The Philosophical Quarterly Vol. 76, No. 2 *2026*
ISSN 0031-8094 https://doi.org/10.1093/pq/pqag008
Advance Access Publication 14th February 2026

Animal rationality through higher-order reasons

MARTIN NIEDERL [1] AND ZSÓFIA VIRÁNYI[2,3]

[1]*University of Chicago, USA,* [2]*Cognition and Applied Ethology, Messerli Research Institute, University of Veterinary Medicine, Vienna, and* [3]*Medical University of Vienna, University of Vienna, Austria*

Melis and Monsó have recently argued that non-human animals are capable of reflectively responding to normative reasons. They argue for this via animals' capacity to respond to undermining epistemic defeaters. Although we are ultimately sympathetic to this idea, the present article develops a principled methodological challenge to the empirical evidence they cite to corroborate their claims. What we call the Acquisition Problem shows that the studies that Melis and Monsó cite cannot provide differential support for the hypothesis that animals reflectively respond to reasons. We close by suggesting two possible ways forward that evade our challenge, while remaining true to the spirit of Melis and Monsó's proposal. First, future studies could aim to model testimony of undermining defeaters in social learning. Second, and even more promisingly, we should focus on the broader phenomenon that Melis and Monsó's proposal points towards, namely, animals' responsiveness to higher-order reasons.

Keywords: rationality; non-human animals; animal agency; higher-order reasons; animal reason; comparative psychology.

I. Introduction

Human beings are still regularly singled out among the larger animal kingdom as *the* rational animal.[1] This is typically understood in terms of humans'

Correspondence to: Martin Niederl, mwniederl@uchicago.edu

[1]We are incredibly grateful to in-depth comments by Niels de Haan, Mikayla Kelley, and especially Giacomo Melis. Special thanks also to Will Hornett, the entire team of Giacomo Melis' ARED project, to two anonymous reviewers for their insightful comments, and to audiences at the LSE *Workshop on Animal Minds*, the FSU *Graduate Conference on Free Will, Moral Responsibility, and Agency*, the University of Laval *Everything Agency Conference*, and the University of Stirling *Rational Animals Conference*. This work was supported by a UK Research and Innovation Future Leaders Fellowship (grant # MR/T042249/1) awarded to Giacomo Melis.

allegedly unique capacity to reflectively respond to normative reasons.[2] As opposed to our fellow creatures, some argue, we can not only act and believe in accordance with the reasons that speak in favour of our doing so, but as a reflective response to these reasons. Against this philosophical orthodoxy, an increasing number of philosophers have recently started to argue that certain animals can in fact respond to reasons in the same (or at least a very similar) sense in which we do.[3] Out of those, Melis & Monsó (2024) argue for the most robust conclusion in the literature to date: that certain non-human animals can reflectively respond to reasons *as reasons*, and that current empirical evidence supports this.

Our aim in this article is 2-fold. First, we develop a principled methodological challenge for the empirical evidence that Melis & Monsó (2024: Section 5) use to support their arguments: the *Acquisition Problem*. It shows that the studies that Melis and Monsó cite cannot provide differential support for the hypothesis that animals reflectively respond to reasons. Crucially, this is so partly in virtue of the very construct these studies supposedly track: undermining epistemic defeaters. Second, we point towards possibilities for future research on animals' capacity for reasons-responsiveness by focusing on the broader phenomenon of animals' sensitivity to higher-order reasons. Our suggestion thus remains true to the spirit (if not the letter) of Melis and Monsó's novel and important contribution.

II. Responding to reasons

To understand Melis and Monsó's proposal, let's first turn to normative reasons and what it means to reflectively respond to them. Normative reasons are typically conceived of as facts that favour responses.[4] The fact that this building is on fire, for instance, is a reason for me to leave it; and the fact that it is raining is a reason for me to believe that this is so. In both cases, a given fact favours some response: an action, or the formation of a belief. For our purposes, we can thus assume that there is no deep ontological difference between practical and epistemic reasons (see Scanlon 1998). Evidence for p, for instance, is simply an epistemic normative reason for believing that p.

If normative reasons are facts that favour responses, then subjects can be reflectively responsive to reasons only if they can be 'credited with the capacity for some sort of normative or axiological thinking' (Alvarez 2022: 197). The thought is that there is an important difference between merely acting or believing in accordance with the relevant reasons, and responding to them. In particular, in reflectively responding to a reason, one must somehow cognise

[2] See e.g. Alvarez (2010), Boyle (2016), Korsgaard (2018), and Alvarez (2022).
[3] Among others, Hurley (2003) and Glock (2019).
[4] See Scanlon (1998), cf. Wedgewood (2017).

the relevant fact *as a reason* (Boghossian 2014; Korsgaard 2018). Given that reasons are facts that favour responses, reflectively responding to a reason *as a reason* just means to respond to a fact *as favouring* the response in question:

[*Responsiveness*$_{R}$] *S* is reflectively responsive to a reason *r* if and only if *S* cognises *r* as favouring φing and (attempts to) φ because of this.[5]

First of all, notice that this is a definition partly in terms of the explanation of a subject's response. It is not only that *S* cognises some fact as favouring the response, but also that this explains *S*'s (attempt at) so responding. Second, for our purposes, we can think of cognizing a reason *as favouring* φing as being *competently attracted*, given the relevant facts, to act in the ways favoured by that reason (Sylvan 2015).[6] In responding to reasons *as reasons*, that is, *S*'s response is explained by *S*'s thought competently tracking the responses favoured by *p*.

Sceptics like Alvarez (2022) thus lay a 2-fold burden of proof on proponents of animal rationality:

[*Principled Question*] Can animals cognise favouring relations?
[*Methodological Question*] When are we licensed to explain animals' behaviour in terms of responses to favouring relations?

Although they do not differentiate between these two questions, Melis & Monsó (2024) are arguably the first to effectively provide answers to both. They argue for animals' capacity to respond to normative reasons *via* their capacity for responses to undermining epistemic defeaters, thereby sketching the beginnings of an empirical paradigm to test for this capacity.

III. Animals as responding to reasons

In a novel treatment of this problem, Melis and Monsó use the notion of *epistemic defeat* to argue that animals can cognise favouring relations.[7] They thus propose the following idea for answering the *Principled Question*:

[*Responsiveness*$_{UD}$] If *S* has the capacity to respond to undermining defeaters, then *S* has the capacity to respond to reasons *as reasons*.

This thesis is about the sufficiency of one capacity for another. To understand it, let's turn to epistemic defeaters. There are at least two kinds of epistemic

[5] See also Dancy (2000), Alvarez (2010), and Korsgaard (2018). We talk about *attempts* because reasons-responsiveness plausibly does not require success in acting. Thanks to Mikayla Kelley for raising this point.

[6] See also Lord (2018). As Boghossian (2014) himself admits, hardly any adult ever explicitly applies the concept 'favouring' when responding to reasons. Thanks to an anonymous reviewer for pressing us on this point.

[7] See Pollock (1974) for the *locus classicus* on epistemic defeaters.

defeaters. *Overriding defeaters* suggest 'that one should replace one's belief in P with a belief in not-P' (Melis and Monsó 2024: 852). If Amy has told me that it is raining, and I see that it is not raining, then the latter fact completely overrides Amy's testimony—in effect, I now have reason to believe that it is *not* raining. Overriding defeaters thus constitute first-order evidence for or against some belief.

Undermining defeaters, on the other hand, are facts that 'provide the subject with a reason to give up one's belief in P without thereby providing a reason to believe not-P' (Melis and Monsó 2024: 852). Typically, that is, undermining defeaters work by 'suggesting that a specific source of evidence is unreliable or that a specific piece of evidence may be misleading' (Melis and Monsó 2024: 852). If Amy is unreliable, their unreliability speaks against *my treating their testimony as a reason* to believe that *p*. As opposed to overriding defeaters, undermining defeaters thus constitute higher-order evidence. Undermining defeaters do not constitute evidence for or against the truth of *p*—rather, they alter the evidential (i.e. normative) status of the available evidence. While that fact that Amy is unreliable itself neither favours believing that it is going to rain, nor believing that it is not going to rain, it alters the normative status of my available evidence (i.e. *that Amy says it is going to rain*). In order to keep track, let's fix the difference between overriding and undermining defeaters thus:

[*OD*] Overriding defeaters constitute evidence for or against some proposition.
[*UD*] Undermining defeaters alter the normative status of evidence.

As Melis and Monsó argue, responding to UDs thus implies the capacity for normative thought. In particular, they hold that thoughts like

> <Source X is unreliable> and similar underminers are instances of so-called 'higher-order' evidence. Because of it, responding to them demands explicit epistemic evaluations: the formulation of *thoughts about evidence* (Melis and Monsó 2024: 853; emphasis ours).

The strategy is quite nifty. Second-order reasons generally evaluate (i.e. favour or disfavour) first-order reasons (see Raz 1975). So, just like first-order reasons influence the normative status of responses, second-order reasons influence the normative status of first-order reasons, that is, facts favouring responses. Amy's unreliability, for instance, diminishes the evidential status of their weather-predictions. So, if I had no way of assessing evidential status, I would be unable to appropriately respond to the fact that Amy is unreliable. In this way, a capacity to respond to higher-order reasons *implies* a sensitivity to the extent to which facts favour certain responses. My withholding judgment after learning that Amy is unreliable thus cannot be explained except by way

of my capacity to track the altered extent to which their testimony favours my believing that it is raining.[8]

This leads straight into Melis and Monsó's proposed strategy for the *Methodological Question*. Sceptics like Alvarez (2022) emphasize that there seems to be no way we can ascertain whether animals respond merely to facts, or to those facts *as favouring* some response. However, if we can show that animals respond to UDs, then this straightforwardly *implies* their capacity to cognise favouring relations. Why? Because it is a conceptual truth that responses to higher-order reasons are responses specifically to altered favouring relations. One thus cannot explain responses to UDs without implying the subject's competency to track the evidential status of the relevant facts. Hence, if we can ascertain whether animals have the capacity to respond to UDs, we *thereby* ascertain whether they have the capacity to respond to favouring relations.

IV. The acquisition problem

While we applaud Melis and Monsó's general approach to this issue, we see a principled methodological challenge for the empirical evidence they cite. To prove their point, Melis and Monsó need it to be the case that animals respond specifically to UDs. This is because only UDs are genuinely second-order reasons and can thus provide us with evidence for animals' capacity for normative thought. Moreover, in order to justify that we are testing for real *transfer* of concept or ability, one would need to show that subjects refuse 'to follow indications coming from the unreliable source *the very first time* they reencounter it in a *new context*' (Melis and Monsó 2024: 853; emphases ours).[9]

As we argue now, however, there is a principled reason why even the most promising studies that Melis & Monsó (2024: Section 5) cite cannot corroborate these claims.[10] To illustrate our point, consider first the studies from the *Social Informant* paradigm that Melis & Monsó (2024: 854) cite. In a first phase, Takaoka et al. (2015) presented dogs with an informant who reliably pointed to the one of the two containers that contained food. In a second phase, this informant sometimes pointed to the empty container—they became unreliable. Finally, in a third phase, the same informant reliably pointed to the rewarded container again. Both Takaoka et al. (2015) and Melis and Monsó suggest that 'dogs tended to stop following the leads in the third phase because they had been exposed to the unreliability of the informant in the second phase' (2024: 855). In a control condition, a new informant was introduced for the third

[8] Thanks to Nathan Howard for pressing us on this point.

[9] See also Singley & Anderson (1989).

[10] Note that, while Melis and Monsó cite studies that do not follow this strict pattern, they discount their results as strictly insufficient (2024: 854). We thus do not engage with these studies here. Thanks to an anonymous reviewer for pressing us on this.

phase and dogs tended to follow their leads again. This, in turn, 'is compatible with the suggestion that dogs were guided by thoughts like <that particular informant is unreliable, but not this other one>' (Melis and Monsó 2024: 855). According to Melis and Monsó, these studies thus suggest dogs' responsiveness to second-order evidence regarding the evidential status of one particular informant's pointing gestures.

However, notice that there is a salient alternative explanation precisely because of the ODs the dogs experienced in the second phase. Pointing is the evidence for the dog to believe that food is in a specific container. In phase two, this evidence is repeatedly *overridden* by the fact *that no food is there*. The association between *this* particular informant's pointing and food could thus be extinguished.[11] Their hesitation in phase three is thus easily explained by their experience with many ODs with regard to that specific informant. Importantly, this is precisely because of the way in which the animals were supposed to learn about the experimenter's unreliability in the first place: by way of their pointing being repeatedly *overridden* by the emptiness of the container.

What is even more important than the specific alternative explanations, however, is what grounds their availability: the fact that this experiment does not create sufficiently *novel* circumstances in the crucial third phase. The features that together constitute the ODs in the training phases all re-occur in the testing phase: the respective experimenter, their pointing, and the reward. The dogs' frustrations and other negative attitudes associated with that experimenter could thus be carried over to the testing phase. The control condition is similarly easily explained by a lack of negative associations towards the new experimenter and dogs' general tendency to follow pointing gestures (Dwyer and Cole 2018). This makes it impossible to tell apart whether subjects' responses are based on reflective engagement with the evidence, or simple positive/negative/neutral associations with the experimenter.

As another example, turn to the *Appearance-Reality* paradigm that Melis and Monsó (2024: 855) consult. In a study by Krachun et al. (2016), chimpanzees could initially see in which of two boxes a smaller grape (x) and a bigger grape (X) were hidden. However, one box had a magnifying and the other minimizing lens in front of them. So, once placed in the boxes, the grapes seemed to have reversed their sizes. After choosing grape x a few times, the chimpanzees started choosing the x-looking grape to get the actually bigger grape X. Melis and Monsó (2024: 855–6) take the chimpanzees' behaviour to be 'best explained as being guided by a thought like <the grape *seems* big/small>', holding that 'to represent a state of affairs as merely seeming true is to make a personal-level epistemic evaluation.'

[11] See Dunsmoor et al. (2015) on extinction.

However, an alternative explanation is once again quite salient. Chimpanzees do not avoid grape x (i.e. apply reverse-contingency rules), nor do they seek grape X (which is the researchers' explanation for the control condition). Rather, they learn to *avoid grape X* when it is in the box because avoiding grape X always gets you the bigger grape. After choosing the bigger-seeming grape a few times and ending up with grape x, this is a simple rule to learn. Again, the acquisition of this rule is explained by their experience with ODs, namely, finding *that grape X is not in the box*. This explanation thus accounts for both the test and the control condition while positing nothing but a simple associative rule the animals acquired due to their exposure with multiple ODs.[12]

Our point is that there is a pattern at work here, and it has to do with how UDs are (putatively) acquired in these studies. What ordinarily grounds one's acquisition of a UD are many first-hand experiences of *structurally related* ODs. Tentatively, consider ODs and UDs structurally related when they concern the same general relations between features: same informants, type of evidence, goals, etc. The reason why I have the UD *that Amy's weather predictions are unreliable* at my disposal is because I have experienced multiple structurally related ODs in the past (e.g. *that it is not actually raining when Amy said it would*). Similarly, the reason why we can explain the animals' behaviour in terms of simpler responses is precisely because of their first-hand experience with multiple structurally related ODs during the tests. Hence, the acquisition of UDs *via* structurally related ODs is sufficient to provide us with an alternative explanation of animals' behaviour. Since this is how all the most promising studies that Melis and Monsó cite operate, these studies cannot corroborate their claims regarding animals' reasons-responsiveness. This gives rise to a principled methodological challenge:

[*Acquisition Problem*] If *S*'s (putative) acquisition of a UD is grounded in first-hand experience with *structurally related* ODs, then there is no *differential support* for the hypothesis that animals respond to UDs.

The notion of *differential support* is borrowed from Fitzpatrick's (2008) formulation of *Evidentialism*. In combination with Evidentialism, an independently plausible and widely accepted principle, the *Acquisition Problem* thus poses a principled methodological challenge for extant studies:

[*Evidentialism*] in no case should we endorse an explanation of animal behaviour in terms of cognitive process X on the basis of the available evidence if that evidence gives us no reason to prefer it to an alternative

[12] Essentially the same complaint arguably extends to a new study by Schleihauf et al. (2025). In brief, they provide chimpanzees with evidence that *overrides* (rather than undermines) their previous evidence, thus not only casting previous information into doubt but effectively making it irrelevant (see our Veit & Niederl et al. ms.)

> explanation in terms of a different cognitive process Y—whether this be lower or higher on the 'psychical scale' (Fitzpatrick 2008: 242).[13]

The problem, in summary, is this. Extant studies (putatively) equip animals with UDs by training them with first-hand experience with many ODs. By design, these UDs and ODs must be structurally related. To learn about the unreliability of *S*'s pointing, for instance, dogs are trained on precisely this: *S*'s pointing in relation to food. This precludes the testing phase from creating sufficiently *novel* circumstances in order for the need for genuine transfer of concepts or abilities to occur.[14] As such, these studies cannot provide differential support for the hypothesis that those animals are responsive to reasons as reasons.

V. Two possible ways forward

In closing, we want to propose two ways of circumventing the problems we pose while staying true to the spirit (even if not necessarily the letter) of Melis and Monsó's account.

First, one could try to model the acquisition of UDs without prior first-hand experience with related ODs. In animals like us, there is a clear case of this: *testimony*. If you *tell me* that Amy is unreliable, I do not need to fall prey to their false weather predictions in order to acquire the UD *that Amy is unreliable*. One piece of information contains the relevant upshot of your experience with multiple ODs. Now, instead of learning about Amy's unreliability through verbal communication, I could also come to know about it by observing how *you* respond to Amy's weather predictions. I would thus acquire a UD by learning from your behaviour. Since I have had no first-hand experience with Amy's predictions, there is no simple rule that could tie my hesitating response to their testimony. Hence, we could avoid the *Acquisition Problem* by creating social learning conditions for animals that model testimony without language.[15]

Moving away from UDs altogether, we have a second, possibly even more promising proposal. As we have argued, responding to UDs is just a particular form of responding to higher-order reasons. Responses to higher-order reasons, in turn, are just responses informed by an altered favouring relations between facts and responses. Since it is this wider phenomenon that does the theoretical work in Melis and Monsó's proposal, we suggest broadening the scope of future experiments beyond the bounds of epistemic defeaters.

[13] This is a more refined version of what *Morgan's Canon* is typically taken to express. Thanks to an anonymous reviewer for drawing our attention to this.

[14] Again, see Melis & Monsó (2024: 853) and Singley & Anderson (1989).

[15] See, for instance, Abdai & Miklósi (2016) on social eavesdropping.

Here is a brief sketch of a more concrete suggestion.[16] Suppose an animal has to choose between retrieving food from tunnel A or B. The only difference between them is that tunnel B is *opaque*. At first, the fact that tunnel B is opaque is not a reason for or against taking it. So, the animal should be indifferent which tunnel to choose to get to the food. That tunnel B is opaque only *becomes* a reason for taking it once another fact is in place: *that a competitor is watching*. This last part is crucial. The fact that a competitor is watching alters the normative status of the fact *that tunnel B is opaque*. This is precisely how higher-order reasons work: they modify the favouring relations between facts and responses.

If the animals respond to this without prior exposure to the interaction between the tunnel's opacity, the competitor, and the reward, then we have sufficiently different conditions for genuine transfer to occur. Crucially, the animals must thus learn about the three relevant properties (reward at the end of the tunnel, opacity/transparency of the tunnel, and competition with the experimenter) in three separate learning phases. This prevents learned associations between those properties to carry over into the testing phase. This design thus also allows us to observe how animals respond *the very first time* they have to make this choice, thereby more effectively '[constraining] the size of the hypothesis space for the possible cognitive mechanisms an agent might be using' (Taylor et al. 2022: 741) than extant studies can.

Melis & Monsó's (2024) proposal is no doubt ground-breaking for proponents of animal agency and rationality. We have argued that the studies they use to support their point suffer from a thoroughgoing methodological impediment. Moreover, this problem is to some extent inherent in the phenomenon these studies aim to test: responses to UDs. However, this should not lead us to overlook or even abandon the deeper insight in Melis and Monsó's importantly novel contribution: that if animals are capable of responding to higher-order reasons, we should start seriously reconsidering our status as *the* rational animal.

References

Abdai, J. and Miklósi, Á. (2016) 'The Origin and Evolution of Social Evaluation, Social Eavesdropping, Reputation Formation, Image Scoring and What You Will', *Frontiers in Psychology*, 7: 1772.

Alvarez, M. (2010) *Kinds of Reasons: an Essay in the Philosophy of Action*. Oxford: OUP.

[16] See Veit & Niederl et al. (ms.) for a fully developed implementation of this design. Note that this study works with practical first-order reasons, rather than epistemic ones. However, note also that this differentiation does not track on the second-order. Second-order reasons are defined by their altering the normative force of first-order reasons, thus being moot on the precise nature of the normativity involved. As such, our proposal still aligns well with the spirit of Melis and Monsó's paper. Thanks to an anonymous reviewer for pressing us on this.

Alvarez, M. (2022) 'Intelligence and Reasons in animals'. In: C. Pfisterer, N. Rathgeb and E. Schmidt (eds.) *Wittgenstein and Beyond: Essays in Honour of Hans-Johann Glock*, pp. 185–203. New York, NY: Routledge.
Boghossian, P. (2014) 'What Is Inference?', *Philosophical Studies*, 169: 1–18.
Boyle, M. (2016) 'Additive Theories of Rationality: a Critique', *European Journal of Philosophy*, 24: 527–55.
Dancy, J. (2000) *Practical Reality*. Oxford: OUP.
Dunsmoor, J. et al. (2015) 'Rethinking Extinction', *Neuron*, 88: 47–63.
Dwyer, C. and Cole, M. (2018) 'Domesticated Dogs (*Canis Familiaris*) Tend to Follow Repeated Deceptive human Cues Even When Food Is Visible', *Learning & Behavior*, 46: 442–8.
Fitzpatrick, S. (2008) 'Doing away with Morgan's Canon', *Mind & Language*, 23: 224–46.
Glock, H.-J. (2019) 'Agency, Intelligence and Reasons in Animals', *Philosophy*, 94: 645–71.
Hurley, S. (2003) 'Animal Action in the Space of Reasons', *Mind & Language*, 18: 231–57.
Korsgaard, C. M. (2018) *Fellow Creatures: Our Obligations to the Other Animals*. Oxford: OUP.
Krachun, C. et al. (2016) 'Smoke and Mirrors: Testing the Scope of Chimpanzees' Appearance–Reality Understanding', *Cognition*, 150: 53–67.
Lord, E. (2018) *The Importance of Being Rational*. Oxford: OUP.
Melis, G. and Monsó, S. (2024) 'Are Humans the Only Rational Animals?', *The Philosophical Quarterly*, 74: 844–64.
Pollock, J. L. (1974) *Knowledge and Justification*. Princeton, NJ: PUP.
Raz, J. (1975) *Practical Reason and Norms*. London: Hutchinson.
Scanlon, T. (1998) *What We Owe to each Other*. Cambridge, MA: HUP.
Schleihauf, H. et al. (2025) 'Chimpanzees Rationally Revise Their Beliefs', *Science*, 390: 521–6.
Singley, M. and Anderson, J. (1989) *The Transfer of Cognitive Skill*. Cambridge, MA: HUP.
Sylvan, K. (2015) 'What Apparent Reasons Appear to be', *Philosophical Studies*, 172: 587–606.
Takaoka, A. et al. (2015) 'Do Dogs Follow Behavioral Cues from an Unreliable Human?', *Animal Cognition*, 18: 475–83.
Taylor, A. et al. (2022) 'The Signature-Testing Approach to Mapping Biological and Artificial Intelligences', *Trends in Cognitive Sciences*, 26: 738–50.
Veit, A. and Niederl, M. (ms) 'There's a Reason to Hide Only When Being Watched: Do Dogs Respond to Higher-order Reasons?'
Wedgwood, R. (2017) *The Value of Rationality*. Oxford: OUP.