits discussion. After all, if you were to acknowledge such a duty, your relationship to the animal kingdom would be radically transformed. Furthermore, if you want to claim that some animals *are* moral actors, but avoid my proposed duty to punish, then your views in other areas of philosophy may well have to give.

In Section 1, I sketch why many now find it plausible that some non-human animals are moral actors. In Section 2, I explore which philosophical accounts of responsiveness to

[9] The only exception that I know of is Mark Rowlands' *Can Animals Be Moral?* (2012), which very briefly uses a potential retributivist duty to punish to reductio the view that animals are truly moral agents (pp. 83-84).

moral reasons are compatible with the prospect of the moral animal and which are not, an exercise that demonstrates how believers in animal morality are constrained in their choice of broader moral theories. In Section 3, I argue that if moral animals do exist, they are likely to be severely limited in their ability to morally evaluate one another, if they are even capable of moral evaluation at all. Furthermore, I argue that the severity of these deficits is such that impartial human observers would generally be more competent to see to matters of animal desert than the moral animals themselves. Building on those arguments, Sections 4 and 5 make the case that if there are moral animals, then, contrary to our intuition, both retributivists and deterrence theorists about punishment ought to recognize a strong reason to punish animal wrongdoers and perhaps even a duty to do so. Finally, Section 6 responds to the objection that considerations from moral relativism tell against my conclusion.

1: The Case for the Moral Animal

Often it can seem as if an animal's actions are guided by moral concerns. Some anecdotal examples of animals who appear to be acting morally are so old as to have their own literary traditions: reports of dolphins rescuing sailors from shipwreck, for example, occur as early as Plutarch, and the trope of the loyal and heroic dog has been a cultural touchstone throughout much of human history, persisting from the Welsh fable of Gelert to Lassie of the silver screen. But the ubiquity of moral animals in popular culture is hardly a reason to take them to actually exist—otherwise we might find ourselves believing in witches, ghosts, and all other manner of folk superstitions. Furthermore, there are few rules considered more fundamental to the study of animal minds than Morgan's Canon, which warns that "in no case is an animal activity to be interpreted in terms of higher

psychological processes if it can be fairly interpreted in terms of processes which stand lower in the scale of psychological evolution and development" (Morgan 1894).

On the face of it, Morgan's Canon seems to tell against the hypothesis that there are moral animals, in that whenever an animal appears to act morally, there are ways to explain that animal's behavior in terms of less sophisticated mental processes. To illustrate, imagine that you are visibly upset and your beloved spaniel Brownie rests his head on your knee in a way that gives you comfort. One explanation of Brownie's behavior is that Brownie, sensitive creature that he is, recognizes your distress and intends to alleviate it. On this explanation, Brownie behaves morally. However, consider, as Morgan's Canon would have us do, the spectrum of alternate, lower-level explanations: Perhaps Brownie, mindless brute that he is, acts purely on instinct (perhaps such behavior was adaptive at some point in his species' developmental history). Or perhaps he has been praised for similar actions in the past, and has thus learned to respond in this manner through operant conditioning. Or perhaps he more mindfully – but still amorally – takes you to be the leader of his pack and recognizes that you are an ineffective guardian in your current state; alleviating your distress, whether or not Brownie recognizes it as distress, is merely a way for him to restore you to the condition of a reliable leader and protector. What is it that has led some researchers to think that the moral for animal behavior like Brownie's is often the best explanation, despite the psychological complexity of the processes involved?

One important factor is that the last twenty years of research in comparative psychology have revealed that non-human animals are capable of more complex cognitive feats than previously thought, such as tool use, social learning, insightful problem-solving, long-term planning, deception, and even (debatably) theory of mind. As a result of such

findings, comparative psychologists are increasingly willing to entertain hypotheses about animal behavior that in previous decades would have been dismissed as requiring too much cognitive complexity to be credible. Accordingly, researchers have directly investigated the question of whether animals can be moral, and some have taken their results to support an answer in the affirmative. Much of this work focuses on animals who act out of concern for others.[10] For example, an early study by Masserman et al. (1964) set an influential precedent by showing that macaques preferred to go without food, sometimes for days on end, rather than accept it from a mechanism that administered a painful electroshock to a conspecific. A more recent (and less grisly) study showed that capuchins reliably choose to deliver desirable foods to capuchins in separate chambers, even when doing so predictably diminishes the desirability of their own reward (Lakshminarayanan and Santos 2008). There has also been a proliferation of work on whether animals are sensitive to considerations of justice or fairness. A notable example is Brosnan and de Waal's 2003 study on inequity aversion in capuchins. Brosnan and de Waal showed that capuchins willingly perform tasks alongside a companion when both are rewarded with pieces of cucumber for their participation. However, if their companion is given a more coveted reward (grapes) while they continue to only receive cucumber, they will refuse to participate in further tasks (with some individuals going so far as to hurl their cucumbers back at the researcher). A similar aversion to inequity has also been found in chimpanzees, our nearer primate relatives (Brosnan, Schiff, and de Waal, 2005).

[10] I describe but a handful of these studies here, but their subjects range from rodents to cetaceans, and curious readers can see Bartal, Decety, and Mason 2011; Church 1959; De Waal, Leimgruber and Greenberg 2008; Masserman, Wechkin, and Terris 1964; Palagi and Norscia 2013; Preston and de Waal 2002a and 2002b; Schino and Aureli 2009; and Silk 2007.

In the realm of neuroscience, Panksepp (1998) and Berridge (2003) have both argued that the basic emotional repertoires of animals, and particularly primates, overlap greatly with our own, owing to shared brain structures (including the prefrontal cortex, cingulate cortex, and amygdala). There is evidence from neuroscience that many animal brains share markers associated with human moral capacities or sentiments. For sentimentalists and others who take morality to be grounded in affective states, such findings are particularly significant.

All of the empirical work outlined above dovetails nicely with an evolutionary understanding of the animal kingdom as a kingdom to which the morally capable *homo sapiens* belongs rather than towers above. Indeed, considerations of evolutionary continuity naturally encourage an openness to the possibility that certain animals besides humans might in fact be moral actors. It is standard best practice to explain the presence of a trait in a variety of populations by positing that the trait arose in a common ancestor of those populations. Humans and many other animals exhibit patterns of behavior consistent with responsiveness to moral reasons. We therefore have good reason to posit a single mechanism or capacity, inherited from a common ancestor, to explain why humans and animals share these patterns of behavior. Additionally, we have firsthand knowledge that humans, at least, are genuinely responding to moral reasons when they exhibit seemingly moral behaviors. So, there is a prima facie case that other species are responding to moral reasons, as well, especially those with whom we have a great deal of common ancestry, like primates.

Although the case for the presence of a moral capacity is strongest in species closely related to humans (since we know for a fact that humans are moral), it is also possible for

species without that kinship to have independently developed the ability to respond to moral reasons via convergent evolution, as a result of facing similar selection pressures. The development of a moral mechanism of the kind possessed by humans may well be the most adaptive response to those shared challenges.

A prominent advocate of evolutionary approaches to morality is primatologist and ethologist Frans de Waal, much of whose work is devoted to identifying precursors to human morality in the social behaviors of primates (de Waal 1996; 2009; Flack and de Waal 2000; Preston and de Waal 2002a). Adopting the framework of convergent evolution, De Waal has asserted that many animal species, not just our close primate cousins, exhibit the markers of a moral framework that closely resembles our own (de Waal 1996).

Finally, it may even be the case that the hypothesis that there are moral animals is most *parsimonious* way to explain the phenomena discussed in this section. The adherent of Morgan's Canon would need to posit a great many diverse lower-level psychological processes to explain some of the experimental results discussed above—whereas, if we are willing to deviate from the Canon and attribute rudimentary moral cognition to animals, it is but a single capacity (albeit a sophisticated one) that shoulders the entirety of the explanatory burden. Positing one capacity with far greater explanatory power may well be preferable to positing many, each of which have lesser explanatory power—even if that one capacity *is* rather sophisticated.

Whether or not one is convinced that some animals are genuine moral agents, it is a view that many scholars take seriously, and moreover one that a rising number consider to be empirically supported. Thus, I hope that what I have to say about the implications of that view will be of interest not just to its adherents but to skeptics and neutral parties alike.

2: Animals and the Meaning of Moral Agency

If animals *are* moral agents, then what does that commit us to? Before we can work out the consequences of ascribing morality to animals, we need to get clear on what exactly it is that we credit animals with doing when we say they are behaving morally. In this section, I demonstrate how believers in animal morality are constrained in their choice of broader moral theories by showing which philosophical accounts of responsiveness to moral reasons are compatible with the prospect of the moral animal and which are not.

There are various ways to understand a responsiveness to moral reasons. On the richest available conception, someone who acts in response to moral reasons performs an action because they believe that, morally speaking, it is the right thing to do.[11] Call this account, famously endorsed by Kant, *rich moral responsiveness*. For Kant, a morally good action is not merely one which happens to "*conform* to the moral law"—rather, "it must also happen *for the sake of this law*" (Kant 2002).

To satisfy the requirements of rich moral responsiveness, one must possess and deploy some concept of morality itself. After all, one cannot act for "the sake of [the moral] law" lest one recognize that a certain action satisfies the mandates of that law. Thus, for the rich moral responsiveness theorist, the prospect of the moral animal is relatively implausible. Most animals will be excluded from the class of moral actors from the outset, namely those which are incapable of such sophisticated or abstract cogitation as higher-order moral reflection. Rats, for instance, need not apply, and even elephants or

[11] Or, in the case of a moral Lucifer, because they recognize that, morally speaking, their chosen action would be *wrong* to do.

chimpanzees are unlikely to meet such a high standard for moral action. Christine Korsgaard demonstrates this reasoning when she writes,

> We have ideas about what we ought to do and to be like and we are constantly trying to live up to them. Apes do not live in that way. We struggle to be honest and courteous and responsible and brave in circumstances where it is difficult. Even if apes are sometimes courteous, responsible, and brave, it is not because they think they should be. (Korsgaard 2006)

And just as rich moral responsiveness theorists like Korsgaard find themselves unable to ascribe morality to animals, so too, by *modus tollens*, the believer in animal morality cannot endorse rich moral responsiveness.

At first, this may not seem too terrible a result. After all, there are some independent reasons to reject rich moral responsiveness in favor of a weaker account. One reason is that its high bar risks ruling out many *human* actors. A bystander might catch a falling baby simply because she doesn't want the baby to be injured, without, in the heat of the moment, recognizing that allowing the baby to fall would be immoral, or even that catching the baby would be right. Although the bystander may choose her action on the basis of its right-making *features*, she isn't responding to the *rightness* of catching the baby, and so, according to rich moral responsiveness, fails to count as responsive to moral reasons. This worry has led most proponents of rich moral responsiveness to retreat to the position that one need only be *capable* of such reflection in order to count as a moral actor, whether or not one does in fact reflect on any particular occasion. Philosophers who endorse something akin to this weakened version of rich moral responsiveness are John McDowell (2009 and 1979), Susan Wolf (1987), and Dana Kay Nelkin (2011). However, even this

weakened version of rich moral responsiveness is still too strong for animal moral agency, since it was the very fact that an animal is unlikely to have even a *capacity* for higher-order moral reflection that made the moral animal hypothesis incompatible with the stronger version of the view.

With both strong and weak rich moral responsiveness accounts off the menu, believers in animal morality have a somewhat restricted diet. They can, however, endorse an interpretation of responsiveness to moral reasons that lowers the bar to mere responsiveness to an action's right- or wrong-making features, if not the rightness or wrongness of the action itself. Call this view *feature-based moral responsiveness*. Its most prominent defender is Nomy Arpaly (2003), but others who are sympathetic to it are Elizabeth Harman (2011 and 2015) and, after a fashion, Julia Markovits (2010). In her *Unprincipled Virtue* (2003), Arpaly writes:

I take a person to be responsive to moral reasons to the extent that she wants noninstrumentally to take courses of action that have those features that are (whether or not she describes them this way) right-making and not to take courses of action that have those features that are (whether or not she describes them this way) wrong-making features.

The non-instrumentality Arpaly emphasizes is an important constraint to the view: if I comfort you because your sobbing is interrupting my nap, I'm obviously not motivated by the moral features of my action. Like the rainstorm that saves lives by staunching a destructive wildfire, I may be achieving a good, but only accidentally. I am not morally praiseworthy. If, on the other hand, I comfort you simply because I want to dispel your distress, regardless of whether you interrupted my nap, then I am directly motivated by the

right-making features of my course of action, and, as such, I manifest feature-based moral responsiveness and, by Arpaly's lights, deserve moral praise.

Unlike the rich moral responsiveness view, the feature-based account allows for the possibility that creatures who lack higher-order moral concepts might still be moral actors. The non-instrumental desire to, say, defend of a victim of bullying is sufficient to make doing so a response to moral reasons, regardless of whether the defender recognizes that their action is right, or even that bullying is wrong.

Furthermore, sentimentalism and related understandings of moral responsiveness are at least as friendly to the prospect of animal morality as the feature-based account. Such views take emotions to ground morality, and their proponents include David Hume (2014), Adam Smith (1817/1759), Jonathan Haidt (2001), Michael Slote (2010), and Jesse Prinz (2016). Sentimentalist accounts and feature-based accounts, though in principle distinct, are conceptually linked, because our emotions are often what spur us to respond to the morally relevant features of a situation in the first place. Therefore, it is unsurprising that both views share an openness to non-human moral agency. Yet sentimentalists also have unique reasons to be open to the prospect of the moral animal, for, as we saw in the previous section, some neuroscientists contest that the emotions of animals, and particularly those of primates, likely have a great deal of overlap with our own. If this is true, then it might sometimes be the very same emotions that motivate humans and animals to act.

To sum up, the proponent of animal morality is restricted to feature-based or sentimentalist accounts of moral responsiveness, while denied the more robust rich moral responsiveness account.

3: Animals and Moral Judgment

So far, I have established that the existence of the moral animal is a hypothesis that the scientific community considers a live option and has investigated accordingly. Yet I have also shown how it is a hypothesis that greatly constrains one's general understanding of what constitutes responding to moral reasons. In this section, I will argue that if moral animals do exist, they are likely to be severely limited in their ability to morally evaluate one another, if they are even capable of doing so at all. My further claim is that the severity of these deficits is such that impartial human observers would generally be more competent to see to matters of animal desert than the moral animals themselves. This will prepare the way for my arguments in the following sections that if moral animals exist, then such human observers, due to their greater competence, have a strong reason and perhaps even an obligation to play this unusual role.

As human beings, we enjoy not only the capacity to act for moral reasons, but also the capacity to judge the actions of actions as moral or immoral. Although they are closely related, these are competencies that can come apart. In some respects, the latter capacity is more cognitively demanding. After all, judging others arguably requires a theory of mind that recognizes not only that another person is an entity that takes in the world from its own unique perspective – and to whom we might incur a moral obligation – but also that they are the kind of entity that can incur moral obligations of their own. In other words, even if we can make a good case for Brownie being a moral actor, we cannot assume without argument that he is also capable of judging the actions of his peers at the dog park. It may

well be that Brownie and creatures like him are wholly incapable of moral evaluation. And if moral animals cannot morally evaluate, then they cannot hold one another to account.

This is a little too quick. Although we cannot assume all animal moral actors are moral evaluators, that is not to say that no argument *can* be made for that further achievement. The ethologist, for instance, might point out that to enforce the norms of their group, social animals often exercise punishment, or 'negative reciprocity', as it is sometimes called (Clutton-Brock and Parker 1995). Canids socially ostracize or otherwise reprimand those who don't 'play fair' when play-fighting by refusing to wrestle with them in the future (Bekoff 2004); chimpanzees retaliate against food thieves (Jensen et al 2007); and marmosets who have watched humans make social exchanges accept food less frequently from those they've observed to be non-reciprocators (Kawai et al 2013). If we grant that behaviors like these are good evidence that some moral animals hold their peers to a standard of moral conduct, do we preempt any argument that we ought to be the ones punishing animals? Unfortunately not. For given their cognitive limitations, such animals are unlikely to be *effective* moral evaluators.

It is easy to see why this is the case. As I said above, the capacity to morally *judge* is more sophisticated than the capacity to morally *act*, probably even requiring at least a rudimentary theory of mind. Even if were to grant theory of mind, the further informational demands of moral evaluation are great. Often, 'internal' factors such as an agent's intentions or beliefs about the situation are relevant to whether we find them worthy of blame or praise. In our Brownie example, for instance, it made a difference whether Brownie comforted his upset owner because he simply wanted to soothe or whether he did so because he, more self-interestedly, wanted to make them a more alert and capable leader.

Many believe that only in the former case is Brownie genuinely praiseworthy. Accordingly, when an animal judges the actions of a conspecific, they may well need to infer facts about that creature's motives or epistemic situation. For creatures of limited intelligence, this will be a difficult task.

Indeed, there are some relevant factors that even very smart animals will have difficulty tracking. Consider, for instance, that false beliefs sometimes excuse one from being fully blameworthy for a given harm: if I had every reason to believe I was giving you medicine rather than poison, I am not culpable for poisoning you. Yet, to single out an animal thought to be among the most sophisticated social cognizers, chimpanzees, by all experimental indications, are incapable of attributing false beliefs to others, notwithstanding whatever other theory of mind abilities they might have (Call and Tomasello 2008, Call and Tomasello 1999, Hare et al 2001, Krachun et al 2010). So, even if a chimpanzee *could* attribute blame, he would likely not be able to take false beliefs into account as a mitigating or exculpatory factor.

There are a range of other cognitive thresholds a competent moral assessor must meet. To name but a few, judgment often involves attending to the actions of others without distraction, remembering past events and histories of behavior, differentiating an unprovoked outburst from one that was given cause. All of these present challenges to members of the animal kingdom, who typically fall short on one or other of these dimensions. Even if there were animals that are capable of moral judgment, then, they would almost certainly be deeply flawed evaluators. Indeed, they could arguably do no better than a human observer with even a rudimentary grasp of their species' behavioral repertoire. Even very young humans, for example, are adept at attributing false beliefs.

Furthermore, generations of successful fieldwork in comparative cognition prove that human observers are well able to observe animal groups, differentiate their members, and systematize and ultimately interpret their behaviors.

These epistemic and cognitive limitations are far from the only reasons humans would fare better than the animals themselves in identifying instances of animal wrongdoing, exculpatory or aggravating circumstances, and so forth. There are also significant practical constraints on the ability of animals to punish one another. Because punishers incur risk, we should expect the circumstances in which animals are *willing* to punish instances of wrongdoing to be limited, and empirical work confirms this prediction. Recall that the chimpanzee was cited above as an animal the ethologist might argue does mete out punishments within the group, at least where thievery is concerned. Even they, it seems, do not mete out *third-party* punishments for this crime. Although they will sometimes aggress against those whom they catch stealing their own food, Riedl and collaborators (2012) have shown that they are unwilling to act when the stolen food belongs to another chimpanzee. Such punishments, then, are personal rather than impartial.

Furthermore, even when their own food has been stolen, chimpanzees are significantly less likely to retaliate when the thief is a dominant individual, presumably because of the risk they would incur by doing so; whereas, when dominant individuals are the victims of food theft, they tend to retaliate swiftly (Riedl et al 2012). This is not the only way animal punishment is likely to go awry where dominance hierarchies are concerned. Some species, such as mandrills (Cheney et al 1986) and Japanese macaques (Aureli et al 1992), practice a form of redirected aggression, whereby instead of punishing

a more dominant individual for a perceived slight, the wronged subordinate attacks the wrongdoer's more vulnerable kin.

So even if animals do sometimes have a disposition to punish, it remains the case that they won't consistently punish whom they should, or will punish in an otherwise inappropriate manner. In these cases, too, an impartial human observer who is not at the mercy of an animal group's dominance hierarchies might well be better suited to punish its wrongdoers than the animals themselves are. Impartiality, to the extent that it can be achieved, has long been taken to be one of the most fundamental desiderata of adjudicative processes. Indeed, the putative impartiality of the human criminal justice system is one of the strongest justifications for its existence.

If moral animals were to exist, then, and humans were best equipped to morally evaluate them, would we be obligated or at least given strong reason to punish animal wrongdoers? One might think we surely cannot be so obligated, that it would be inappropriate to attempt anything of the kind. In many respects, the prospect of actually discharging such an obligation is absurd. It calls to mind images from the movie *The Advocate*, in which Colin Firth plays a public defender in 15th century France whose client is a pig accused of murder.[12] Can we possibly have reason to fastidiously monitor the social world of animals, occasionally stepping in to, say, dethrone a particularly sadistic

[12] Such historical 'animal trials' are a curious but well-documented phenomenon in European history, mostly occurring between the 12th and 18th centuries. As one might expect, many of the trials were religious in nature and concerned with alleged instances demonic possession (Girgen 2003). Others, however, featured animals accused of less supernatural offenses, such as the criminal destruction of property or participation in acts of bestiality (Girgen 2003; Srivastava 2007).

chimpanzee alpha or to save from retaliation a low-ranking member of his tribe whom he mistakenly believes to have slighted him in some way? Despite commonsense intuitions to the contrary, I will argue that if one accepts moral animals, then the answer to this question is *yes*, first by the lights of the retributivist, and then, even more plausibly, by the lights of the deterrence theorist. In this respect, the believer in animal morality may be biting off more than they are prepared to chew.

4 Retributivism and Animal Punishment

In this section, I argue that you cannot both accept moral animals and retributivist views in the philosophy of punishment without also recognizing strong reasons to punish animal wrongdoers.

Retributivism is the view that punishing wrongdoers is intrinsically good, regardless of any extrinsic benefit it confers, such as deterrence. On strong versions of the view, we are invested not only with the right but also the duty to punish wrongdoers. Depending on the extent to which animals are thought to attain standing as moral agents, then, we might expect strong retributivists to hold that humans are obligated to play the role of the world's policeman in the animal kingdom if they are (within reason) able to do so. Weaker forms of retributivism, which acknowledge a reason but not an obligation to punish the guilty, may sometimes escape incurring the obligation to punish animals, but can still be expected to concede that it is often permissible to do so, depending on the circumstances and what other reasons are in play.

Just how many retributivists fall within the strong as opposed to the weak camp is a matter of some debate, but it is likely that a great deal of them do, perhaps even the

majority. For example, if Michael Moore is right, there is a sense in which *all* forms of retributivism entail not merely a right to punish but also an obligation to do so. Moore writes:

> Retributivism is a very straightforward theory of punishment: We are justified in punishing because and only because offenders deserve it. Moral responsibility ('desert') in such a view is not only necessary for justified punishment, it is sufficient. Such sufficiency of justification gives society more than merely a *right* to punish culpable offenders. … For a retributivist, the moral responsibility of an offender also gives society the *duty* to punish. (Moore 2010)

Moore is certainly painting with a broad brush, and not all retributivists will assent to his characterization of their position. Still, many retributivists *do* explicitly endorse the form of reasoning Moore sets out and situate themselves in the strong retributivist camp. Take Immanuel Kant, for instance, who argues that

> [e]ven if a civil society were to be dissolved by the consent of all its members (e.g., if a people inhabiting an island decided to separate and disperse throughout the world), the last murderer remaining in prison would first have to be executed, so that each has done to him what his deeds deserve and blood guilt does not cling to the people for not having insisted on his punishment; for otherwise the people can be regarded as collaborators in this public violation of justice. (Kant 1797)

Retributivism goes hand in hand with the view that there is a sense in which all instances of moral wrongdoing, both public and private, merit criminalization, though the question of whether we should in fact criminalize a given type of wrongdoing is informed by further considerations, both principled and pragmatic (Duff 2014). One might expect these further considerations, especially those of principle, to adjudicate against animal punishment and thus let the retributivist off the hook—but let us see how far we can go.

As it turns out, it is surprisingly difficult for the retributivist to shirk this putative responsibility for any reason other than the pragmatic one of costs. A more principled hall pass is difficult to secure.

Perhaps the most obvious candidate reason that the retributivist might be able to refrain from dispensing animal punishments on principled grounds concerns jurisdiction. Even if animal punishment is a good, the retributivist might argue, surely humans do not have the jurisdiction to mete it out. This sounds very plausible on its face. With the possible exception of certain domesticated animals, human life proceeds largely at a remove from the lives of the animals who are candidates for punishment. Historically, we have not participated in their systems of punishment and reward, nor have they participated in ours. Thus, there is a sense in which alleged wrongdoing in the animal kingdom seems to be 'none of our business'—it is regrettable, perhaps, but not something we have the authority to address, given that we are in no sense members of their communities. For a human interloper to hold an animal to account would be to overstep or interfere somehow, the intuition goes – similar to if extraterrestrials were to suddenly alight on Earth and attempt to try us for our own misdeeds.

This turns out to be a tricky response for the retributivist to retreat towards. This is because retributivists tend to take a straightforwardly instrumentalist approach to the question of *who* should punish wrongdoers, holding that punishment should be meted out however is most effective. Indeed, similar approaches are more or less conventional in the philosophy of punishment in general, and not just within the retributivist literature. As it manifests in the retributivist framework, however, the instrumentalist argument runs as follows. Given that the punishment of wrongdoers is a *prima facie* good, its value is not

contingent on who realizes that good. Wrongdoers should therefore be punished by whoever is best positioned to punish them, i.e., whoever can deliver that punishment most effectively and efficiently. In civil society, this will often be the state, but it needn't be.

John Locke is perhaps the earliest philosopher to explicitly subscribe to this instrumentalist line of reasoning. When it comes to punishing wrongdoers, Locke (2014/1689) held that "every man hath a right to punish the offender, and be executioner of the law of nature." For Locke, whatever special authority and jurisdiction a state has to punish its subjects is grounded *only* in its ability to "restrain the violence and partiality of men", i.e. to dispense proportionate punishment more reliably and with less bias than can individuals who independently hold each other accountable (Locke 2014/1689). Andrew von Hirsch expresses a similar sentiment in *Doing Justice* (1976): "[T]here will be less social disruption," he writes, "if offenders are punished by the state rather than left to private retaliation."

Not only does this argument tend to favor state arbiters of punishment to their vigilante counterparts, it also tends to support the restriction of a state's jurisdiction to wrongdoings that occur within its own borders, again purely on instrumental grounds. For the instrumentalist, there is no *principled* reason why the government of the United States is entitled to mete out punishment to its own residents but not to those of the United Kingdom, Austria, or Zaire. Rather, there is only the pragmatic reason that, in our world of sovereign states, wrongdoers are typically punished more effectively and efficiently when countries 'police their own'.

Of course, by their very nature, instrumentalist restrictions on jurisdiction are sound only insofar as things *do* go better when justice is dispensed officially by a state rather than

informally by its citizens, or when it is dispensed autonomously within states rather than internationally. If a country were to fail to dispense justice to its people, for example, this might constitute a reason for some other state or group of states to take up that burden.[13]

Similarly, if a group of people existed outside the jurisdictional authority of any state and relied on a system of vigilante justice inferior to that which a state could provide, neighboring states might have sufficient moral reason to extend their jurisdictions to encompass members of that group. This, I argue, is more or less the situation in the case of animals. As I have discussed, even insofar as animals might be moral actors, they would be extremely limited in their abilities to reliably call wrongdoers to account. Furthermore, animals are often brought low in their attempts to sanction one another by what Locke would call their innate "violence and partiality" (Locke 2014/1689): the chimpanzees discussed earlier who punished only *subordinate* food thieves are but one example of this phenomenon. Humans have the advantage of not being embedded within the animal dominance hierarchies that make retribution difficult to exact against high-status individuals. Indeed, one can take as a proof of concept the ways in which humans benefit from official systems of law and order that are designed to curtail our own baser instincts.

Thus, if animals *are* moral actors deserving of punishment, those retributivists who are instrumentalists about jurisdiction (as most of them are) ought to recognize a duty to step in to punish animal wrongdoers when doing so is more likely than non-intervention to implement a principle of retributive justice. This is of course only a prima facie duty—

[13] One can see this reasoning at work in the Rome Statute of the International Criminal Court: Article 17 states that a case is inadmissible if it is already "being investigated or prosecuted by a State which has jurisdiction over it, *unless the State is unwilling or unable genuinely to carry out the investigation or prosecution*" (emphasis mine) (Rome Statute 1998).

there are a number of possible pragmatic reasons to leave animals to their own devices. However, if the only reasons not to punish moral animals are pragmatic, that is still an extremely interesting result, as, presumably, those pragmatic reasons won't always be decisive.

Are there other principled grounds besides that of jurisdiction on which the retributivist might resist incurring the obligation to punish animals? One common proposal about moral responsibility is that an agent must possess certain *capacities* in order to be morally responsible (or at least in order to be liable for punishment, should liability and moral responsibility come apart) (See, e.g., Wallace 1994 and Fischer 1999). Which capacities are named as the relevant ones varies across accounts, but the capacity that is most often singled out is the capacity to respond to moral reasons in a way that is consistent (Wallace 1994; Fischer 1999). Someone who believes in the existence of moral animals but wants to avoid punishing them might argue, then, that while some animals are genuine moral actors, they are too hit-or-miss when it comes to responding to moral reasons to lay claim to the *consistency* that is required to be fully possessed of the relevant capacity.

This response raises some interesting questions, although I am skeptical that it can easily succeed. For the sake of argument, I will grant the assumption that any capacity to respond to moral reasons that an animal might possess is likely to be a fairly local, contingent capacity. An animal that is motivated to prevent another animal from being hurt may not be motivated to prevent it from being stolen from, or even painlessly killed; alternately, an animal that is motivated to prevent all of these things normally may fail to be so motivated when they are themselves hungry or distracted. Furthermore, as discussed

previously, animals might be more consistent at moderating their own behavior than sanctioning the behavior of others, for various epistemic and pragmatic reasons.

But while I grant that animals' capacities for responding to moral reasons are plausibly local, I reject the premise that a global capacity to respond to moral reasons is necessary for punishment to be justified. When circumstances *do* permit the exercise of an animal's (local) moral capacity, it follows that animal should be held liable for failing to act as they ought. This is in line with how we think about human agency and responsibility. A kleptomaniac, for example, might be unable to consistently respond to reasons not to steal. Nonetheless, she may be able to consistently respond to reasons generated by the welfare of other, and for that reason can be held responsible for striking the mall cop who catches her stealing. To the extent that we can single out a range of circumstances within which an animal is responsive to moral reasons, then within that range one is licensed to punish the animal's performance accordingly.

What of animals' other shortcomings? A retributivist might have more luck avoiding the putative duty to punish animals by invoking certain strategies that have been used to shield young children and the criminally insane from retributivist punishment. This strain of response is in keeping with the standard retributivist literature, which typically endorses a relatively high punishment liability threshold, one that children are often thought to be incapable of meeting because of low intelligence, irrationality, or poor impulse control. Yet even so, here it can be argued that if a creature truly does have the capacity to respond to moral reasons, as the proponent of animal morality claims, then these other factors are insufficient to fully excuse that creature's wrongful deeds, though they can certainly be mitigating. This position is in parallel with that of those courts and

legal scholars who have argued that diminished rationality only mitigates or partially excuses criminal culpability in the case of human offenders (e.g., *Atkins v. Virginia* 2002, Morse 2003, Steinberg and Scott 2003). See also Cynthia Ward's less mainstream but thought-provoking arguments that even children as young as six can sometimes meet the standard of *mens rea* when they act deliberately and with knowledge of the inflicted harm (Ward 2006). Depending on where one stands on these matters, it may be difficult for the retributivist to secure animals a complete reprieve from punishment on these grounds.

On a further note, pursuant to the comparison between animals and young children, it may be self-undermining to both grant that animals are moral actors and to exculpate them from wrongdoing on the basis of their diminished mental capacities. In the 1970s and '80s, an active 'children's rights movement' sought and won legal victories for minors, such as the right to abortion without parental consent. Part of their strategy was to draw on work in psychology and the social sciences that showed adolescents were less impulsive and more rational than previously assumed. Their successes, however, had unexpected consequences—the same research they relied on was soon used to argue on behalf of stricter punishments for children, and even for the abolishment of a separate juvenile justice system. As Elizabeth Cauffman and colleagues write, the children's rights movement's appeal to the rational agency of the young proved a "double-edged sword", for "[h]ow could adolescents be mature enough to make their own decisions about abortion, but not mature enough to face the consequences of committing armed robbery or using marijuana?" (Cauffman et al 1999).[14] Similarly, those who want to argue that animals can

[14] This historical background, and the quotation from Cauffman et al, is drawn from Ward 2006, where a longer discussion can be found on the interesting dilemmas faced by the children's rights movement.

respond to moral reasons but are not morally responsible may not be able to have their cake and eat it, too. Research showing animals to be capable of genuine altruism or malice, especially when coupled with recent, more general work on animal rationality and decision-making, might just as easily be repurposed to argue that those animals are liable for the results of their actions.

5 Deterrence and Animal Punishment

If even the retributivist believer in the moral animal has reason to acknowledge a duty to punish animal wrongdoers, then the case for the deterrence theorist is stronger still. Deterrence theory is the view that punishment is permissible in light of its extrinsic benefits, most prominently its ability to deter further wrongdoing, either by the particular wrongdoer who is punished (individual deterrence), or by the general population who witnesses his punishment (general deterrence).[15] Because deterrence theory is fundamentally a consequentialist view, it is less concerned than retributivism with the culpability of the punished. Some proponents even go so far as to argue that punishing a known innocent is justified when doing so is truly what produces the greatest amount of wellbeing (e.g., Bagaric and Amaraskekara 2000).

When it comes to animal punishment, the deterrence theorist is not burdened by the retributivist's worries about whether animals enjoy full and informed moral autonomy over their actions, or whether humans have the jurisdiction to intervene in their affairs. Even if animals were morally insensitive, the deterrence theorist could justify punishing them

[15] Note, however, that deterrence is not the only extrinsic benefit of punishment, though it may be the most prominent. Other benefits include the incapacitation or reform of wrongdoers, as well as whatever benefits victims enjoy in seeing their assailants punished.

purely in virtue of the consequent reduction in the overall level of animal suffering. In this respect, the deterrence theorist has no *in principle* misgivings about animal punishment. For them, whether we ought to punish animals is a straightforwardly empirical question.

How should deterrence theorists empirically adjudicate the question of animal punishment? How can they know whether intervention will have the desired effect of reducing animal wrongdoing and hence animal suffering? Even when it comes to punishing other humans, the data on whether and which punishments are effective is complex and contradictory. As the authors of one meta-analysis warn, "[d]espite the apparent simplicity of the theoretical framework offered by deterrence theory, determining its empirical validity has not been so easy" (Pratt, Cullen, et al 2006). Is the situation more difficult when we must decide whether to punish a species removed from our own, whose behavior and motivations we may never fully understand? Or is it perhaps easier, given that misbehaving animals are less psychologically sophisticated than their human counterparts? At the very least, given the cognitive limitations of even very clever animals, there are special constraints on which punishments are likely to be efficacious. The deterrence theorist might expect that empirical considerations will dismiss animal punishment out of hand: Surely such punishments will never be effective enough to be justified! Yet I contend that this is not obviously the case. At the very least, substantiating it requires making some very specific commitments about animals' capabilities.

Deterrence comes in two varieties, individual and general. Individual deterrence concerns the individual being punished: Is *x*'s punishment likely to deter *x* from future wrongdoing? General deterrence concerns the effect of punishment on the broader community: Is the example of *x*'s punishment likely to deter the public at large from future

wrongdoing? These two aims of punishment are distinct and it might be the case that the human-administered punishment of moral animals would achieve both, neither, or just one of them.

First, I will consider whether punishing moral animals could achieve individual deterrence. Focusing on individual deterrence has the advantage that we needn't worry about whether animals can properly interpret and learn from the punishments others receive—all that matters is that they are able to learn from being punished themselves. Here, the research is clear that punishment can effectively shape animal behavior, with two important caveats. First, the punishment must be relatively consistent; second, it must be concurrent with or swiftly follow the offense (Mackintosh 1983, Lindsay 2000). Improperly administered animal punishments are associated with neurotic behaviors, increased levels of stress, and even the exacerbation of existing behavioral issues (Schalke et al 2007, McGreevy and McLean, 2009, Dale et al 2017). For this reason, when they are feasible, reward-based approaches to shaping animal behavior tend to be safer and more reliable than punitive ones. Nevertheless, sometimes punishment is the only available intervention. When an animal *does* commit a serious wrong, rendering reward-based approaches to good behavior unavailable, administering a punishment will have a greater effect on that animal's future behavior than doing nothing at all.

Can we also make a case for the efficacy of general deterrence, whereby an animal learns not only from punishment but also from witnessing the punishment of its peers? I think that we can. Let us look at what is required for general deterrence to succeed. First, for the offender's punishment to affect my future actions, I need to connect what the offender has done to what is being done to the offender. This doesn't have to be a conscious

or even a cognitive connection: simple associative mechanisms could do the necessary work here, and so this requirement, at least, can plausibly be met by an animal. Secondly, the offender's punishment needs to be understood as or linked with a negative outcome. If I experience distress merely by witnessing the offender's punishment (perhaps via emotional contagion), then I needn't be very cognitively sophisticated at all to meet this second requirement. Observing the offender's act of wrongdoing, and myself suffering when I observe his punishment, may suffice to steer me clear of following in his footsteps. Alternatively, perhaps I *don't* suffer when I observe the offender's punishment, but rather grasp that what is happening to him is undesirable (by hearing him scream, say) and grasp, too, that it is something that might also happen to me if I were to act as he did. At first glance, this second route to general deterrence seems to presume a theory of mind, as it appears to require the observer to understand that the offender is suffering. This need not be the case, however. [16] Suppose I observe one animal push another, and that I then observe the first animal being placed in a cage. Even if I do not realize that the caged animal has a mental life that is affected by being placed in a cage, I might form the connection between *X pushing Y* and *X's being placed in a cage*, realize that *I* would not want to be caged, and refrain from pushing other animals on that basis. So a case can be made that even general deterrence might prove effective with animals, if they are capable of the thought processes outlined above.

The takeaway here for the deterrence theorist is that animals can indeed be deterred by punishment, not only by being punished themselves but plausibly also by witnessing

[16] Note that even if theory of mind and causal cognition *are* required for one to be deterred by the punishment of others, some have argued that certain animals, like corvids and great apes, *do* possess those capacities.

the punishment of their peers. Of course, whether punishment is the most effective intervention—or even more effective than no intervention at all—will be highly variable across situations. A particularly recalcitrant bad actor may merit punitive killing or removal from his group, for the sake of his peers; the perpetrator of a wrong may merit punishment in the moment of his misdeed so as to best immediately curtail him; but a more borderline miscreant might be best responded to with some other, non-punitive action, or simply left alone. So, although the deterrence theorist won't be pressed to punish *every* case of animal wrongdoing, and will certainly be pressed to punish less than the retributivist might be, there is indeed a subset of cases in which punishment *is* the deterrence theorist's optimal course of action.

Clearly, then, the possibility that there are moral animals has an extremely peculiar consequence in the philosophy of punishment, namely the consequence that humans ought to sometimes occupy themselves with the business of punishing them. This appears to follow whether one is a retributivist or whether one favors the more consequentialist alternative of deterrence theory. What is a believer in animal morality to do, if they accept the arguments I have given? For one, they might try to convince their opponents that this consequence is *not* so unpalatable as it might appear. Indeed, there are those who would be all too happy to use it as a *reductio* against their belief in animal morality; a good argument for why it is not a *reductio*, then, would serve them well. I won't attempt to sketch out how this kind of response might go. I *do* think the claim that we ought to punish animals is both counterintuitive and unpalatable, and so I prefer simply to invite those with contrary intuitions to convince me otherwise.

Another option for the believer in animal morality is to show that, although they believe in animal morality, and do accept that many believers in animal morality indeed ought to punish animals, there is some special reason that they do not fall prey to this entailment themselves, one not discussed at in the previous sections. In the next section, I anticipate one such way in which the believer in animal morality might take this second approach, namely with an argument from moral relativism, and sketch why I think this argument does not work.

6 Can the Relativist Avoid Animal Punishment?

Here is one objection to the claim that the existence of moral animals entails an obligation to punish them. This objection begins with the thought that the moral norms of animals may be very different from our own, and concludes with the claim that such differences would release us from any obligation to punish animal "wrongdoers," and perhaps even prohibit us from doing so. We can call this the objection from relativism.

Relativism can take various different forms (see Gowans 2015 for a helpful overview). Some forms do not in fact pose a problem for the claim that we would have reason to punish moral animals, and so the objection from relativism is not an objection from all forms of relativism. To illustrate: On one version of moral relativism, which John Tilley calls *Appraiser Relativism* (Tilley 2000), the truth of a moral judgment depends upon the norms of the community to which the one making the judgment belongs.[17] According to Appraiser Relativism, the sentence "Slavery is wrong" is true when uttered by someone in 21st century America, but false when uttered in ancient Greece. If anything, Appraiser

[17] The name Appraiser Relativism, and that of its counterpart Agent Relativism, were adapted by Tilley from David Lyons' "appraiser's-group" and "agent's-group relativism" (Lyons 1976, Tilley 2000).

Relativism actually vindicates the notion that humans should punish animals they take to be immoral. For the ones making such judgments are human and, according to human standards, we should indeed punish those who act as the animals in question have acted.

The objection from relativism is grounded not in Appraiser Relativism, then, but rather in its contrast, *Agent Relativism* (Tilley 2000). According to Agent Relativism, it is not the moral norms of the judge that matter, but rather those of the agent being judged. The truth of moral judgments is grounded in the cultural norms not of the judge but rather of the agent who committed the action (Tilley 2000). If Agent Relativism is correct, then it is wrong for a person to impose the moral norms of her own community on those living in communities with differing norms, for the two communities have equally valid yet irreconcilable systems. So, for example, perhaps we should not attempt to stop some society from torturing their criminals, even if such punishments are unjust by our own society's standards. Or in the animal case, even if animals do turn out to be moral beings, we humans ought not to punish them, for the animals' moral communities adhere to different moral principles than our own. This is the objection from relativism.

I will not attempt to argue against Agent Relativism, on the truth of which the objection from relativism depends. For one, there is already work in the literature to this effect (Tilley 2000). But more importantly, I contend that the objection from relativism is flawed even if Agent Relativism is true.

I see two basic problems with the objection from relativism. First, the fact that we cannot punish immoral animals according to our own norms does not entail that we cannot punish them at all. For we may be able to identify some of the basic moral principles of

their societies, and then be fully licensed in bringing their wrongdoers to justice, even if their notions of wrongdoing or justice don't quite line up with our own. In section 3, I briefly discussed some observable patterns of negative reciprocity in the animal kingdom—wolves who won't play with unfair players, chimpanzees who punish theft, et cetera. Even if we can't ascertain an animal community's *full* set of moral norms, it seems plausible that observable behaviors like these could help us piece together some proper subset of those norms. If this is right, then the arguments of my paper would still apply for breaches of norms belonging to that subset. Ergo, the believer in animal morality who embraces Agent Relativism cannot fully shirk the arguments of this paper unless the moral norms of animals are truly so alien as to be fully beyond our discernment. Yet if the moral norms of animals are truly so alien as to be fully beyond our discernment, then the objection from relativism will *still* fail, as the second problem I raise will make clear.

The second problem with the objection from relativism is that it is self-undermining. Since the objection is intended to save the believer in animal morality from also incurring reason to punish bad animal actors, it takes as a premise the claim that some animals are moral. Section 1 of this paper outlined some reasons one might want to endorse this premise, yet all of those reasons are fundamentally grounded in the connection between animal behavior and human morality. For example, animals seem to punish behavior that is recognizably immoral – i.e. immoral *by our own lights*. Or, animals seem to endure pain in order to avoid causing harm to their peers, something that *we take to be good.* The assumption that these behaviors are evidence for animal morality is underwritten by the deeper assumption that if animals were moral, then their sense of morality would observably resemble our own. If the relativist accepts this deeper assumption, they can

justify their belief in the existence of moral animals by referencing the literature reviewed in Section 1, yet must then accept my arguments above that there is still animal conduct worthy of punishment. If the relativist rejects this deeper assumption, they can avoid my arguments above, but at the cost of now having no clear evidence to justify their prior belief in the existence of moral animals. At best, they now defend only an exotic counterfactual: *If* there were moral animals (whose morality would fail to line up with our own in any observable sense), *then* humans ought not punish them. Since we have no reason to believe the antecedent of that conditional, the objection from relativism turns out to be of limited interest.

7 Summing Up

I hope to have shown that a commitment to the existence of moral animals comes at a cost. First, there is the cost of adopting a particularly weak notion of responsiveness to moral reason. Second, there is either the cost of accepting that animal morality begets animal punishment, or the cost of avoiding that entailment. The latter involves having your choice of more general views in ethics and the philosophy of punishment greatly constrained. The consequentialist believer in animal morality decides whether to punish animals on a case-by-case basis according to pragmatic factors, and so may avoid having to punish many animal wrongdoers—but likely not all of them, and certainly not in principle. The retributivist believer in animal morality may partly rebuff animal punishment on more principled grounds, for instance by claiming that animals are inconsistent moral actors and therefore less blameworthy. However, that strategy is in tension with the claim that animals are wholeheartedly moral in the first place. Likewise, although a form of moral relativism

at first appeared to lift the duty to punish animal wrongdoers, it, too, undermined the most commonly cited reasons for taking animals to be moral.

If the literature concerning the moral animal is to develop, we must increasingly discuss not only whether some animals are moral but also what it would mean for human beings if they were. I hope this paper has made a start in that direction.

Bibliography

Andrews, K., & Gruen, L. (2014). Empathy in other apes. Empathy and Morality, 193.

Atkins v. Virginia. (2002), 260 Va. 375, 534S. E. 2d 312.

Audi, R. (2015). Reasons, Rights, and Values. Cambridge University Press.

Audi, R. (2013). Moral perception. Princeton University Press.

Aureli, F., Cozzolino, R., Cordischi, C., and Scucchi, S. (1992), 'Kin-oriented redirection among Japanese macaques — an expression of a revenge system?', Animal Behaviour, 44, pp. 283–291.

Bagaric, M., & Amaraskekara, K. (2000). The errors of retributivism. Melb. UL Rev., 24, 124.

Bar M. (2007). "The proactive brain: using analogies and associations to generate predictions." Trends Cognitive Science.

Barbur, J. L., Watson, J. D. G., Frackowiak, R. S., and Zeki, S. (1993). "Conscious visual perception without V1." Brain.

Bartal, I. B. A., Decety, J., & Mason, P. (2011). Empathy and pro-social behavior in rats. Science, 334(6061), 1427-1430.

Bekoff, M., & Pierce, J. (2009). Wild justice: The moral lives of animals. University of Chicago Press.

Bekoff, M. (2004). Wild justice and fair play: Cooperation, forgiveness, and morality in animals. Biology and Philosophy, 19(4), 489-520.

Bekoff, M. (2000). Animal Emotions: Exploring Passionate Natures. Current interdisciplinary research provides compelling evidence that many animals experience such emotions as joy, fear, love, despair, and grief—we are not alone. BioScience, 50(10), 861-870.

Berridge, K.C. (2003). Comparing the emotional brain of humans and other animals. In R.J. Davidson, K.R. Scherer, & H.H. Goldsmith (Eds.), Handbook of affective sciences (pp. 25–51). New York: Oxford University Press.

Bird, C. D., & Emery, N. J. (2009). Insightful problem solving and creative tool modification by captive nontool-using rooks. Proceedings of the National Academy of Sciences, 106(25), 10370-10375.

Bird, C. D., & Emery, N. J. (2009). Rooks use stones to raise the water level to reach a floating worm. Current Biology, 19(16), 1410-1414.

Bowden, E. M., Jung-Beeman, M., Fleck, J., & Kounios, J. (2005). New approaches to demystifying insight. Trends in cognitive sciences, 9(7), 322-328.

Brosnan, S. F., & De Waal, F. B. (2003). Monkeys reject unequal pay. Nature, 425(6955), 297-299.

Byrne, A. (2009). “Experience and content.” The Philosophical Quarterly.

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. Trends in cognitive sciences, 12(5), 187-192.

Call, J., & Tomasello, M. (1999). A nonverbal false belief task: The performance of children and great apes. Child development, 70(2), 381-395.

Cauffman, E., Woolard, J., & Reppucci, N. D. (1999). Justice for Juveniles: New Perspectives on Adolescents' Competence and Culpability, 18 Quinnipiac L. Rev, 403, 405.

Chronicle, E. P., Ormerod, T. C., & MacGregor, J. N. (2001). When insight just won't come: The failure of visual cues in the nine-dot problem. The Quarterly Journal of Experimental Psychology: Section A, 54(3), 903-919.

Church, R. M. (1959). Emotional reactions of rats to the pain of others. Journal of comparative and physiological psychology, 52(2), 132.

Clark, S. R. L. (1984). The nature of the beast: Are animals moral? New York, NY: Oxford University Press.

Clement, G. (2013). Animals and moral agency: The recent debate and its implications. Journal of Animal ethics, 3(1), 1-14.

Clutton-Brock, T. H., & Parker, G. A. (1995). Punishment in animal societies. Nature, 373(6511), 209-216.

Cook, R. G., & Fowler, C. (2014). “Insight” in pigeons: absence of means–end processing in displacement tests. Animal cognition, 17(2), 207-220.

Cowan, Robert (2015). “Perceptual Intuitionism”. Philosophy and Phenomenological Research. Dretske, F. I. (1995). "Meaningful perception." An Invitation to Cognitive Science: Visual Cognition.

Dale, A. R., Podlesnik, C. A., & Elliffe, D. (2017). Evaluation of an aversion-based program designed to reduce predation of native birds by dogs: An analysis of training records for 1156 dogs. Applied Animal Behaviour Science, 191, 59-66.

DeGrazia, D. (1996). Taking animals seriously: mental life and moral status. Cambridge University Press.

DeGrazia, D. (1997). Great apes, dolphins, and the concept of personhood. The Southern journal of philosophy, 35(3), 301-320.

De Waal, F. (2009). Primates and philosophers: How morality evolved: How morality evolved. Princeton University Press

De Waal, F. B. (1996). Good natured (No. 87). Harvard University Press.

Dixon, B. A. (2008). Animals, emotion, and morality: Marking the boundary. Amherst, NY: Prometheus Books.

Duff, R. A. (2014) "Towards a modest legal moralism." Criminal Law and Philosophy 8.1: 217-235.

Emery, N. J. (2013). 4 Insight, imagination and invention: Tool understanding in corvid. Tool use in animals: Cognition and ecology, 67.

Epstein R, Kirshnit CE, Lanza RP, Rubin LC. (1984). 'Insight' in the pigeon: antecedents and determinants of an intelligent performance. Nature.

Faraci, D. (2014). "A hard look at moral perception." Philosophical Studies.

Finke, R. A., & Slayton, K. (1988). Explorations of creative visual synthesis in mental imagery. Memory & cognition, 16(3), 252-257.

Fischer, J. M. (1999). Recent work on moral responsibility. Ethics, 110(1), 93-139.

Fitzpatrick, S. (2017). Animal morality: What is the debate about?. Biology & Philosophy, 1-33.

Flack, J. C., & De Waal, F. B. (2000). 'Any animal whatever'. Darwinian building blocks of morality in monkeys and apes. Journal of Consciousness Studies, 7(1-2), 1-29.

Foerder, P., Galloway, M., Barthel, T., Moore III, D. E., & Reiss, D. (2011). Insightful problem solving in an Asian elephant. PloS one, 6(8).

Friston K., and Kiebel S. (2009). "Predictive coding under the free-energy principle." Philosophical Transactions of the Royal Society: Biological Sciences.

Girgen, J. (2003). The Historical and Contemporary Prosecution and Punishment of Animals. Animal Law Review at Lewis & Clark Law School.

Gowans, C. (2004). Moral relativism. Stanford Encyclopedia of Philosophy.

Gruber, R., Schiestl, M., Boeckle, M., Frohnwieser, A., Miller, R., Gray, R. D., ... & Taylor, A. H. (2019). New Caledonian crows use mental representations to solve metatool problems. Current Biology, 29(4), 686-692.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychological review, 108(4), 814.

Hanus, D., Mendes, N., Tennie, C., & Call, J. (2011). Comparing the performances of apes (Gorilla gorilla, Pan troglodytes, Pongo pygmaeus) and human children (Homo sapiens) in the floating peanut task. PloS one, 6(6).

Harman, E. (2015). The irrelevance of moral uncertainty. Oxford studies in metaethics, 10, 53-79.

Harman, E. (2011). Does moral ignorance exculpate? Ratio, 24(4), 443-468.

Heinrich, B. (1995). An experimental investigation of insight in common ravens (Corvus corax). The Auk, 112(4), 994-1003.

Huber, L., & Gajdon, G. K. (2006). Technical intelligence in animals: the kea model. Animal cognition, 9(4), 295-305.

Hume, D. (2014). A treatise of human nature. Simon and Schuster.

Jensen, K., Call, J., & Tomasello, M. (2007). Chimpanzees are vengeful but not spiteful. Proceedings of the National Academy of Sciences, 104(32), 13046-13050.

Kant, I. (1797), The Metaphysics of Morals, M. Gregor (trans.), Cambridge University Press (1991).

Kant, I., Wood, A. W., & Schneewind, J. B. (2002). Groundwork for the Metaphysics of Morals. Yale University Press.

Kawai, N., Yasue, M., Banno, T., & Ichinohe, N. (2014). Marmoset monkeys evaluate third-party reciprocity. Biology letters, 10(5), 20140058.

Kershaw, T. C., & Ohlsson, S. (2004). Multiple causes of difficulty in insight: the case of the nine-dot problem. Journal of experimental psychology: learning, memory, and cognition, 30(1), 3.

Köhler, W. (2013). Intelligenzprüfungen an Menschenaffen: Mit einem Anhang: Zur Psychologie des Schimpansen (Vol. 134). Springer-Verlag.

Korka, B., Schröger, E., & Widmann, A. (2019). Action Intention-based and Stimulus Regularity-based Predictions: Same or Different? Journal of Cognitive Neuroscience.

Korsgaard, C. (2006). Morality and the distinctiveness of human action. Primates and philosophers: How morality evolved, 98-119. Chicago

Krasheninnikova, A. (2018). Means-End Reasoning. In Encyclopedia of Animal Cognition and Behavior (pp. 1-6). Springer, Cham.

Lakshminarayanan, V. R., & Santos, L. R. (2008). Capuchin monkeys are sensitive to others' welfare. Current Biology, 18(21), R999-R1000.

Lin, W. L., & Lien, Y. W. (2013). The different role of working memory in open-ended versus closed-ended creative problem solving: a dual-process theory account. Creativity Research Journal, 25(1), 85-96.

Lindsay, S.R., 2000. Adaptation and learning. In: Handbook of Applied Dog Behaviour and Training., Vol. 1. Iowa State University Press, Ames, IA, pp. 298-305.

Locke, J. (2014). Second Treatise of Government: An Essay Concerning the True Original, Extent and End of Civil Government. John Wiley & Sons.

David Lyons, “Ethical Relativism and the Problem of Incoherence,” Ethics 86(2) (1976): 107–21.

MacGregor, J. N., Ormerod, T. C., & Chronicle, E. P. (2001). Information processing and insight: a process model of performance on the nine-dot and related problems. Journal of Experimental Psychology: Learning, Memory, and Cognition, 27(1), 176.

Mackintosh N. J. Conditioning and Associative Learning. Oxford Univ. Press, New York 1983).

Mandelbaum, M. (1955). “Societal facts.” The British Journal of Sociology. McHugh, C. and Way, J. (2016). “Fittingness First.” Ethics.

Markovits, J. (2010). “Acting for the Right Reasons.” Philosophical Review, Vol. 119, No. 2: 201-242.

Masserman, J. H., Wechkin, S., & Terris, W. (1964). "Altruistic" behavior in rhesus monkeys. The American journal of psychiatry.

McDowell, J. (2009). Conceptual capacities in perception. Having the world in view: essays on Kant, Hegel, and Sellars, 127-44.

McDowell, J. (1979). Virtue and reason. The monist, 62(3), 331-350.

McGreevy, P. D., & McLean, A. N. (2009). Punishment in horse-training and the concept of ethical equitation. Journal of Veterinary Behavior: Clinical Applications and Research, 4(5), 193-197.

Mednick, S. (1962). The associative basis of the creative process. Psychological review, 69(3), 220.

Mendes, N., Hanus, D., & Call, J. (2007). Raising the level: orangutans use water as a tool. Biology letters, 3(5), 453-455.

Mole, C. (2015). J. Zembeikis and T. Raftopoulos (eds.) The Cognitive Penetrability of Perception. Oxford University Press.

Monsó, S. (2015). Empathy and morality in behaviour readers. Biology and Philosophy, 30, 671-690.

Moore, M. S. (2010). Placing blame: A general theory of the criminal law. Oxford University Press.

Moreau, P., Jolicœur, P., & Peretz, I. (2009). “Automatic brain responses to pitch changes in congenital amusia.” Annals of the New York Academy of Sciences.

Morgan, C. L. (1894). An introduction to comparative psychology. London: The Walter Scott Publishing Co.

Morse, S. J. (2003). Diminished rationality, diminished responsibility. Ohio St. J. Crim. L., 1, 289.

Musschenga, B. (2013). Animal morality and human morality. In B. Musschenga and A. van Harskamp (eds.), What Makes us Moral?: On The Capacities and Conditions for Being Moral. Dordrecht: Springer.

Näätänen, R., Jacobsen, T., and Winkler, I. (2005). “Memory-based or afferent processes in mismatch negativity (MMN): A review of the evidence.” Psychophysiology.

Näätänen, R., and Gaillard, A. W. K. (1983). “The orienting reflex and the N2 deflection of the event-related potential (ERP).” Advances in psychology.

Nanay, B. (2011). “Do we see apples as edible?” Pacific Philosophical Quarterly.

Neilands, P. D., Jelbert, S. A., Breen, A. J., Schiestl, M., & Taylor, A. H. (2016). How insightful is 'insight'? New Caledonian crows do not attend to object weight during spontaneous stone dropping. PloS one, 11(12).

Nelkin, D. K. (2011). Making sense of freedom and responsibility. Oxford University Press.

Paavilainen, P. (2013). "The mismatch-negativity (MMN) component of the auditory event-related potential to violations of abstract regularities: a review". International journal of psychophysiology.

Panksepp, J. (1998). Affective neuroscience: The foundations of human and animal emotions. Oxford university press.

Pearce, J. M. S. (2005). "Selected observations on amusia." European Neurology.

Peterson, D. 2011. The moral lives of animals. New York: Bloomsbury Press.

Piaget, J., & Cook, M. (1952). The origins of intelligence in children. New York: International Universities Press

Pieszek M., Widmann A., Gruber T., and Schröger E. (2013). "The human brain maintains contradictory and redundant auditory sensory predictions." PLoS ONE.

Pieszek, M., Schröger, E., and Widmann, A. (2014). "Separate and concurrent symbolic predictions of sound features are processed differently." Frontiers in psychology.

Evelyn Pluhar, Beyond Prejudice: The Moral Significance of Human and Nonhuman Animals (Durham, NC: Duke University Press, 1995).

Pratt, T. C., Cullen, F. T., Blevins, K. R., Daigle, L. E., & Madensen, T. D. (2006). The empirical status of deterrence theory: A meta-analysis. Taking stock: The status of criminological theory, 15, 367-396.

Preston, S. D., & De Waal, F. B. (2002a). Empathy: Its ultimate and proximate bases. Behavioral and brain sciences, 25(1), 1-20.

Preston, S. D., & de Waal, F. B. (2002b). The communication of emotions and the possibility of empathy in animals. Altruistic love: Science, philosophy, and religion in dialogue, 284-308.

Prinz, J. (2016). Sentimentalism and the Moral Brain. Moral Brains: The Neuroscience of Morality, 45.

Redish, A. D. (2016). Vicarious trial and error. Nature Reviews Neuroscience, 17(3), 147.

Renner, E., Abramo, A. M., Hambright, M. K., & Phillips, K. A. (2017). Insightful problem solving and emulation in brown capuchin monkeys. Animal cognition, 20(3), 531-536.

Riedl, Katrin, et al. "No third-party punishment in chimpanzees." Proceedings of the National Academy of Sciences 109.37 (2012): 14824-14829.

Rome Statute of the International Criminal Court. (1998). UN Doc. A/CONF. 183/9; 37 ILM 1002 (1998); 2187 UNTS 90.

Rowlands, M. (2013). Animals and moral motivation: A response to Clement. Journal of Animal Ethics, 3(1), 15-24.

Rowlands, M. (2012). Can animals be moral?. Oxford University Press.

Rowlands, M. (2011). Animals that act for moral reasons. The Oxford handbook of animal ethics, 519-546.

Sanz, C. M., Call, J., & Boesch, C. (Eds.). (2013). Tool use in animals: cognition and ecology. Cambridge University Press.

Sapontzis, S. F. (1987). Morals, reason, and animals. Philadelphia, PA: Temple University Press.

Schalke, E., Stichnoth, J., Ott, S., & Jones-Baade, R. (2007). Clinical signs caused by the use of electric training collars on dogs in everyday life situations. Applied Animal Behaviour Science, 105(4), 369-380.

Schino, G., & Aureli, F. (2009). Reciprocal altruism in primates: partner choice, cognition, and emotions. Advances in the Study of Behavior, 39, 45-69.

Shapiro, P. (2006). Moral agency in other animals. Theoretical Medicine and Bioethics, 27(4), 357-373.

Shettleworth, S. J. (2009). Animal cognition: deconstructing avian insight. Current Biology, 19(22), R1039-R1040.

Shettleworth, Sara J. (2012) "Do animals have insight, and what is insight anyway?." Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale.

Siegel, S. (2006). “Which Properties are Represented in Perception?” in T. Gendler and J. Hawthorne (eds) Perceptual Experience. Oxford University Press.

Siegel, S. (2007) "How can we discover the contents of experience?" The Southern Journal of Philosophy.

Siegel, S. (2009), “The Visual Experience of Causation.” The Philosophical Quarterly.

Siegel, S. (2011). The contents of visual experience. Oxford University Press. Tye, M. (1995). Ten Problems of Consciousness. MIT Press.

Silk, J. B. (2007). Empathy, sympathy, and prosocial preferences in primates. Oxford handbook of evolutionary psychology. Oxford University Press, New York, 115-126.

Smith, A. (1817 [1759]). The theory of moral sentiments (Vols. I & II). Boston: Wells and Lilly.

Sommerville, J. A., & Woodward, A. L. (2005). Infants' sensitivity to the causal features of means-end support sequences in action and perception. Infancy, 8(2), 119-145.

Sowden, P. T., Pringle, A., & Gabora, L. (2015). The shifting sands of creative thinking: Connections to dual-process theory. Thinking & Reasoning, 21(1), 40-60.

Srivastava, Anila. (2007) Mean, dangerous, and uncontrollable beasts: Mediaeval Animal Trials. Mosaic: A Journal for the Interdisciplinary Study of Literature, Volume 40, issue 1, page 127.

Steinberg, L., & Scott, E. S. (2003). Less guilty by reason of adolescence: developmental immaturity, diminished responsibility, and the juvenile death penalty. American Psychologist, 58(12), 1009.

Sternberg, R. J., & Davidson, J. E. (1995). The nature of insight. The MIT Press.

Taylor, A. H., Knaebe, B., & Gray, R. D. (2012). An end to insight? New Caledonian crows can spontaneously solve problems without planning their actions. Proceedings of the Royal Society B: Biological Sciences, 279(1749), 4977-4981.

Thorpe, W. H. (1956). Learning and instinct in animals.

Tilley, J. J. (2000). Cultural relativism. Human Rights Quarterly, 22(2), 501-547.

Van Gulick, R. (1994). "Deficit studies and the function of phenomenal consciousness." In G. Graham and G. Lynn Stephens (eds.) Philosophical Psychopathology. MIT Press.

Von Hirsch, Andrew, Committee for the Study of Incarceration (Etats-Unis), and Willard Gaylin. "Doing justice: The choice of punishments." (1976): 19-35.

Wallace, R. J. (1994). Responsibility and the moral sentiments. Harvard University Press.

Ward, C. V. (2006). Punishing children in the criminal law. Notre Dame L. Rev., 82, 429.

Werner, P. forthcoming."Moral Perception and the Contents of Experience." Journal of Moral Philosophy.

Widmann, A., Gruber, T., Kujala, T., Tervaniemi, M., and Schröger, E. (2007). "Binding symbols and sounds: evidence from event-related oscillatory gamma-band activity." Cerebral Cortex.

Willatts P (1999) Development of means-end behavior in young infants: pulling a support to retrieve a distant object. Dev Psychol 35:651–667.

Wolf, S. (1987). Sanity and the Metaphysics of Responsibility. 1987, 46-62.