In Praise of Animals[1]

Rhys Borchert & Aliya R. Dewey

Department of Philosophy, University of Arizona

Reasons-responsive accounts of praiseworthiness say, roughly, that an agent is praiseworthy for an action just in case the reasons that explain why they acted are also the reasons that explain why the action is right. In this paper, we argue that reasons-responsive accounts imply that some actions of non-human animals are praiseworthy. Trying to exclude non-human animals, we argue, risks neglecting cases of inadvertent virtue in human action and undermining the anti-intellectualist commitments that are typically associated with reasons-responsive accounts. Of course, this could be taken as a reason to reject reasons-responsive accounts, rather than as a reason to attribute praiseworthiness to non-human animal action. We respond to two reasons that one might resist the implication that non-human animal action can be praiseworthy. The first appeals to intuition: it's too counterintuitive to attribute praiseworthiness to non-human animal action. In response, we argue that once the factors that determine an action's praiseworthiness are disambiguated from the factors that determine whether an agent should be praised, the intuitive objection loses much of its force. The second appeals to empirical evidence: attributing praiseworthiness to non-human animal action involves a problematic kind of anthropomorphizing. First, we point out that this objection is mostly an *a priori* objection in *a posteriori* clothes: whether we give anthropomorphic vs. anthropectic explanations is a methodological choice, not an empirical one. Second, we argue that considerations from the literature on rational analysis and radical interpretation actually support anthropomorphic explanations over anthropectic explanations.

Let us begin by telling you two true stories. In the first, a mother, Angel, woke up in the middle of the night to the smell of smoke. Her house was on fire, and she immediately rushed out with two of her children. Her third child, however, was still inside so she ran back into the flames to save her. She ultimately suffered third-degree burns over most of her body, but she successfully rescued all of her children.

In the second, a mother, Scarlett, rushes into a burning building to save her five children. By the time she gets all five to safety, her body is covered with burns, including burns to both of her eyelids preventing her from opening them. When the fire department arrives at the scene, they see her blindly tending to her children, making sure they are all safe before collapsing from exhaustion.

As described, we would expect that everyone would judge that both of these stories involve heroic action. The actions of Angel and Scarlett seem to be paradigmatic cases of moral praiseworthi-

[1] Penultimate draft of paper forthcoming in *Biology & Philosophy*. Please cite the published paper, not this draft.

ness.[2] However, nearly every philosopher in the history of the Western tradition, from antiquity to the present day, would maintain that Scarlett's action was not praiseworthy to any degree. This is due to a difference in the cases that we have not yet mentioned: Angel was a human being and Scarlett was a cat.

Why is Angel praiseworthy in the strongest sense, while Scarlett is not praiseworthy in any sense? The specific reason philosophers give for denying the praiseworthiness of non-human animals may differ. We are rational, they are not. We are self-conscious, they are not. We have language, they do not. We have morally relevant concepts, they do not. We have autonomy, they do not. The philosophy of moral responsibility typically takes it as an unquestioned assumption that only human beings can be praiseworthy.[3,4] So the background assumption for theorizing about praise is that any proposed property, or set of properties, sufficient for praiseworthiness must be something that we have but that non-human animals cannot have, and any property, or set of properties, that non-human animals could have cannot be sufficient for praiseworthiness.

We aim to cast doubt on this assumption. Here is our reasoning. First, we think that the correct account of praise for human beings is a reasons-responsive account. Second, we will argue that if we take the motivations for a reasons-responsive account of human praiseworthiness seriously, we will end up committed to extending praiseworthiness to non-human animals. By itself, this reasoning might seem to be a *reductio ad absurdum* against taking the motivations for a reasons-responsive account of human praiseworthiness seriously. Rather than extending praiseworthiness to non-human animals (*modus ponens*), as we wish to do, the reader may wish to reject, or restrict, a reasons-responsiveness account of human praiseworthiness (*modus tollens*).

To encourage *modus ponens*, rather than *modus tollens*, we aim to respond to two objections that we anticipate against our proposal to extend praiseworthiness to non-human animals. The first objection is on intuitive or conceptual grounds. This objection insists that the intuition that only human beings can be praiseworthy is too strong to be doubted by philosophical argumentation. We think that the strong reluctance about accepting the praiseworthiness of non-human animals on intuitive or conceptual grounds often confuses factors that determine whether a being is *praiseworthy* with factors that determine whether a being *ought to be praised*. When these considerations are disambiguated, we will argue that attempts at a *reductio ad absurdum* of our position have less force.

The second objection is allegedly on empirical or scientific grounds. This objection says that attributing praiseworthiness to non-human animals, like Scarlett, involves a problematic kind of anthropomorphizing: one is projecting human qualities (e.g., responsiveness to reasons) onto beings that cannot possibly possess them. However, we show that a reasons-responsive approach can be taken to explain any kind of behavior—even instinctive behavior. Hence, our approach is

[2] In this paper, it should always be assumed that the terms 'praise', 'praiseworthy', etc., mean 'moral praise', 'moral praiseworthiness', etc. unless otherwise specified.

[3] One may instead say that it is a conjunction of unquestioned assumptions: only *persons* can be praiseworthy and non-human animals cannot be persons.

[4] One recent exception to this is Korsgaard (2018). She excludes non-human animals as praiseworthy, but this is not an unquestioned assumption. She takes the possibility seriously but her account of responsibility, which retains much of her account in Korsgaard (1996), does exclude non-human animals (as well as many human animals).

anthropomorphic in a modest sense: we prefer the most *charitable* (reasons-responsive) explanation of a behavior that is consistent with the evidence, regardless of whether that behavior is the result of a human or non-human agent. This point is theoretical, not empirical: we claim that reasons-responsive explanations are preferable to non-anthropomorphic (or "anthropectic")[5] explanations when both are consistent with the evidence.[6] Thus, we'll argue that theoretical considerations from the literature on rational analysis and radical interpretation justify our anthropomorphic approach to behavioral explanation.

This worry that we subscribe to a naive kind of anthropomorphism may be deepened by our methodological decision to use anecdotal examples of animal behavior in the introduction and throughout this paper, rather than examples from laboratory or field research (see Footnote 16 for more). But our arguments aren't sensitive to the empirical contents of these examples: we argue that theoretical considerations favor anthropomorphic over anthropectic explanations, when both are consistent with the evidence.[7] This point stands, regardless of the quality of the evidence. As a result, our use of anecdotal examples is meant to *illustrate—not* support—our theoretical reasons for favoring anthropomorphic explanations. Since fields and especially laboratories provide fewer opportunities for clear cases of heroic action by non-human animals, we find that anecdotal reports provide the most convenient illustrations of our anthropomorphic approach.

If our responses to these objections succeed, then we believe that the reasons to accept a reasons-responsive account of praise will override any reasons to doubt the implication that non-human animals can achieve praiseworthiness. We hope that this will refute any *reductio ad absurdum* and convince the reader to take seriously the prospect that both Angel and Scarlett were praiseworthy for rescuing their children.

§1. Reasons-responsiveness for humans

We mentioned in the introduction that philosophers have given many explanations for why only human beings can be praiseworthy. One of the more popular explanations in contemporary, analytic philosophy has to do with autonomy and/or the possession of moral concepts. In this paper, we focus on the Kantian-inspired condition which says that in order for an agent to be morally praiseworthy for an action, the agent must be motivated out of concern for morality. A necessary condition for being 'motivated out of concern for morality' seems to be that the agent must represent their action to themselves as morally right.[8] In other words, the agent must conceive of their morally right action *as* morally right. This condition excludes non-human animals as

[5] Andrews & Huss (2014) introduced the helpful term 'anthropectomy' (derived from the Greek phrase for "cutting out the human") to refer to the methodological preference for full explanations of non-human animal behavior that are sufficiently or maximally different from full explanations of corresponding human behavior.

[6] Though the preference for non-anthropomorphic explanations has been prevalent in the history of animal psychology and cognitive science, this says little in favor of continuing the practice, since the history of animal psychology and cognitive science is rooted in human exceptionalism.

[7] For the most part, we take for granted that there are already a variety of explanations that are empirically adequate, some of which are anthropomorphic and the rest of which are anthropectic. In this way, we've screened off empirical concerns. However, one exception is that we do argue in Section IV that a reasons-responsiveness approach can be taken to explain even so-called instinctual behavior.

[8] Here is how Herman (1993, p. 6) expresses this type of condition: "when we say that an action has moral worth, we mean to indicate (at the very least) that the agent acted dutifully from an interest in the rightness of his action."

candidates for moral praise since, plausibly, in order to represent an action as morally right an agent must possess the concept MORALLY RIGHT. The presumption is that non-human animals cannot possess the concept MORALLY RIGHT, so cannot satisfy this Kantian condition.

We expect many are familiar with the motivating case for this kind of condition. We are asked to imagine a shopkeeper who never tries to cheat his customers. Not cheating your customers is the morally right thing to do. However, this is not sufficient to make the actions of the shopkeeper praiseworthy. In one case, we are told that the shopkeeper is only motivated by profit; he does not try to cheat his customers because, if he were to get caught, then that would be bad for business. In another case, we are told that the shopkeeper is motivated by doing the right thing. He treats his customers fairly because he judges that he is morally obligated to treat his customers fairly. Only under the latter description is the shopkeeper praiseworthy. According to Kant, this is because only under the latter description does he act out of respect for the moral law. How can one act out of respect for the moral law if one does not even have the concept of a moral law? Intuitively, one cannot, which is why the Kantian inspired accounts of responsibility tend to require the possession of moral concepts.[9]

Though there does seem to be room to argue that non-human animals do have moral concepts, or, perhaps, proto-moral concepts, we will grant that only human beings have them for the purposes of this paper.[10] Perhaps the strongest reason for rejecting this Kantian condition is that there are good reasons for thinking that it is false for human actions. Arpaly (2003), for instance, argues that Kantian conceptions of praiseworthiness seem to go badly wrong in cases of inadvertent virtue. Cases where an agent does what they think is the morally wrong action, but, plausibly, they are still praiseworthy for acting the way that they did. The traditional example of inadvertent virtue is the case of Huckleberry Finn. At one point in the story, Huck Finn has a chance to turn Jim in. He thinks he is morally obligated to turn Jim in, but he fails to do so. He considers his failure to turn Jim in as a sign of moral weakness, not moral strength. Huck Finn clearly cannot be motivated by concern for morality in the Kantian sense, because he takes himself to be acting contrary to it. Huck Finn clearly violated the Kantian requirement to represent his action as morally right since he represents his action as morally wrong. In spite of this, it seems clear that Huck Finn not only

[9] Notice that this condition is, presumably, a necessary, but not sufficient, aspect of a full Kantian account, since we expect that the Kantian would say that merely representing your action as morally right is insufficient to count as acting out of respect for the moral law. So, it is not as though by targeting this condition, we are targeting an overly-strong version of a Kantian account.

[10] A number of authors have argued that there are moral psychological capacities that are widespread among social animals (e.g., Bekoff & Pierce 2009; Piece & Bekoff 2012; Rowlands 2012, 2017, 2019; Andrews & Gruen 2014) or, at least, proto-moral psychological capacities (e.g., de Waal 1996, 2006; Flack & de Waal 2000). While these authors fall short of attributing moral (or proto-moral) concept possession, we think that, given some anti-intellectualist assumptions about concept possession in general, it is possible to use the considerations of these authors as a basis for arguing for moral (or proto-moral) concept possession. It is possible to question the assumption that only humans have moral concepts. This questioning, of course, depends on very difficult questions about what exactly it takes to possess moral concepts (or even just the possession of concepts *simpliciter*). Yet, even so, it is difficult to see how these arguments would extend beyond a relatively small number of types of animals outside of human beings — perhaps apes, elephants, and dolphins. Ultimately, our view is not that humans and some human-like animals can be praiseworthy, but rather that *animals* can be praiseworthy. Furthermore, our arguments are compatible with the possibility that only human beings possess moral concepts, so we will not question the assumption regarding moral concept possession in this paper. Though we do emphasize that it is important to distinguish between moral concepts and morally relevant concepts. See our discussion in section II.

does the morally right thing, but that he is morally praiseworthy for doing so. Condemnation or indifference toward Huck Finn's actions seems inappropriate. Of course, Huck Finn is not a perfect moral agent, since he has false beliefs about the demands of morality, but he still has a good moral character that manifests when he acts contrary to these false beliefs.

Huck Finn suffers from a kind of *akrasia*, he acts contrary to his best judgment. But we do not need to appeal to literature to find examples of practical *akrasia*. Perhaps a parent thinks the morally right thing to do is to be cold and distant to their child, but they cannot bring themselves to be, so are very engaged and attentive. They think that they are doing the morally wrong thing due to a weakness in their character. But really, they are sensitive to their child's needs and desires, as any good parent would be. Arpaly gives an example of a student who "waving his copy of *Atlas Shrugged* in one's face, preaches that one should be selfish and then proceeds to lose sleep generously helping his peers" (Arpaly 2003, p. 78). This parent is praiseworthy for being a good parent and the student is praiseworthy for helping his peers, in spite of their incorrect evaluations of their behavior.

The lesson to learn from cases of inadvertent virtue is that we cannot put too much emphasis on conscious representation, autonomy, or deliberation, else we risk saying Huck Finn, and other cases of practical *akrasia,* are not praiseworthy. Reasons-responsive theorists have taken this lesson to heart.

A broad, mostly uncontroversial statement of a reasons-responsive account says that an agent is morally praiseworthy for an action if, and only if, they do the right thing for the right reasons. Uncontroversial, because what the account really says depends on how we are to understand 'for the right reasons'. Nomy Arpaly makes a distinction between two senses of 'acting for the right reasons'. One sense, which she calls *de dicto* representation, is the aforementioned Kantian condition that an agent must represent their action as morally right. The other sense, which she calls *de re* sensitivity, is understood as sensitivity to the right-making reasons of an action.[11] Huck Finn fails the condition of *de dicto* representation, since he does not represent his action as morally right. Huck Finn satisfies the condition of *de re* sensitivity, since he is sensitive to Jim's humanity, his friendship with Jim, etc., when he decides not to turn Jim in.

For Arpaly, praiseworthiness only requires that the reasons that explain their actions are also the reasons that explain why the action is morally right. In other words, an agent's *motivating* reasons are aligned with the *right-making* reasons of the action. The fact that an agent represents an action as morally right *de dicto* may be a motivating reason, but it is never a right-making reason — countless nefarious actors throughout history have done terrible things whilst thinking that they are in the right. So, a *de dicto* representation may be part of the explanation for why an agent *acts*, but it is never part of the explanation for why the action is *right*.

[11] The terms *de dicto* and *de re* are often used in philosophy to distinguish between a word/concept (*de dicto* translates as 'about what is said') and its object/instantiation (*de re* translates to 'about the thing'). The Kantian condition is focused on the concept of right action, while Arpaly's account is focused on the instantiation of right reasons, which is why the *de dicto*/*de re* distinction is invoked here. Admittedly, the use of the distinction in this context is somewhat peculiar, however, it has become the norm in the moral responsibility literature to distinguish between views using this distinction.

One upshot of this account is that sensitivity to the morally relevant reasons can be opaque to the agent. That is, they could think they are doing the wrong thing and/or acting on the wrong reasons, yet still be praiseworthy so long as they responded to the right reasons *de re*—which is precisely Huck Finn's situation. How and whether the agent represents the action *de dicto* is irrelevant to the praiseworthiness of the action because it is irrelevant to the rightness of the action.[12] In Arpaly's (2003) words, "[f]or a right action to have (positive) moral worth, it is neither sufficient nor necessary that it stem from the agent's interest in the rightness of his action" (p. 73).[13,14]

§2. Reasons-responsiveness for non-humans animals

In the previous section, we sketched one of the primary motivations for a reasons-responsive account of praise; namely, the possibility of inadvertent virtue. There are different varieties of reasons-responsive accounts, but we think that they are fundamentally committed to the following.

(RR) An agent S is praiseworthy for action A just in case:
- (i) A is the morally right action.
- (ii) Reasons that explain why the action is morally right are identical to reasons that explain why S did A.

The two most representative accounts of (RR) come from Arpaly (2003) and Markovits (2010). Arpaly says that "for an agent to be morally praiseworthy for doing the right thing is for her to have done the right thing for the relevant moral reasons—that is, the reasons for which she acts are identical to the reasons for which the action is right" (2003, p. 72). Markovits argues in favor of the Coincident Reasons Thesis, which says "my action is morally worthy if and only if my motivating reasons for acting coincide with the reasons morally justifying the action—that is, if and only if I perform the action I morally ought to perform, for the (normative) reasons why it morally ought to be performed" (2010, p. 203).

More must be added to (RR) to give a *complete* account of praiseworthiness. For instance, Arpaly (2003) argues that the degree to which an agent is morally praiseworthy is determined by the agent's depth of moral concern. Markovits' (2010, 2012) account also says that praiseworthiness comes in degrees: an agent is praiseworthy to the degree that the non-instrumental reasons motivating the action coincide with the non-instrumental reasons that morally justify its performance. Arpaly & Schroeder (2014) give an account that appeals to the intrinsic desires of

[12] There may be reasons for thinking that it would be better to have all these things align, however this lack of alignment doesn't diminish the praiseworthiness for a particular action.

[13] As we understand it, *de dicto* representation is a weaker condition than 'interest in the rightness of his action' (or 'caring about morality *de dicto*' as Arpaly puts it in Arpaly 2015) since the latter presumes the former but the former does not necessarily presume the latter.

[14] We take the reasons that count in favor of a reasons-responsive account of praise for humans to be particularly strong. As a result, we are particularly inclined to take seriously its implications, even if they are *prima facie* counterintuitive. However, we recognize that the reader may disagree: they may be unimpressed by Huck Finn cases and sympathetic to the Kantian view that *de dicto* representation is necessary for praiseworthiness (e.g. Sliwa 2016, Johnson-King 2020). We don't have much to say to this reader in this paper: they will take our ensuing argument to be a *reductio ad absurdum* against reasons-responsive accounts of praise. Instead, we mean to be addressing the reader who is impressed by Huck Finn cases yet resists attributing praiseworthiness to non-human animals. We aim to argue that this is an untenable position to hold.

the agent. Fischer and Ravizza (1998) flesh out (ii) in terms of guidance control, which, at the very least, implies a counterfactual sensitivity to the reasons for action.[15] So (RR) is a skeleton of an account of moral praise. There are many ways of fleshing it out, but fleshing out is only possible if there are bones to attach the flesh to. Insofar as reasons-responsive theorists are motivated by cases of inadvertent virtue, we think that they ought to proceed with caution when adding substantive conditions to (i) and (ii) that skew intellectual, else they risk undermining the original motivations for a reasons-responsive account.

What we aim to show in this section is (a) that there are cases of non-human actions that seem to satisfy (RR) as stated, (b) amending (RR) to exclude non-human animals would require adding a strong intellectual condition for praise, but (c) a strong intellectual condition would make (RR) effectively indistinguishable from the Kantian-inspired accounts that the reasons-responsive theorist rejects. Furthermore, the differing ways that reasons-responsive theorists have fleshed out (RR) do not, in principle, have barriers that exclude the praiseworthiness of non-human animals.

Recall the two cases from the introduction. To us, it seems fairly obvious that there are, at the very least, some reasons that both explain why Scarlett acts the way she did and explains why her act was morally right. That her kittens would be harmed or killed is part of the explanation for why she acted and the fact that she prevented her kittens from being harmed or killed is part of the explanation for why the action was morally right.[16] It seems, then, that (RR) implies that this is a case of praiseworthy action.

The case of Scarlett is a dramatic case but not unique. Another cat, Zatara, was recorded protecting a young child from aggressive dogs. A stray dog in Chile waded into a busy road to drag another dog that had been hit by a car to safety.[17] There are less dramatic examples too, like a dog rescuing and caring for a rabbit, a chimpanzee protecting an injured bird, and, in experimental settings, monkeys and rats have been shown to refuse food upon learning that taking the food would cause harm to one of their companions. And there are countless examples of the mundane sort, like a dog helping a ferret who is struggling to jump onto a couch or a buffalo using their horn to flip over a tortoise that was stuck on their back. In all of these cases, it seems like we have actions

[15] Fischer & Ravizza (1998) also make a distinction between *receptivity* to reasons and *responsiveness* to reasons. Being receptive to reasons means being able to recognize factual considerations that count as reasons for action. One might try to exclude non-human animals from this account of responsibility by arguing for a demanding constraint for what it takes to recognize factual considerations as reasons for action, but this will just lead to a similar dilemma one faces when trying to make the constraint of responsiveness to reasons more demanding. Making receptivity more demanding means that fewer human beings are responsible for their actions, fewer human actions are ones that we are responsible for, and the less room there will be for allowing cases of inadvertent virtue.

[16] We assume without argument that preventing Scarlett's kittens from burning to death is morally right. This presumably implies that Scarlett's kittens have moral standing, which we expect most readers to accept. Nevertheless, we think it is important to mention that we reject that the moral agency of non-human animals (or, at least, their capacity for being morally praiseworthy) depends on their moral standing. On the contrary, we think that the moral agency and moral standing of non-human animals are mutually independent. After all, suppose we rejected that non-human animals had moral standing. Even so, our argument would be exactly the same: non-human animals could still behave in ways that count as morally right vis-à-vis humans (who have uncontroversial moral standing), as when a dog rescues their human companion. In these cases, a reasons-responsive account of praise would require us to attribute praiseworthiness to this dog's action.

[17] More examples of purportedly heroic behavior can be found in Gruen (2002), Rowlands (2012), and Andrews (2015).

where it is plausible that the reasons that explain why these actions are morally right are also reasons that explain why the animals did what they did.[18] So (RR) would say that these actions are all praiseworthy to at least *some* degree.

This implication, however, is one that is often not even recognized as a possibility by reasons-responsive theorists and, if it is, the implication is resisted. The most natural move is to concede that non-human animals can act for *reasons* but insist that they cannot act for *moral reasons*. The explanation given is that non-human animals lack the morally relevant concepts that are necessary for acting for moral reasons in addition to acting for mere reasons. Arpaly herself suggests that the relative impoverishment of morally relevant concepts explains why non-human animals cannot act for moral reasons.

> "[T]he dog's mind presumably cannot grasp — nor can it track, the way even unsophisticated people can — such things as increasing utility, respecting persons, or even friendship… Thus, even if this animal can act for reasons, to some extent, it cannot respond to moral reasons, even though it may occasionally come close…to judge a dog vicious for not responding to moral reasons would be similar to judging a dog a philistine for not being able to appreciate Mahler."[19]

This strikes us as an unsatisfying response for a number of reasons. First, it seems to stem from an insistence on intellectualizing the demands for moral praise that Arpaly is at pains to argue against.

Second, her examples of what a dog cannot track are not persuasive. Not only is it plausible to think that a dog can track *increasing utility*, but it is also plausible that dogs can track *maximization of expected utility*. For example, our stray dog in Chile could have improved the injured dog's situation in any number of suboptimal ways: they could have dragged the injured dog to a less busy part of the road, they could have brought them food while leaving them in the road, etc. Instead, they competently singled out the action that maximizes the expected utility of the injured dog: they dragged them completely off the road. It seems plausible to think that insofar as we can attribute maximization of expected utility reasoning to human beings in general, then we can attribute maximization of expected utility reasoning to non-human animals as well. A dog cannot recite the Von Neumann–Morgenstern axioms, of course, but neither can most humans.

While it is true that there is a sense in which a dog cannot conceptualize *what it means* to respect persons, at least if we mean persons in the philosophically substantive sense, this seems to be irrelevant to the question of whether they can respect persons. Most humans do not conceptualize persons in a philosophically substantive sense, but we think that most humans can respect and disrespect persons. And even if we say that dogs cannot respect *persons*, they certainly can respect conscious individuals. For example, domestic dogs, and canids in general, reliably play bow in order to initiate play. This gesture conveys their intentions to play and invites another dog to play

[18] One might be suspicious of the anecdotal nature of these examples. There were no first-hand witnesses to Scarlett's heroism, and many of the other examples are recordings of isolated animal actions. One might complain that it is illegitimate to conclude anything substantive from these examples. We expect that this suspicion is really just a version of the objection from anthropomorphism that we address in section IV, since we doubt that this suspicion would be assuaged by pointing to seemingly morally valenced actions observed in natural settings (see examples in Bekoff 2004; Bekoff & Pierce 2009; Pierce & Bekoff 2012; Rowlands 2012) or in controlled laboratory settings (e.g. Church 1959; Rice 1964; Ben-Ami Bartal, *et al*. 2011; Sato, *et al* 2015).

[19] Arpaly (2003, p. 146).

with them. Throughout play, there are a variety of techniques — play markers — that are used to ensure that the playful mood is maintained. For infant canids, transgressions in play agreements are tolerated, however, violating the playful mood beyond infancy results in punishment. If individuals are perceived as cheaters or play unfairly, then they are less likely to find play partners.[20] It is difficult to make sense of these delicate and complex social behaviors if we are forbidden from invoking any kind of respect.

Finally, it is strange to say that a dog cannot track friendship. Otherwise, it would be difficult to explain why dogs are so much more protective, caring, and even sometimes jealous of "their humans" and even of "their pets" than of human and non-human strangers. Of course, the way that they track friendship may not be identical to the way that we do, but again, it is not enough to say that a dog does not conceptualize a morally relevant concept *in exactly the same way we do,* but, rather, one must say that it is impossible for a dog to track *anything* morally relevant. When it comes to friendship, that strikes us as doubtful. Many people would report that they share a deep bond with their dog, and it is hard to imagine that the friendship is entirely one-way.[21]

So even if a dog cannot track concepts like INCREASING UTILITY, RESPECTING PERSONS, or FRIENDSHIP in *precisely the same way* we do, it still makes sense to invoke these concepts when explaining their behavior or, at the very least, it makes sense to invoke morally-relevant concepts in the neighborhood to explain their behavior. Perhaps a dog cannot respond to *all* of the properties of a person, but they can respond to *some* of the properties (e.g., individuality, sentience, preferences). So even if it is not the exact same concepts, so long as their behavior is still explained with morally relevant concepts, then it is possible for them to be praiseworthy or blameworthy on a reasons-responsive account. It would seem that the only way to block this would be to have a higher cognitive demand for what it takes to have a reason to be part of the explanation for one's behavior. But notice the higher the demand, the rarer cases of inadvertent virtue become. If we say that one has to think about friendship with a particular concept of FRIENDSHIP, then a person who does not have that particular concept, or who represents their action in a different way, could not be considered to be acting out of a concern for friendship. Yet a constraint of this form resembles the Kantian condition Arpaly has aspired to refute.

But even if Arpaly is right that dogs neither grasp nor track complex concepts like INCREASING UTILITY, RESPECTING PERSONS, or FRIENDSHIP, nor any concepts in the neighborhood that are also morally relevant, this would not imply that non-human animals cannot be praiseworthy, since they still track (and possibly grasp) basic concepts that are morally relevant. For instance, it would be difficult to deny that Scarlett tracks the concept PAIN, and, furthermore, that tracking the concept PAIN can feature in the explanation of her actions. And not just in a trivial way, but in a way that satisfies the requirement of *de re* sensitivity.[22] In general, it is implausible to claim that animals cannot track pain. And it would be similarly implausible to suggest that, even though they track

[20] For more on the social play of canids and other non-human animals, see Bekoff (1975, 1977, 1995, 2001), Bekoff & Allen (1992), and Horowitz (2009).

[21] For an interesting discussion of the nature of friendship between humans and non-human animals, see Fröding & Peterson (2011a, 2011b), and Rowlands (2011a, 2011b).

[22] One could claim that in this specific case Scarlett intended to save her kitten's lives, not protect them from pain. We would think that both are part of the explanation, however, we'll note that we do not need such a dramatic case to motivate this point, since cats, and non-human animals in general, are often very responsive to the pain, or potential pain, of their children.

these concepts, these concepts are never employed in explaining what they do. And, finally, it is implausible to suggest that these concepts are not part of the explanation for why saving kittens from a fire is a morally right action. Ultimately, then, there is still an identity between at least some of the reasons that make Scarlett's action morally right and the reasons that explain why she did what she did. The fact that her action prevented her kittens from suffering pain explains why it is morally right and explains why she did what she did. PAIN is an important moral concept, and it is one that very many non-human animals can track (and possibly possess).[23]

Note that, if we are correct about Scarlett, this does not mean that moral praise extends from humans to very human-like creatures (e.g. chimpanzees and gorillas) or from humans to non-human animals with remarkably complex cognitive structures (e.g. dolphins and elephants), but rather it extends to any non-human animal that can track the pain of other beings in a way that makes it so the pain of other beings is an explanation for why they act. Even if Arpaly is right that only human beings track or possess a rich set of morally relevant concepts, this, we think, only implies that we can be praiseworthy for more actions and that we can act for more reasons when compared to non-human animals, not that we alone can be praiseworthy.[24]

Here is another way of putting the point. Accounts of praise will typically construe the possession of or sensitivity to morally relevant concepts as a threshold that only humans can meet. The above criticism of Arpaly amounts to saying that sketching such a threshold such that *only* humans can meet it will invariably result in a highly intellectualized account of praise. Instead of imagining a threshold that only humans can meet, we should imagine that humans possess and track a particularly rich set of morally relevant concepts, whereas non-human animals track or even possess a relatively impoverished set of morally relevant concepts. Perhaps one thinks that only concepts like PAIN, PLEASURE, and DESIRES can be used in explaining the actions of dogs. Or perhaps one thinks that it should also include concepts like INCREASING UTILITY, RESPECTING

[23] The reader might prefer an even simpler explanation: Scarlett was merely compelled by instinct to save her kittens. We have no sympathy for this response. For one, it is an empty explanation: it denies that Scarlett's behavior was reasons-responsive while saying nothing positive about what caused Scarlett's behavior. After all, the causal powers of instincts are left completely unspecified. For another, the notion of instinct here is at odds with the notion of instinct in psychology, which might define instinctive behavior as, e.g., including "highly stereotyped, coordinated movements, the neuromotor apparatus of which belongs, in its complete form, to the hereditary constitution of the animal" (Tinbergen, 1942; cf., Lorenz, 1939). In such an explanation, the instinct is supposed to be the heritable neuromotor apparatus itself. Clearly, Scarlett's behavior cannot be instinctive on this definition: it is a complex response to a unique and hence, unprecedented situation. This is why we have no sympathy for the instinct explanation of Scarlett's behavior: it is a pseudo-scientific explanation that amounts to nothing more than a refusal to take the causal structure of non-human animal behavior seriously. Moreover, note that even the notion of instinct advocated by Tinbergen has fallen out of favor among psychologists for being insufficiently explanatory (e.g., Lehrman, 1953). More on this in section IV.

[24] Our primary claim in this section is that non-human animals can satisfy (RR) since they can be said to act for morally relevant reasons. The defense of this claim is not unique to us. For instance, Rowlands (2012) extensively argues for the claim that non-human animals can be motivated to act by moral reasons. This, claims Rowlands, makes them moral subjects, which is to be distinguished from moral patients, beings who are legitimate objections of moral concern, and moral agents, beings that are responsible for, and can be morally evaluated based on, their actions. In Rowlands' terminology, we are arguing that many non-human animals are moral patients, moral subjects, and moral agents, whereas Rowlands only argues that many non-human animals are moral patients and moral subjects. However, the reasons that Rowlands rejects the possibility of non-human animals being moral agents is a combination of Kantian assumptions and a specious *reductio ad absurdum.* We gave some reasons for rejecting a Kantian account in section I and we criticize Rowlands' *reductio* in section III.

PERSONS, or FRIENDSHIP. Or perhaps one thinks that not those concepts, but rather concepts in the neighborhood — INCREASING UTILITY*, RESPECTING PERSONS*, FRIENDSHIP* — can be used in explaining dog behavior. Regardless, what our arguments imply is that while the degree and sophistication of moral responsibility may scale with the amount and complexity of moral concepts tracked or possessed. The amount and sophistication of concepts one attributes to non-human animals like dogs, elephants, and gorillas will not determine whether or not they meet some threshold, but, rather, will determine *what* they can be responsible for and the *degree* to which they are responsible.

So far, we have mostly been discussing Arpaly's account of moral responsibility. However, we do not think that our arguments are restricted to Arpaly's account. Indeed, our above considerations seem to fit well with Markovits' account. Given that non-human animals are less cognitively sophisticated in certain respects may imply that the degree to which their motivating reasons for an action cannot coincide with *all* of the justifying reasons, however, it does not imply that there is no coincidence.

Consider the influential account from Fischer and Ravizza (1998). According to Fischer and Ravizza, what is required for moral responsibility is (a) that the agent possesses a psychological mechanism M that is sufficiently reasons-responsive and (b) M is "owned" by the agent. If one were to endorse this kind of account and wished to exclude the possibility of non-human animals being praiseworthy, one or both of (a) and (b) would have to posit a significantly high cognitive requirement. In light of our arguments in this section, it seems the only way to construe (a) such that it rejects the praiseworthiness of non-human animals is to construe reasons responsiveness as responsiveness to moral reasons *de dicto.*

Regarding (b), it may be tempting to posit a high demand for when a psychological mechanism is "owned" by an agent. Perhaps they would have to conclude from rational deliberation that they approve of this psychological mechanism. There are many problems with this. First, this seems to presuppose an implausible amount of transparency of psychological mechanisms. Even when given time for rational reflection, introspection is quite limited in understanding what psychological mechanisms are actually implemented in reasoning and action. Second, it is strikingly odd. Why should a psychological mechanism only be considered mine when I engage in complex reflection?

Consider an analogy with epistemology. We all have psychological mechanisms that implement deductive, inductive, and abductive reasoning. These mechanisms are present in infants. It would be absurd to suggest that these reasoning mechanisms are not owned by a person until they engage in a certain kind of rational reflection where they recognize that they use induction and reflectively endorse their use of induction. Third, it would imply that few human beings are ever responsible for what they do, since few human beings have engaged in such reasoning. Note that this is not merely a requirement that a person engage in reflective moral reasoning, but, rather, that a person engage in reflective reasoning *about the psychological mechanism responsible for their ethical actions* and, after coming to know the nature of this mechanism, they reflectively endorse this mechanism. Fourth, it would seem to undermine cases of inadvertent virtue. Insofar as Huck Finn has a psychological mechanism that is responsive to the right moral reasons *de re*, he does not

approve of this mechanism. So, such an account would say that Huck Finn is not responsible for his refusal to turn Jim in.

Instead, it is far more plausible to give a non-deliberative, anti-intellectual account of "ownership" of a psychological mechanism. Arpaly & Schroeder (2014), for instance, argue for an account of "ownership" of this kind. Instead of psychological mechanisms, their target is desires. They wish to distinguish desires "owned" by the agent, *intrinsic desires*, from desires not owned by the agent. On their account, what makes an intrinsic desire an *intrinsic* desire is that it is a state of an unconscious learning system, one that plays certain causal roles in acting, feeling, and cognizing. An account of ownership such as this one would avoid all four of the problems sketched above, but it would also allow for the possibility of the praiseworthiness of non-human animals. It is not as though these unconscious learning systems are unique to humans. Indeed, part of their argument relies on dispelling the idea that reward-and-punishment-based learning only arises in rats navigating mazes.[25]

It is, of course, possible for one to try to discriminate between learning systems such that it only counts for humans, so to speak. But a pattern should now be clear. We would again see a tension between the motivations for rejecting Kantian constraints on moral responsibility and constraints that would exclude the possibility of the praiseworthiness of non-human animals. Reasons-responsive theorists, we think, would do well to give up on trying to thread this needle and instead fully embrace the consequences of rejecting Kantianism, one of which is to accept our fellow animals as moral agents.

§3. Response to the objection from intuition

Recall from the introduction that we expect that some readers will reject the direction of our argument. According to this response, we should not be arguing, via *modus ponens*, to the conclusion that non-human animals can be praiseworthy, but, rather, we should be arguing, via *modus tollens*, that the aforementioned reasons-responsive accounts of praise are false. In other words, our argument could be taken as a *reductio ad absurdum* for reasons-responsive accounts of praise. We will consider and reject two motivations for this response. In this section, we'll consider and reject the strong intuitive judgment that non-human animals simply cannot be praiseworthy, so any philosophical theory that says they can should be rejected.

It is difficult to argue against an intuitive judgment. We cannot attack the premises that support it because an intuition is, in general, not something (explicitly) supported by premises, but rather something used as a premise. We do not mean to suggest that philosophical intuitions are misguided or are things that we should always treat with suspicion, but rather we mean to point out that in order to respond to an intuitive judgment, the best we can do is try to offer some reasons for reconsidering the weight given to such a judgment.

Our attempt at softening this intuitive judgment starts with making a distinction between *praiseworthiness* and *praise*. When an agent is praiseworthy, we take it to mean that a certain kind of attitude toward that agent is fitting. It requires no extrinsic relation between the thing having

[25] Arpaly & Schroeder (2014, p. 128-9).

the attitude and the target of the attitude. For instance, if a person were to read about morally praiseworthy actions in a history book, it would be fitting for them to have a certain kind of attitude toward the person being described in the book — namely, the attitude of attributing to the person the property of being praiseworthy. This is true even if the person has been dead for thousands of years. Compare this attitude with the attitude of admiration. Two people can debate whether it is fitting to admire Napoleon Bonaparte, even though neither person bears any significant relationship to Napoleon.

To *praise* is to *act.* Acts are justified or unjustified. Perhaps an agent's being praiseworthy is necessary to justify praise, but it is not sufficient. Our *practices* of praise and blame are deeply social, so relations to others play a distinctive role for praise and blame that is absent for praiseworthiness and blameworthiness. For example, some people think that blame (and presumably also praise) requires a certain form of *standing*.[26] However, it is implausible to think that this kind of standing is required for one to merely *judge* that someone is blameworthy.

In terms of Watson's (1996) distinction between responsibility as *attributability* and responsibility as *accountability*, we are arguing, at minimum, that non-human animals are candidates for responsibility as attributability. This has certain implications for the possibility of non-human animals as candidates for accountability, however, there is clearly a gap to be filled between these two and there are many different ways of filling this gap. We think that what people report as an aversion to the claim that animals can be praiseworthy (responsibility as attributability) might really be an aversion to the claim that non-human animals ought to be praised (responsibility as accountability). If this is right, then this may take away some of the force of the intuitive objection since non-human animals are often not embedded in many of our social practices. It is not as though the actions of non-human animals are not worthy of being praiseworthy or blameworthy, but rather the acts of praising and blaming often serve social functions that often makes it unnecessary or unjustified to praise or blame non-human animals. For instance, Rowlands' aversion to treating non-human animals as moral agents seems to rest on this conflation between attributability and accountability.

> The claim that animals can be moral agents is, I shall argue, deeply problematic…the concept of agency is inseparable from that of responsibility, and hence from the concepts of praise and blame. If animals are moral agents, it follows they must be responsible for what they do. But if they are responsible for what they do, then, it seems, they can be held accountable for what they do. At one time, courts of law—both nonsecular and secular—set up to try (and subsequently execute) animals for perceived crimes were not uncommon. I assume few would wish to recommend a return to this practice. At the core of this unwillingness is the thought that animals are not responsible, and so cannot be held culpable, for what they do.[27]

It should be clear that the inference from *being responsible* to *being held accountable in a court of law* is a *non-sequitur*, but so is the inference from *being responsible* to *being held accountable.* It ignores the plethora of social and practical considerations that feature in determining whether actions are justified. For example, you may (rightly) judge that your friend's partner is treating

[26] See, e.g., Wertheimer (1998, p. 499), Cohen (2006, p. 118).
[27] Rowlands (2012, p. 83-4).

them unfairly, but this does not imply that you are thereby justified in intervening in their relationship. You may (rightly) judge that a certain social practice from a community to which you don't belong is archaic, but, so long as it is not seriously morally wrong, you would not be justified in intervening in that community and sabotaging the social practice. On the contrary, you would be met with the rightful charge of imperialism.

We need not take a stand on what *exactly* connects correct judgments of attributability and correct actions of accountability. Our point is simply that there is a substantive connection here, not a trivial or straightforward one, yet this seems to be forgotten when criticizing the possibility of non-human animals being responsible for some of their behavior. Attempts at a *reductio* of our view seem to simply ignore the complex and difficult pathway from *correctly judging* that X is responsible for Y and *taking retributive action* against X on the basis that X is responsible for Y. Once this difficult pathway is reflected upon, we think that the intuitive objection to our account loses much of its force.

Our thesis, that non-human animals can be praiseworthy, is certainly a controversial philosophical position. Often controversial philosophical positions, if true, call for extreme changes in how we think or act if we wish to have our thoughts and actions align with the truth. For example, if the free will skeptics are correct, then this calls for extreme changes in how we think about and how we act toward those people who perform actions that are morally wrong. However, we wish to emphasize that it is not so clear that our thesis would fall into this category. Ultimately, what we have argued is that a certain attitude is fitting to have toward many non-human animals (of course, this affects how we ought to act as well, but as we emphasized earlier there are many considerations that determine how one should act, not just the initial reactive attitude). What is interesting is that, for many people, the kinds of attitudes that our theory deems fitting are often the kinds of attitudes that people already form toward non-human animals. And these attitudes are formed automatically. We take it that when many people hear about Scarlett the cat or watch a video of a dog protecting a child, they automatically form a positive appraisal attitude toward these non-human animals. It is the orthodox philosophical position in moral responsibility that deems this attitude infelicitous. According to orthodoxy, these attitudes are tantamount to being angry at your Roomba for getting stuck under your sofa or feeling sorry for Boston Dynamics robots when they are kicked and pushed. 'Maybe one cannot help oneself in forming these attitudes,' orthodoxy says, 'but, at the very least, one ought not reflectively endorse them. One ought to at least recognize that these automatic reactions we have toward the acts of non-human animals are not appropriate.'

While orthodoxy is orthodoxy, there is a significant sense in which *it* is the revisionary thesis, not ours, since it calls for drastic changes in our thoughts towards non-human animals, if we wish those thoughts to be fitting. Of course, our view is revisionary in the sense that many people would be inclined to 'take back' their reactive attitudes toward the actions of non-human animals, but note that changing reflective endorsements of unreflective attitudes to align with those attitudes is far easier than changing our unreflective attitudes to align with reflective endorsements of those attitudes. It seems that it is our view, not orthodoxy, that vindicates the natural attitudes that many people have toward the acts of non-human animals.

Our thesis in this paper is that the actions of non-human animals can be *praiseworthy*, yet one might object that this thesis implies a less palatable claim: that the actions of non-human animals

can also be *blameworthy*. It would be quite counterintuitive to attribute blameworthiness to the lion who kills his rival's cubs after claiming a pride even though it is much more intuitive to attribute praiseworthiness to the lionesses who assiduously hide their cubs from the new lion (to varying success). This asymmetry might seem like a problem for our account: we have drummed up intuitive support for our thesis by foregrounding its intuitive implications vis-à-vis praiseworthiness while backgrounding its counterintuitive implications vis-à-vis blameworthiness.

We agree that there is an important asymmetry between the praiseworthiness and blameworthiness of non-human animal action, but we reject that this is a problem for our account. Before we spell out this asymmetry, though, we'd like to caution against conflating blameworthiness with blame—just like we've cautioned against conflating praiseworthiness with praise. After all, this might lead us to exaggerate the problem. On the one hand, it would be counterintuitive to *blame* the lion for killing his rival's cubs—at least partly because we lack the appropriate standing vis-à-vis the lion. On the other hand, we think that it is more intuitive (or less counterintuitive, at least) to say that a negative attitude towards the lion's killing his rival's cubs is fitting (that his killing is *blameworthy*). If we consider merely holding attitudes toward non-human animals, such attitudes of negative appraisal seem to be quite common.[28]

Even so, we agree that it's less intuitive to attribute blameworthiness than praiseworthiness to the actions of non-human animals. However, we reject that this is a problem for our account because there are deep asymmetries between praiseworthiness and blameworthiness. Some of these asymmetries are very general. For instance, while the structure of (RR) as an account of praiseworthiness is intuitively plausible, an analogous account of blameworthiness is not: an agent can be blameworthy even if the reasons that explain why they performed their action are completely different from the reasons that explain why their action is wrong. For example, a pickpocket might steal an item just because it thrills them, and they are blameworthy for doing so even though the thrill of stealing isn't what explains why their stealing is wrong. After all, blameworthiness involves some kind of failure in responsiveness to moral reasons, not a responsiveness to immoral reasons. This is an important asymmetry, which is one of the reasons for why we set aside the issue of blameworthiness in favor of the issue of praiseworthiness.

Other asymmetries are specific to the actions of non-human animals. Compared to human actions, moral evaluations of the actions of non-human animals, especially those in the wild, are less clear when we are talking about morally bad outcomes when compared to morally good outcomes. For instance, a small benefit conferred from one animal to another — a tortoise flipping upright a fellow tortoise — is easy to evaluate as morally (and all things considered) right. However, even something as appalling as a wolf killing a baby deer is not so easy to evaluate as morally (or all things considered) wrong. After all, if wolves were to stop hunting, or were to be more discriminate toward potential prey, they risk starvation and death. In general, wild animals rarely face situations

[28] For instance, one popular subreddit on the website Reddit.com is /r/animalsbeingjerks and is described as "A place for sharing videos, gifs, and images of animals being jerks." Thinking someone is a jerk is one way of having an attitude of negative appraisal. So, again, it would seem that our view vindicates the natural attitude people have towards animals behaving like jerks, which is that they *are* being jerks. Again, similarly to the attitudes of positive appraisal, many people might 'take back' this attitude if the correctness of this attitude were challenged. We think that the justification for 'taking back' the attitude is likely to be a worry about problematically anthropomorphizing the actions of non-human animals. This worry is the one we dispel in section IV.

where their actions can be straightforwardly evaluated as wrong: they usually face moral dilemmas that make the moral status of their actions either mysterious or indeterminate.[29] When it's counterintuitive to evaluate their actions as blameworthy, we're tracking a deeper problem that has nothing to do with blameworthiness *per se*: that there's no clear way to evaluate their actions as wrong in the first place.

Wrapping up, what we have tried to show in this section is that the intuitive objection is mostly predicated on intuitions about what would be appropriate *behavior* whereas our thesis is about appropriate *attitudes*. Our thesis does not necessarily undermine these intuitions, since there are many considerations that determine the appropriate behavior toward an agent in addition to a judgment of praiseworthiness (or blameworthiness). Furthermore, when we focus on the reactive attitudes we naturally have toward the actions of non-human animals, it is not so clear that our thesis is radically revisionary. At the very least, we hope that the considerations of this section will give pause to those who would initially dismiss our thesis on intuitive grounds.

§4. Response to the objection from anthropomorphism

Now we will finally address the elephant in the room. The second motivation for treating our argument as a *reductio* of reasons-responsive accounts of praise is the vague but deep worry that our various examples of non-human animal actions attribute too much agency to non-human animals. A salient way to formulate this objection is to claim that we are guilty of a problematic kind of *anthropomorphism*: our view unjustifiably projects distinctively human traits onto non-human animals and so, purports to license the unjustifiable projection of a distinctively human achievement (i.e., praiseworthiness) onto non-human animals. This worry is vague because it doesn't draw a sharp line that demarcates how much agency counts as "too much" for non-human animals. However, in a similar manner to the intuitive objection, it insists that that sharp line must be drawn in such a way that it excludes non-human animals from moral praise. We aim to defuse this objection. First, we'll argue that the anthropomorphizing objection begs the question: it is an *a priori* objection on intuitive grounds disguised as an *a posteriori* objection on scientific grounds. Second, we'll integrate this concept of reasons-responsiveness with the methods of rational analysis and radical interpretation to develop a novel response to the anthropomorphism objection.

§4.1. Anthropomorphizing cognitive processes

Those who worry about anthropomorphism often note that animal behavior can be fully explained without appealing to any of the cognitive resources that would be required for them to be morally praiseworthy (even on a reasons-responsive account). Often, the objection goes, we can fully

[29] One of the privileges of human civilization is that it has progressively removed us from moral dilemmas like this. For example, humans in most parts of the world during the 18th century faced deep moral dilemmas between respecting the moral status of non-human animals and acquiring sufficient protein to survive and flourish. It can be argued that the moral status of their killing animals for meat is either mysterious or indeterminate. By comparison, agricultural technology and infrastructure has made it possible for humans in many parts of the world during the 21st century to acquire sufficient protein from plant sources. It can be argued that the moral status of their killing animals is no longer mysterious or indeterminate: it's straightforwardly impermissible. Obviously, though, non-human animals have no such access to these privileges of civilization, so they continue to face moral dilemmas as the default, rather than the exception.

explain animal behavior just by appealing to a complex of instincts, conditioning, and features of the environment—none of which are cognitive or agential resources. But this isn't a remarkable observation: human behavior can be fully explained without appealing to cognitive resources too. For example, any human behavior can be fully explained just by identifying the cascade of action potentials in the neurons of the human's brain that responds to a set of sensory inputs and produces a set of behavioral outputs. This explanation makes only true claims, yet it doesn't appeal to cognitive resources at all: it identifies only activities that happen below the level of cognition.

Clearly, though, we'd reject an explanation of human behavior that merely identified the cascade of action potentials that caused the behavior in response to sensation. The reason is simple: it lacks generality. It's too sensitive to the particularities of a single (albeit large) conjunctive neural event: slight differences in the sensory stimuli may cause significant differences in the exact cascades of action potentials, even though they resolve into slight differences in the behavioral outputs (if any). To achieve generality in behavioral explanation, we need to attribute cognitive states and processes to humans, not neurobiological ones. The same reasoning extends to animal behavior: sure, it might be true that we can fully explain any particular token of animal behavior without attributing cognitive states and processes to them, but the explanation probably won't achieve the requisite generality. The lesson here is simple: the mere fact that behavior *can* be fully explained without appealing to a certain set of cognitive resources doesn't mean that it *should* be explained in this way.[30]

As it turns out, full explanations of behavior are very cheap. In fact, there are infinitely many ways that a given set of sensory inputs could be transformed into a given set of behavioral outputs, so there are infinitely many ways to fully explain any behavior (Anderson 1990). Moreover, the actual cognitive states and processes that mediate between sensory inputs and behavioral outputs are unobservable, so observation is consistent with infinitely many explanations. This creates the notorious *black-box problem* in the philosophy of cognitive science: cognitive explanation is infinitely underdetermined by observation (Sober 1998).[31] Therefore, the black-box problem must be solved *a priori*: we must consult philosophical considerations to decide how human and non-human animal behavior *ought* to be explained.

Outside of animal psychology, most cognitive scientists today accept some version of Anderson's (1990) answer: we should select the explanation of behavior that rationalizes the behavior under

[30] In fact, this isn't a surprising lesson. After all, we often encounter specious yet full explanations of human behavior. For example, consider the psychological egoist who explains all human behavior in terms of self-interest or the Freudian psychoanalyst who explains all human behavior in terms of conflict between the id, ego, and super-ego. These explanations can be modified *ad hoc* to fully explain any human behavior. If we challenge the psychological egoist to explain various great and small achievements of altruism, for example, they will respond that it only *seems* like the person is acting altruistically, but *really* they are acting in self-interest. We reject explanations like these because we recognize that these modifications are *ad hoc*. Therefore, we already intuitively accept that fully explaining behavior is insufficient for correctly explaining it.

[31] The black-box problem is closely related to but different from various problems of radical interpretation in the philosophy of mind and language. Davidson (1973) raises the problem with interpreting speech behavior in particular, vs. behavior in general. Lewis (1974) raises the problem of interpreting behavior by attributing propositional attitudes to agents, vs. cognitive states. Williams (2020) raises the problem of interpreting behavior by interpreting pre-individuated symbols in the language of thought, whereas the black-box problem includes the problem of individuating cognitive entities (such as symbols in the language of thought) in the first place.

cognitive constraints.[32] This methodology is known as *rational analysis*. For example, suppose that we see Angel or Scarlett walking into a house on fire. There are any number of ways that we could fully explain both behaviors: we could appeal to neural cascades, self-interests, instincts, conflicts between the id, ego, and super-ego, etc. But rational analysis requires us to explain their behaviors by rationalizing them: we know that Angel and Scarlett both have reasons to save their children that override their reasons to avoid the threats to their own safety, so rational analysis requires us to infer that Angel and Scarlett must be both registering their reasons and rationally responding to them.[33]

But what exactly are the cognitive processes in Angel and Scarlett that realize their responsiveness to their reasons? Are they equivalent between Angel and Scarlett? The answers to these questions depend on the cognitive constraints that we attribute to Angel and Scarlett. We can infer these from rationally analyzing their behaviors more broadly.[34] [35]A relevant difference between Angel and Scarlett is that Angel is responsive to more kinds of reasons than Scarlett is. For example, there is a sense in which all mothers have reasons to seek out medical help when their children are sick but only Angel will be responsive to these reasons. Out of charity to Scarlett, we rationalize her behavior by attributing more cognitive constraints to her: she isn't responsive to these reasons only because she *cannot* be, given her cognitive constraints. That is, we infer that Scarlett has a cognitive constraint that prevents her from grasping the concept MEDICAL HELP (or, plausibly, any other concept in its neighborhood), such that she can't be responsive to reasons concerning it.

In the most extreme cases, rational analysis might require us to attribute some notion of instincts to non-human animals. For example, many male animals will attempt to mate with any object that bears even a superficial resemblance to a female conspecific. These highly stereotyped mating behaviors aren't responsive to reasons: that the object is not a female conspecific, that it's a female conspecific who has died, that it's a lure designed to resemble a female conspecific, that the object

[32] Rational analysis is often compared to Davidson's (1984) principle of charity, which roughly claims that we should assign beliefs and meaning to a speaker in a way that maximizes the number of true beliefs and true assertions that the speaker is prepared to assert. By comparison, Anderson's rational analysis is more general, and it emphasizes rationality over (or, in addition to) truth: it claims that we should assign anything to an animal (not just beliefs and meanings) that maximizes the rationality of the animal (not just the truth of their beliefs and assertions). This emphasis on rationality over (or, in addition to) truth can also be seen in solutions to related problems of radical interpretation (see Lewis 1974 and Williams 2020).

[33] The particular conception of rationality used in rational analysis is rarely (if ever) explicated within the psychological literature. Instead, it's used quite flexibly to single out any way of responding to any given situation that seems uniquely optimal. We do the same here: we use a reasons-responsiveness conception of rationality here to single out a way of responding to a situation for Angel and Scarlett that seems uniquely optimal. In general, though, we maintain neutrality on the relationship between various conceptions of rationality and rational analysis. See Williams (2020) for a related discussion on the relationship between substantive conceptions of rationality and radical interpretation.

[34] A gap in Anderson's (1990) description of rational analysis is that it's somewhat unclear about how to assign constraints to cognitive agents, which leaves open the possibility that we could be uncharitable to non-human animals by assigning too many constraints to them. So, we emphasize here that constraints should be assigned to rationalize the total set of behaviors of the agent. This emphasis on rationalizing the total set of behaviors (rather than some subset of them) can be found in the literature on radical interpretation (Davidson 1973, Lewis 1974, Williams 2020).

[35] Our proposal that the best explanations of animal behavior are those that best rationalize it (with a reasons-responsiveness conception of rationality) goes beyond other anthropomorphic proposals that the best explanations of animal behavior are those that best account for empirical evidence (e.g., Fitzpatrick, 2008, 2018).

is a female conspecific who is prepared to consume the male, etc. In such cases, rational analysis requires us to attribute some notion of instinct to the male animals out of charity: to rationalize their nearly complete unresponsiveness to reasons. However, such cases are relatively rare. Even lions, for example, won't kill their rival's cubs when they are uncertain about their paternity—a fact that female Asiatic lions have been found to exploit (Chakrabarti & Jhala, 2019).

Thus, there is a certain virtuous sense in which rational analysis is anthropomorphic: it treats humans and non-human animals with the same charity when we explain human and non-human animal behavior. [36] If it recommends different explanations for humans vs. non-human animals, it does so only because it cannot rationalize the unresponsiveness of non-human animals to certain kinds of reasons and so must attribute further cognitive constraints to them. As a result, rational analysis minimizes the differences between human and non-human animal cognition: it attributes the least cognitive differences to humans and non-human animals that are necessary to interpret their behavioral differences (i.e., their differences in rational, or appropriate, responsiveness to reasons).

By comparison, many animal psychologists continue attributing cognitive constraints to non-human animals even when doing so is uncharitable—i.e., when they aren't necessary to rationalize animal behavior more broadly. They tend to explain animal behavior by attributing cognitive states that are lower on a "cognitive hierarchy" and to explain human behavior by attributing cognitive states that are higher on a "cognitive hierarchy". We could call this methodology *behaviorism for animals, representationalism for people*.[37] This "cognitive hierarchy" is rarely spelled out, but it typically seems to be underwritten by an implicit conception of rationality.[38] Andrews & Huss (2014) introduced the helpful term 'anthropectomy' (derived from the Greek phrase for "cutting out the human") to individuate this bias, which we construe as being uncharitable to non-human animals in order to vindicate human exceptionalism.

Finally, we propose that our view is anthropomorphic in the same way that rational analysis is. When an animal — human or non-human — is equally responsive to a particular kind of reason, we treat their behavior with equal charity: we rationalize it in the same way, and we attribute the same cognitive activity to the animal. It is rational for both Angel and Scarlett to *immediately* recognize and respond to their overriding reasons to rescue their children. And it would be irrational for either of them to, e.g., distance themselves from their inclinations, reflect on what

[36] Different versions of Anderson's (1990) rational analysis are extremely influential in the literature on cognitive modeling, where specific algorithms are needed to generate formal models of cognition (for an influential example, see Oaksford & Chater, 2007). Unfortunately, a lot of animal psychology isn't informed by the literature on rational analysis. As a result, many animal psychologists use implicit solutions to the black-box problem that are undermotivated (compared to rational analysis) and typically, uncharitable to non-human animals.

[37] This is a reference to Robert Nozick (1974), who called the attitude that treated the moral status of animals and non-human animals radically different *utilitarianism for animals, Kantianism for people.* The approach to non-human animals is behaviorist in the loose sense that it fully explains behavior by attributing non-rational states (e.g., instincts and reflexes) and rational states lowest on the cognitive hierarchy. And the approach to humans is representationalist in the loose sense that it fully explains behavior by attributing rational states highest in the cognitive hierarchy (e.g., representations).

[38] For example, "Thorndike uses 'lower' to refer to those animals whose behavior can be accounted for in terms of 'a bundle of original and acquired connections between situation and response' whereas human behavior is more appropriately described in terms of consciousness and insight (Thorndike 1911, 4)," quoted from Andrews & Huss (2014, p. 715).

duty requires of them, and respond to their reasons only once they've confirmed that this is what duty requires of them. Therefore, we should treat Angel and Scarlett with equal charity: both immediately do what is rationally required of them.

§4.2. Anthropomorphizing cognitive capacities

Still, we may be tempted to say that Angel's response is different from Scarlett's insofar as Angel has the capacity to respond to the situation using her concepts DUTY and PARENTHOOD, whereas Scarlett lacks this capacity. Of course, it is possible that Angel does have this capacity and could exercise it: she does grasp DUTY and PARENTHOOD and reasoning with those concepts would probably lead Angel to take the course of action that she actually took. However, we think that there is an important sense in which it would be unreasonable for Angel to respond to her situation by deploying DUTY and PARENTHOOD. In the famous words of Bernard Williams (1981), it would be "one thought too many" for her to think that risking the flames would save her children and that saving her children is a duty for her qua parent. Instead, we think that the most reasonable response is for Angel to respond as Scarlett does: to *immediately* respond to the plight of her children by risking the flames, without a second thought.

Ironically, then, it would be uncharitable to Angel if we inferred that she uses DUTY and PARENTHOOD to rescue her children and thereby fetishizes morality in a distinctively human (and perhaps Kantian) way. In this case, some would prefer to be uncharitable to humans in order to maintain that human responses are "higher" in the cognitive hierarchy than non-human animal responses. This reveals the absurdity of anthropectomy and human exceptionalism: it leads us to be uncharitable not only to non-human animals but also to humans.

To be clear, though, there are possible situations where Angel could deserve this unsavory interpretation: if Angel was extremely principled about her decisions and never demonstrated inadvertent virtue, we would have to conclude that Angel has become unresponsive to reasons before she has conceptualized them. And that might lead us to conclude that Angel will pause to conceptualize her duties and her parenthood before she risks the fire to save her children. Again, we think this is irrational, but critically, this unsavory interpretation wouldn't be uncharitable: it would require us to assign cognitive constraints to Angel in order to rationalize her unresponsiveness to reasons that she hasn't conceptualized yet. The important point is that most humans (and all non-human animals) lack these cognitive constraints and hence, rationally respond in an immediate way to urgent, overriding reasons.

There seems to be only one way to attribute praiseworthiness to human behavior like Angel's yet withhold attributing praiseworthiness to non-human behavior like Scarlett's without being uncharitable to human or non-human animals. And that would be to endorse a counterfactual account, which claims that behavior by an agent is apt to be praiseworthy only if that behavior either was—or could have been—performed by the agent in response to *de dicto* moral reasons. After all, humans are always able to behave for *de dicto* moral reasons, even if they don't actually do so, whereas non-human animals are never able to do so (*ex hypothesi*). But we think that this counterfactual addendum is *ad hoc*: when we're evaluating a behavior as praiseworthy, we're not

evaluating how it could have been—we're evaluating how it actually was.[39] What makes the Huck Finn case interesting (and hence, what motivates the reasons-responsive account of praise) is that he refused to turn Jim in despite his actual *de dicto* representation that this refusal is morally wrong, not that he could have represented this refusal as morally right.

If we refuse to attribute "higher" cognitive states to humans when it would be uncharitable to do so, then there will be many cases (such as Angel and Scarlett's) where we must attribute the same (or, at least, very similar) cognitive states and processes to human and non-human animals. In these cases, a reasons-responsive account of praiseworthiness will produce the same verdict if we evaluate the actions as they actually were caused (not as they could have been caused): Angel and Scarlett will both come out as praiseworthy (even perhaps, *equally* praiseworthy). Still, there will be many cases where humans will be responsive to more reasons than non-human animals. Then it will be charitable for us to attribute more sophisticated cognitive states and processes to humans than to non-human animals. In these cases, our account insists that non-human animal action will still achieve a degree of praiseworthiness, even though human action is able to achieve greater degrees of praiseworthiness.

§5. Conclusion

Let's revisit Angel and Scarlett. Recall that we began this paper by recounting their behaviors as stories of heroic action and invited the reader to evaluate them as paradigmatic cases of praiseworthiness. But we didn't tell the whole truth: we waited till the end of both stories before we revealed that Angel was a human and Scarlett was a cat. We expect that an interesting change would have happened to the skeptical reader. They would continue to say that Angel acted from love for her children, but they would suddenly start to demur for Scarlett. They would say that she "acted from love for her children" in some sense, but she didn't *really* act from love for her children. Likewise, they would say that Scarlett acted for "reasons" in some sense, but she didn't *really* act for reasons. And so on.

By leaving out the information that Scarlett was feline, did we deceive our skeptical reader? Our skeptical reader may think so, but we maintain that we did not. When we revealed that Scarlett was a feline, we haven't revealed any information relevant to the actual way that she responded to her reasons. After all, Angel and Scarlett's situation gave them reasons that they were equally capable of responding to. So, rational analysis requires us to be equally charitable to both Angel and Scarlett: we are committed to explaining their behaviors in the same way. That is, we are committed to explaining that Angel and Scarlett both recognized that their children were in danger, recognized that rescuing them would mean considerable risk and harm to themselves, and yet responded to the powerful overriding reasons to rescue their children.[40]

---

[39] This point is emphasized in Markovitz (2012).

[40] The skeptical reader might wonder: what if we had revealed that Scarlett was a robot designed for rescue? Wouldn't that be deceptive? Yet how would that be any different? Our response to this line of questioning is to point to the fact that a robot designed for rescue isn't internally responsive to any reasons whatsoever: such a robot won't do anything, much less respond to reasons for rescuing children, unless it is being steered by a human controller (or programmed by a human programmer). If we had failed to reveal from the beginning that Scarlett was a robot, we would have deceived the reader by misattributing the agency of the behavior to the robot, rather than the remote controller. This would have misconstrued the relevant reasons in the situation: e.g., Scarlett's aversion to the fire

The change in the skeptical reader, we propose, is that they withdrew their charity to Scarlett. This is evident from the way that they refuse to seriously engage with the way that she actually acted. They might make dismissive appeals to instinct. Or they might focus on the way that Scarlett couldn't have acted, the kinds of reasons that she couldn't have responded to, or the kinds of situations that would have presented reasons that she couldn't have responded to. We think that this refusal to seriously and charitably engage with Scarlett's agency is wrong (epistemically and morally). Once we recognize this, we've argued, we are committed to recognizing Scarlett's actions as praiseworthy if, and only if, we recognize Angel's actions as praiseworthy. We think that a reasons-responsive account is right to recognize both as praiseworthy, but we recognize that a Kantian-inspired account might only recognize Scarlett's actions as praiseworthy.

At this point, the reader may worry that human agency is being demoted in our argument. We have two things to say about this worry. First, we maintain that human agency does represent a unique achievement, albeit in a weaker sense than we might have thought: most (but certainly not all) humans have the capacity to respond to many more kinds of reasons than non-human animals do, so most (but again, not all) humans have a far greater capacity to achieve praiseworthiness than non-human animals do. Second, we reject the framing: recognizing that non-human animal behavior can be praiseworthy is a *promotion* for non-human animals, not a *demotion* for humans. We aren't rejecting the idea that humans are special—we are insisting that non-human animals are special too.

If the skeptical reader reacts to our stories of Angel and Scarlett by making pseudo-scientific appeals to instinct and dismissing Scarlett's actual achievements to talk about the limitations of her capacity, we think the antidote for the hesitant but open-minded reader is to consciously take Scarlett's agency very seriously. Once we do, we face the black-box problem, and we must turn to principles of charity in order to solve that problem and single out the best explanation of her behavior. And the rest comes along for the ride.

Speaking for ourselves, we have come to experience a radical shift in our philosophical intuitions. It strikes us as intuitive that animals could be praiseworthy because we think humans can be praiseworthy and humans *are* animals. A rat solving a maze is not that different from a human solving a sudoku puzzle. An elephant raising a calf is not that different from a human raising a child. A dog comforting their owner is not that different from a child comforting their parent. When we think of Angel and Scarlett, we just think about *mothers* saving their *children*. Both are heroic. Both displayed courage. Both are praiseworthy.

would have been explained by the reasons for the controller to mitigate the costs of fire damage, not by Scarlett's own reasons to preserve her health and safety and to avoid pain. This clarifies that no such deception was used in our retelling of Scarlett's actual story.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Lawrence Erlbaum Associates, Inc.

Andrews, K. (2015). *The animal mind: An introduction to the philosophy of animal cognition*. London: Routledge.

Andrews, K. & Huss, B. (2014). Anthropomorphism, anthropectomy, and the null hypothesis, *Biology & Philosophy*, 29, 771–729.

Andrews, K., and Gruen, L. (2014). Empathy in other apes. In H. Maibom (ed.), *Empathy and Morality*. New York: Oxford University Press.

Arpaly, N. (2003). *Unprincipled virtue*: *An inquiry into moral agency*. New York: Oxford University Press.

Arpaly, N. (2015). Huckleberry Finn revisited: Inverse akrasia and moral ignorance, in *The Nature of Moral Responsibility: New Essays* (R. Clarke, M. McKenna & A. Smith, eds.). New York: Oxford University Press

Arpaly, N. & Schroeder, T. (2014). *In Praise of Desire*. New York: Oxford University Press.

Bekoff, M. (1975), 'The communication of play intention: Are play signals functional?' *Semiotica*, 15, pp. 231–9.

Bekoff, M. (1977), 'Social communication in canids: Evidence for the evolution of a stereotyped mammalian display', *Science*, 197, pp. 1097–9.

Bekoff, M. (1995), 'Play signals as punctuation: The structure of social play in canids', *Behaviour*, 132, pp. 419–29.

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

Bekoff, M. and Allen, C. (1992), 'Intentional icons: towards an evolutionary cognitive ethology', *Ethology*, 91, pp. 1–16.

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

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

Chakrabarti, S. & Jhala, Y. V. (2019). Battle of the sexes: A multi-male mating strategy helps lionesses win the gender war of fitness. *Behavioral Ecology*, 30(4), 1050–1061. https://doi.org/10.1093/beheco/arz048

Church, R. M. (1959). Emotional reactions of rats to the pain of others, *Journal of Comparative and Physiological Psychology*, 52(2), 132–134.

Cohen, G. A. (2006). Casting the first stone: Who can, and who can't, condemn the terrorists? *Royal Institute of Philosophy Supplement*, 58, 113–136.

Davidson, D. (1973). Radical interpretation, *Dialectica*, 27(3/4), 313–328.

de Waal, F. (1999). Anthropomorphism and anthropodenial: consistency in our thinking about humans and other animals, *Philosophical Topics*, 27, 255-280.

de Waal, F. (2016). *Are we smart enough to know how smart Animals are?* New York: W. W. Norton & Company.

Fitzpatrick, S. (2008). Doing away with Morgan's Canon. *Mind & Language*, 23, 224–246.

Fitzpatrick, S. (2018). Against Morgan's Canon, in *The Routledge Handbook of Philosophy of Animal Minds*, K. Andrews and J. Beck (eds.). New York: Routledge.

Fischer, J. M. & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge University Press.

Fröding, B. & Peterson, M. (2011a). Animal ethics based on friendship. *Journal of Animal Ethics*, 1(1), 58-69.

Fröding, B. & Peterson, M. (2011a). Animals and friendship: A reply to Rowlands. *Journal of Animal Ethics*, 1(2), 87-189.

Gruen, L. (2002). The morals of animal minds. In *The Cognitive Animal: Empirical and Theoretical Perspectives on Animal Cognition*, M. Bekoff, C. Allen, & G. M. Burghardt (eds.). Cambridge, MA: MIT Press.

Herman, B. (1993). *The practice of moral judgment*. Cambridge, MA: Harvard University Press.

Horowitz, A. (2009). Attention to attention in domestic dog (*Canis familiaris*) dyadic play. *Animal Cognition*, 12, 107-118.

Johnson-King, Z. (2020). Accidentally doing the right thing. *Philosophy and Phenomenological Research*, *100*(1), 186–206.

Kant, I. (1994). *Ethical philosophy.* 2nd ed. (J. Ellington, trans.). Indianapolis: Hacking.

Korsgaard, C. (1996). *The sources of normativity*. Cambridge: Cambridge University Press.

Korsgaard, C. (2018). *Fellow creatures: Our obligations to the other animals*. Oxford University Press.

Lehrman, D. S. (1953). A critique of Konrad Lorenz's theory of instinctive behavior. *The Quarterly Review of Biology*, *28*(4), 337–363.

Lewis, D. (1974). Radical interpretation, *Synthese*, 27(3/4), 331–344.

Lorenz, K. (1939). Vergleichende verhaltensforschung. *Zoologischer Anzeiger*, *12*(Suppl. band): 69–102.

Markovits, J. (2010). Acting for the right reasons. *Philosophical Review*, 119(2), 201–242.

Markovits, J. (2012). Saints, heroes, sages, and villains. *Philosophical Studies*, 158, 289-311.

Nozick, R. (1974). *Anarchy, state, and utopia*. Basic Books.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford: Oxford University Press.

Pierce, J., & Bekoff, M. (2012). Wild justice redux: What we know about social justice in animals and why it matters. *Social Justice Research*, 25, 122-139.

Rice, G. E. J. (1964). Aiding behavior vs. fear in the albino rat. *The Psychological Record*, 14(2), 165–170.

Rowlands, M. (2011a). Friendship and animals: A reply to Fröding and Peterson, *Journal of Animal Ethics*, 1(1), 70-79.

Rowlands, M. (2011b). Friendship and animals, again: A response to Fröding and Peterson, *Journal of Animal Ethics*, 1(2), 90-194.

Rowlands, M. (2012). *Can animals be moral?* New York: Oxford University Press.

Rowlands, M. (2017). Moral subjects. In K. Andrews and J. Beck (eds.), *Routledge Handbook of the Philosophy of Animal Minds.* New York: Routledge

Rowlands, M. (2019). *Can animals be persons?* New York: Oxford University Press.

Sato, N., Tan, L., Tate, K. & Okada, M. (2015). Rats demonstrate helping behavior toward a soaked conspecific, *Animal Cognition*, 18(5), 1039–1047.

Sliwa, P. (2016). Moral worth and moral knowledge. *Philosophy and Phenomenological Research,* 93(2), 393–418.

Sober, E. (1998). Black box inference: When should intervening variables be postulated? *The British Journal for the Philosophy of Science*, 49, 469–498.

Thorndike, E. L. (1911). *Animal intelligence*. Macmillan, New York.

Tinbergen, N. (1942). An objectivistic study of the innate behavior of animals. *Bibliotheca biotheoretica, Leiden*, D, 1: 39–98.

Van Bourg, J., Patterson, J. E., & Wynne, C. D. (2020). Pet dogs (*Canis lupus familiaris*) release their trapped and distressed owners: Individual variation and evidence of emotional contagion. *PLoS One*, 15(4), e0231742.

Watson, G. (1996). Two faces of responsibility. *Philosophical Topics*, 24(2), 227–248.

Williams, B. (1981). *Moral luck*. Cambridge: Cambridge University Press.

Williams, B. (1985). *Ethics and the limits of philosophy*. London: Fontana.

Williams, J. R. G. (2020). *The metaphysics of representation*. Oxford University Press.