EXPERIMENT AND ANIMAL MINDS: WHY STATISTICAL CHOICES MATTER

PSA 2014
Forthcoming in *Philosophy of Science*

Irina Mikhalevich, Washington University- St. Louis

ABSTRACT

Comparative cognition is the interdisciplinary study of nonhuman animal cognition. It has been criticized for systematically underattributing sophisticated cognition to nonhuman animals, a problem that I refer to as the *underattribution bias*. In this paper, I show that philosophical treatments of this bias at the experimental level have emphasized one feature of the experimental-statistical methodology (the preferential guarding against false positives over false negatives) at the expense of neglecting another feature (the default, or *null*, hypothesis). In order to eliminate this bias, I propose a reformulation of the standard statistical framework in comparative cognition. My proposal identifies and removes a problematic reliance on the value of parsimony in the calibration of the null hypothesis, replacing it with relevant empirical and theoretical information. In so doing, I illustrate how epistemic and non-epistemic values can covertly enter scientific methodology through features of statistical models, potentially biasing the products of scientific research. Broadly construed, this paper calls for increased philosophical attention to the experimental methodology and statistical choices.

INTRODUCTION

Comparative cognition is the interdisciplinary study of the evolution, development, and function of cognitive processes and mechanisms in nonhuman animals. A central controversy within the field concerns ways of guarding against bias in the course of interpreting nonhuman animal (henceforth, animal) behavior. As many philosophers and scientists have written, the worry is disproportionally aimed at guarding against overattribution of sophisticated cognition to animals (Sober 2005, Andrews 2011, de Waal 1998). For the purposes of this paper, I take it as a given that comparative cognition researchers as a group prefer explanations with the most austere cognitive ontologies, and that this practice results in an *underattribution bias*, or the systematic underascription of putatively complex cognition to animals (Andrews 2011, Fitzpatrick 2008, Meketa 2014). This preference is typically cashed out in terms of taking putatively simple mechanisms, processes, or abilities as the default experimental hypothesis.

In this essay, I locate the mechanism that drives the underattribution bias within the choice of the statistical *null* hypothesis (H_0). I argue that the manner in which the null hypothesis is currently chosen embeds a parsimony-based preference for simple cognitive ontologies. Having identified the mechanism driving the underattribution bias, I recommend removing that mechanism from the statistical methodology in which it is embedded, and replacing it with a procedure that is sensitive to empirical information. In so doing, I offer a case study of how values, such as parsimony, may come to play a central, though implicit, role in scientific methodology.

1. LOCATING THE SOURCE OF THE UNDERATTRIBUTION BIAS: THE NEYMAN-PEARSON METHOD OF HYPOTHESIS-TESTING

In the case of comparative cognition, the dominant statistical analysis method is what is known as the Neyman-Pearson Method (NPM) of hypothesis testing. Although the NPM is not the only statistical system available to science – there are also Bayesian and likelihoodist methods – it is the orthodoxy in comparative psychology, and, by extension, in comparative cognition.[1] The NPM includes what Peter Godfrey-Smith (1994) calls the *error-rate asymmetry*, which calls for preferring one type of hypothesis over another. Because the hypothesis typically preferred in comparative cognition is the one positing the simplest cognitive ontology, the error-rate asymmetry results in the underattribution bias. The remainder of this section explores the results of the error-rate asymmetry and sets the stage for my proposed solution for eliminating the underattribution bias. My solution works within the dominant paradigm of the NPM, retaining its desirable features, but offering a means of eliminating the parsimony-based underattribution bias. Put another way, my solution should be viewed as a *reformation* rather than as a *revolution*.

What exactly is the NPM? Put simply, the NPM is a method for controlling the error-rates (long-run relative frequencies) of two types of errors, which are labeled Type I errors and Type II errors. Type I errors, in general, are *defined* as those that are most serious. In the Neyman-Pearson tradition, the assumption is that the most serious type of error is the one that rejects the null hypothesis (H_0) when the H_0 is true. Within this paradigm, accepting the H_0 when

[1] For challenges to the orthodoxy of the NPM, see Kruschke (2010), who advocates replacing it with Bayesian analysis, and Anderson et al. (2000), who favor a version of likelihoodist methods to the allegedly "unscientific" null hypothesis testing methods. For philosophical defenses of the NPM, see Mayo (1981) and Mayo (1992). For a "severity-analysis" reformulation of the NPM see Mayo (2004) and Mayo and Spanos (2009, 2011).

it is false is a Type II error, and it is treated as less serious. Type I error *rates* are denoted as α and Type II error rates as β. To modify slightly Dienes's (2008) formalization of the relationship between error types and their relative frequencies, we may say that

$$\alpha =_{def} \mathbf{P(rejecting\ H_0 \mid H_0)}$$
$$\beta =_{def} \mathbf{P(accepting\ H_0 \mid \neg H_0).}$$

Although the NPM provides a means for controlling error rates in a way that minimizes the risk of making both types of errors, researchers have traditionally set the risk of a Type I errors lower than Type II errors. Typical values for α are .05, .01, and sometimes .001. Treating Type I errors as more serious translates into controlling for Type I errors by making α very small, while keeping β either large or not controlled at all. This preference for making Type II errors over Type I errors is the *error-rate asymmetry*.

2. The Importance of the Null Hypothesis

With this background complete, it is now possible to see how earlier assessments of the underattribution bias in comparative cognition have correctly located the source of the bias at the statistical level, but preferentially emphasized the error-rate asymmetry over what I will now argue is the real mechanism: the *null hypothesis*.

Earlier Solutions: Locating the Underattribution Bias in the Error-Rate Asymmetry

Elliott Sober (2001, 2005) argues that there is no reason to prefer making Type II errors over Type I errors in comparative cognition and that this preference is furthermore a misapplication of MC, understood as a parsimony principle. According to Sober, not only is MC *not* a parsimony principle, but both types of errors are equally undesirable, since both errors are equally wrong. Sober advocates ridding the field entirely of the error-rate asymmetry, arguing that, “the only prophylactic [against risk of error] is empiricism” (Sober 2005, 97).

Building on Sober’s work, Andrews (2011) identifies the preference for Type I error with an exaggerated and damaging worry over the alleged systematic overattribution of sophisticated cognition to animals.[2] She agrees with Sober that the matter is an empirical one, but departs from his conclusion regarding the seriousness of each type of error. She argues that Type I errors are in fact *more* damaging than Type II errors because they foreclose on the possibility of future research (Andrews 2011). On her view, preferring to make Type I errors means preferring to wrongly conclude that, e.g., the New Caledonian crows do not use planning to solve puzzles. Once such a judgment has been made, it no longer makes sense to ask further questions about the features of e.g., the crows’ future-planning abilities, such as whether they are domain-specific or general, available only with appropriate environmental scaffolding, and so on. As a result, a potentially fruitful research program never gets a chance to get off the ground. Type II errors, on the other hand, promote a further refinement of experimental questions. These questions may produce results that conflict with the original (mistaken) judgment, but, argues Andrews, science must be willing to take such risks.

Moving to the Null Hypothesis

Despite discussing the biasing effects of the error-rate asymmetry, neither Sober (2001, 2005) nor Andrews (2011) question the fact that the H_0 is treated as the absence of the mental feature

[2] Andrews refers to this overattribution fear as the fear of so-called anthropomorphism, or the attribution of allegedly uniquely human cognitive complexity to nonhuman animals. For in-depth analyses of the alleged mistake of anthropomorphism, or the mistaken attribution of human properties to nonhuman entities, in comparative psychology, see Fisher (1990; 1991). See also Keeley (2004) for an update to Fisher’s arguments.

under investigation.[3] However, as I will now show, the choice of the H_0 is just as likely to be a source of bias as the error-rate asymmetry. Once the H_0 is understood to be a source of bias, a solution to the underattribution problem will become clear.

Consider Andrews's claim that the error-rate asymmetry results in what she calls a behavioristic bias. She is right as long as the H_0 is defined as the absence of a cognitive feature. However, if the H_0 were defined as the *presence* of rich cognitive abilities, the result would be the *opposite* of a bias toward underattribution: comparative cognition would be biased toward *overattribution*. Such a dramatic difference in the outcome of the application of our procedural rules points to the significance of the construction of the $H_{0,}$ i.e., the choice of how it is to be defined. If the construction of the H_0 is so important to the final outcome of a given hypothesis-testing procedure, then we must pay more careful attention to how we come to identify something as the H_0.

To illustrate the importance of attending to the construction of the H_0 more concretely, consider a case where replacing the H_0 while retaining the error-rate asymmetry results in a bias toward sophisticated cognitive explanations. Let us take a closer look at an experiment by Allison Foote and Jonathom Crystal (2007), which used a duration-discrimination task to test for metacognition – awareness of one's own mental states – among rats. In this case, the mental state in question was that of uncertainty. Meketa (2014) describes the experiment as follows:

> "[Rats] were presented with audio tones of different durations and trained to classify the tones into the categories of "short" or "long." The rats were then presented with a range of tones, some clearly short and others ambiguous. Correct responses were rewarded with food, and incorrect responses were not rewarded at all. Next, the rats were given

---

[3] Since the present essay was written, Andrews and Brian Huss have written, but, to my knowledge, not yet published a manuscript that includes an explicit discussion of the role of the null hypothesis (Andrews & Huss *unpublished manuscript*).

> the same test, but were given a third option: to decline a test. Declined tests allowed the subjects to move on to more tests with the prospect of getting more food. When given the choice to decline tests, the rats consistently opted to decline the ambiguous ("harder") tests but not the unambiguous tests, even though declining a test resulted in a smaller food reward than answering correctly. Moreover, the overall accuracy improved when rats were allowed to opt out of difficult tests. Foote and Crystal (2007) concluded that the rats were aware of their own uncertainty." (Meketa 2014)

In other words, Crystal and Foote concluded that this behavior showed that the rats were metacognitive.

Let us now abstract away from the details of the experimental setup and just consider the hypothesis being tested**.** We see that their H_0 was that the rats do *not* possess metacognition. The alternative hypothesis – the one they wished to demonstrate – was that the rats *are* capable of metacognition. Given the error-rate asymmetry, the burden of proof falls on the metacognitive hypothesis. In fact, in a follow-up paper on metacognition, Crystal and Foote (2009) clearly state that the default hypothesis – the H_0 – is *and should be* that rats lack metacognitive capacities. The reason, they argue, is that the behaviors they observed in the 2007 trials *could* be explained by allegedly simpler mechanisms, such as associative learning, which is presumed to be incompatible with metacognition.[4] This means that the metacognitive explanations bear the burden of proof.

But now consider what would happen if the H_0 in Foote and Crystal's experiments were a rich cognitive explanation of the rats' behavior (e.g., H_0 = "rats are capable of metacognition").

---

[4] The standard view that associative mechanisms are different from and simpler than cognitive processes has been coming under scrutiny in recent years. For example, Cameron Buckner (2012*)* argues against the view that cognitive and associative systems are incompatible. Taking issue with the assumption that association is simple, Gallistel (2008) argues that associative mechanisms are more demanding than cognitive systems insofar as they would require far greater energy expenditures than alternative mechanisms. He uses the honeybee navigation system to argue that the honeybee brain does not have enough computing power to process information through associations alone, and must require a representational system of mental maps (Gallistel 2008).

Then the burden of proof would be on the hypotheses positing less sophisticated cognition (e.g., H_1 = "rats are relying on stimulus-response learning"). In this case, a preference for Type I errors over Type II errors would mean a preference for accepting (or failing to reject) the hypothesis that rats possess metacognitive abilities when, in fact, the rats do not. As a result, the underattribution bias would be inverted.

The metacognition example suggests that the way that the H_0 is constructed is *at least as important* as the error-rate asymmetry when it comes to assessing an experimental methodology for built-in theoretical commitments. What attending to the construction of the H_0 reveals is that, while the asymmetry introduces a bias, the nature of this bias is specified by the content of the H_0. In one sense, the role of the H_0 may be *more* important than the error-rate asymmetry: while the asymmetry can only be made more or less pronounced, the content of the H_0 can embed any number of problematic assumptions.

This conclusion prompts the question: Why, if the content of H_0 is so important, have scientists and philosophers of science assumed that the H_0 is naturally defined as "non-presence" or "no effect"? It is curious that a feature that carries such powerful implications for inference from experiment should be casually assumed to be globally fixed at the non-presence of the target cognitive property. In order to explain why the content of the H_0 has been systematically overlooked, I turn to Peter Godfrey-Smith's (1994) analysis of the NPM and the possible justifications for its use. Placing the NPM into its historical context will, furthermore, motivate my suggestion that the NPM can be modified as I suggest in §5.

3. The NPM: A Changing Justification and the Introduction of Parsimony

According to Godfrey-Smith (1994), the original justification for the NPM was pragmatic, but that justification was rejected shortly after its introduction while the method of preferentially controlling for Type I errors was retained. Contrary to Andrews's claim, the original NPM included an accept/reject procedure. However, the original, pragmatic, justification of the 'accept/reject' decision-procedure was intended as a *behavioral* strategy, where "accepting" an hypothesis meant *acting as if* the hypothesis were true (Godfrey-Smith 1994, 280-82). This pragmatic justification meant that the NPM could not be used to support belief in the *truth* of an hypothesis or even in the *probability* of the hypothesis being true. This pragmatic justification did not sit well with subsequent scientists and statisticians (e.g., R.A. Fisher), who wanted a statistical system to provide *evidence* for the truth or falsity of an hypothesis – something that the original NPM explicitly avoided (Dienes 2008, Anderson et al., 2000, Gigerenzer 2004). The result, according to Godfrey-Smith, was a proliferation of alternative justifications that have in turn altered the method in unexpected ways.

One alternative justification – which Godfrey-Smith labels the 'semantic' justification[5] – includes the concept of what he calls a 'natural null,' or H_n, which is typically defined as the hypothesis of no effect or no difference. On the semantic justification, the Type I error is a wrong rejection of the hypothesis of no effect, or no difference. Since Type I errors are considered more serious, the semantic justification advises erring on the side of concluding that no effect or difference was detected. Moreover, the accept/reject procedure is interpreted both behaviorally and epistemically (Godfrey-Smith 1994, 287). I wish to focus on what Godfrey-Smith calls the 'semantic' justification for the NPM, because this is the most popular justification in psychology and, hence, also in comparative psychology and comparative cognition. It is worth noting that the other two justifications that Godfrey-Smith discusses, the 'pragmatic' and the 'doxastic,' do not specify a value for the H_0, holding that "H_0 is 'true' if the world is in a state

[5] Godfrey-Smith labels justification as "semantic" because it specifies the semantic content of the null.

such that the action associated with H_0 is better than the alternative action" (Godfrey-Smith 1994, 281).[6]

Crucially, Godfrey-Smith identifies a curious metaphysical principle embedded in the semantic NPM: When combined with the error-rate asymmetry, the H_n results in a preference for *nothing* over *something*, that is an Occamist commitment to maximally simple ontologies and theories that favor such ontologies. Since the semantic justification is, according to Godfrey-Smith's analysis, the most common interpretation in science – sometimes combined with the doxastic justification – it follows that the sciences that use it encode a commitment to Occamist metaphysics. Godfrey-Smith argues that psychology uses the semantic justification almost exclusively, though this is sometimes combined with a doxastic justification. Since comparative cognition is to a large extent constituted by comparative psychology, it is no surprise that Occamism is present in comparative cognition's statistical methodology as well.

Finally, if indeed the content of the H_0 is at least as significant as the error-rate asymmetry for identifying bias in comparative cognition, then Sober and Andrews have been focusing on a feature that only becomes a problem under conditions in which the null hypothesis is biased. It is possible to retain the asymmetry found in the NPM without accepting the question-begging conservatism about animal minds. My analysis recommends that the justifications for the NPM be carefully re-examined to avoid smuggling in a priori theoretical commitments.

---

[6] By contrast with this behavioral "pragmatic" interpretation, the doxastic justification for the NPM is epistemic. It replaces the pragmatic component with the rule that "when an observation in the critical region [the set of values that would cause us to reject the hypothesis] occurs the researcher rejects H_0. But when an observation falls outside the critical region the researcher *merely suspends judgment*" (Godfrey-Smith 1994, 282; emphasis added).

It is evident that statistical methods are often considered value-neutral and, in that respect, objective. However, I have shown that values may be embedded in these statistical methods. In the case of comparative cognition, this value is Occamist parsimony, and it is located in the choice of the null hypothesis within the NPM. Moreover, as a statistical methodology comes to be used as a standard in a given field, the values embedded in the method fade from scientific consciousness. The result is that, while researchers and philosophers appreciate the potential for a gerrymandering of data by cherry-picking statistical analyses, values, such as parsimony, continue to operate in the background methodology itself, without being subject to direct scrutiny. Analyses such as the one I offer here, are, therefore, crucial for uncovering and assessing the effects of values even in such inconspicuous places as tools for statistical analysis of experimental data. So much for a partial account of the invisibility of the null in the philosophy and science of comparative cognition. The next question is how the underattribution bias may be corrected. Given that comparative cognition researchers are unlikely to abandon the NPM in the near future, I propose a reformation of the semantic NPM.

4. The NPM Reformed: Replacing the Natural Null with a Contextual Null

I have argued that the underattribution bias is driven by a parsimony-based preference for purportedly simple cognitive ontologies, and that this practice is regimented in the preference for a H_n. I will now show how to alter the "semantic" NPM in a manner that eliminates this bias. My strategy involves modifying the semantic NPM to replace the natural null with what I call a "contextual null" (H_c), which reflects a broader epistemic context for the animals under investigation.

To begin, note that the semantic view does not require that the H_0 must always be that the feature under investigation is absent. It is, however, this particular definition that lends the semantic view its bias toward Occamist parsimony. It follows that removing the Occamism from the semantic NPM requires removing the natural null. This is precisely what I now suggest: the natural null should be replaced with a contextual null. In contrast to the H_n, the H_c is defined against a suite of background information about the research subjects, such as ontogenetic and phylogenetic information against the background of developmental and evolutionary theories, information about species-typical and individual behavioral profiles, neuroanatomical homologies[7] and homoplasies,[8] ecological context, and information from earlier studies and observational data.[9] I call such a null "contextual" for two reasons: (1) it respects the differences among experimental settings and the organisms being studied, and (2) it does not presuppose either cognitive complexity or cognitive simplicity. It is sensitive to the changing conditions between experiments, both in terms of the kinds of questions that are asked and with respect to how much is known about the target system.

The evolutionary considerations that I propose to be taken into account in constructing the H_c include the species' phylogenetic proximity to species about whom more is known in order to gauge the likelihood of homologous cognitive structures and abilities. These considerations already enter into decisions about which species to study when searching for a given ability, but not into the decision to cast a given hypothesis as the presumptive null. For example, chimpanzees' close phylogenetic proximity to humans is a frequently cited reason to test them for the presence of human-like abilities, such as tool use and metacognition.

[7] A homology is "a similarity inherited from a common ancestor" (Sober 2005, 94).

[8] A homoplasy is "a similarity that is the result of two or more independent derivations of the trait" (ibid).

[9] Fitzpatrick (2008) draws a very similar conclusion about the need for background information in hypothesis testing. However, he does not frame his case in statistical terms. His account is intended to displace the parsimony-based reading of Morgan's canon, understood as a heuristic, with a principle he calls "Evidentialism."

The developmental considerations relevant to constructing the H_c include, inter alia, hypotheses regarding developmental constraints on the evolution of the relevant cognitive and behavioral traits and the effects of the environment on gene expression. Ecological context would include information about the test subjects' behavior in its natural habitat, such as whether it is a social or solitary animal, whether it hunts or stores its food, whether it uses tools, and so on. Once again, these considerations already drive the research projects, suggesting that researchers consider such information to be probability-conferring. Consider the following example of research that has been guided by both ecological and developmental considerations. Furlong et al. (2008) tested chimpanzees for their ability to use tools based on the knowledge that chimpanzees use tools under natural conditions (e.g., dipping sticks into ant mounds to catch ants; using leaves to scoop up water). Based on the negative results obtained by a previous study by Daniel Povinelli (2000), which concluded that chimpanzees lack the competency for flexible use of implements, Furlong et al. hypothesized that Povinelli's chimps were developmentally stunted as a result of being brought up under socially impoverished conditions. Furlong et al. tested chimpanzees with different socialization backgrounds, and found that the ability to manipulate tools in a flexible manner (i.e., one suggestive of causal understanding) was positively correlated with social histories. This example shows the value of social ecology and its effects on chimpanzee intellectual development. Building such considerations into the null hypothesis would ensure that crucial information is not left out of the experimental design.

Finally, the H_c should include information from previous studies and observational data. However, given the arguments of the foregoing sections, including the work of earlier experimental studies would require re-evaluating them for the presence of bias. This can be achieved by analyzing the choice of null hypothesis to ensure that unwarranted metaphysical preferences have not been smuggled in and that relevant empirical and theoretical information was included.

The upshot is that my proposed H_c ensures that Type I errors will always be more epistemically serious because probability-conferring evidence is built into the H_0. Sometimes this method will produce a H_0 positing a simple cognitive ontology, but this will no longer be based on a blanket Occamist preference for simple ontologies, but on an empirically-informed expectation. This suggestion respects the intuition that default hypotheses ought to be those that we have the best reason to adopt. In the end, my account preserves the risk-controlling structure of the semantic account of the NPM while eliminating the questionable metaphysics and replacing it with empirical information.

CONCLUSION

I have argued that earlier assessments of bias in comparative cognition at the level of statistical data evaluation have ignored an important feature of the orthodox NPM –namely the null hypothesis. I have suggested that this lacuna may be attributed to a specific interpretation of the NPM – the semantic justification – which assumes that the null hypothesis must be universally set to a natural null of "no difference" or "no effect." I have suggested that if the semantic version of the NPM commits the researcher to a position supported only by a problematic Occamist metaphysics, then the semantic version needs to be modified. My proposed modification to the NMP maintains the error-rate asymmetry, but replaces the H_n with the H_c. This change respects the intuition that the burden of proof should be on the hypothesis that has the least empirical and theoretical evidence on its side.

At a more general level, the foregoing discussion provides a case study of how metaphysical assumptions, such as a preference for simple ontologies, can enter science at the level of statistical model choice. The fact that such metaphysical assumptions can be grafted onto the standard statistical models of an entire field suggests the need for a more careful scrutiny of statistical models, as philosophers of statistics, such as Deborah Mayo have been arguing for years. I showed that the data processing instruments used to generate inferences may not be value-free. Whether a value-free statistical instrument is desirable is an open question, but recognizing its value-laden dimensions is a necessary step in evaluating the conclusions drawn from scientific experiments.

REFERENCES

Anderson, David R., Kenneth P Burnham and William L. Thompson. 2000. "Null Hypothesis Testing: Problems Prevalence, and an Alternative." *The Journal of Wildlife Management* 64: 912-23

Andrews, Kristin. 2011. "Beyond Anthropomorphism: Attributing Psychological Properties to Animals." In *The Oxford Handbook of Animal Ethics,* edited by Tom Beauchamp and R.G. Frey. Oxford: Oxford U.P.

Buckner, C. 2011. "Two Approaches to the Distinction between Cognition and 'Mere Association.'" *International Journal of Comparative Psychology* 24: 314 – 348

Crystal, Jonathon D. and Allison L. Foote. 2009. "Metacognition in animals." *Comparative Cognition & Behavior* 4: 1 – 16.

Dienes, Zoltan. 2008. Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference. *Palgrave Macmillan*.

De Waal, Frans B. M. 1999. "Anthropomorphism and Anthropodenial." *Philosophical Topics* 27: 255 – 280

———. 1997. Foreword in *Anthropomorphism, Anecdotes, and Animals,* edited by Robert W. Mitchell, Nicolas S. Thompson, and H Lyn Miles, xii – xvii. Albany: SUNY Press.

Fitzpatrick, Simon. 2009. "The primate mindreading controversy: a case study in simplicity and methodology in animal psychology." In *The Philosophy of Animal Minds,* edited by Robert W. Lurz, 237 – 257. New York: Cambridge U.P.

———. 2008. "Doing Away with Morgan's Canon." *Mind & Language* 23: 224-246.

Foote, Allison L. & Jonathon D. Crystal. 2012. "'Play it Again': a new method for testing metacognition in animals" *Animal Cognition* 15: 187 – 199

———. 2007. "Metacognition in the rat." *Current Biology* 17: 551 – 555.

Furlong, E.E, K.J. Boose, and S. T. Boysen. 2008. "Raking it in: the impact of enculturation on chimpanzee tool use." *Animal Cognition* 11: 83 – 97.

Godfrey-Smith, Peter. 1994. "Of Nulls and Norms." *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1: 280 – 290.

Heyes, Cecilia. 2012. "Simple Minds: A Qualified Defense of Associative Learning." *Philosophical Transactions of the Royal Society: Biological Sciences* 367: 2695–2703

Keeley, Brian L. 2004. "Anthropomorphism, Primatomorphism, Mammalomorphism: Understanding cross-species comparisons." *Biology & Philosophy* 19: 521 – 540.

Kruschke, J. K. 2010. "What to believe: Bayesian methods for data analysis." *Trends in Cognitive Sciences* 14: 293-300.

Le Pelley, M. E. 2012. "Metacognitive Monkeys or Associative Animals simple Reinforcement Learning Explains Uncertainty in Nonhuman Animals." *Journal of Experimental Psychology: Learning, Memory, and Cognition.* Accessed January 16, 2012. doi: 10.1037.a0026478

———. 1886. "On the Study of Animal Intelligence" *Mind* 11: 174 – 185.

Mayo, Deborah G. 2004. "An Error-Statistical Philosophy of Evidence." in *The Nature of Scientific Evidence,* ed. Mark L. Taper and Subhash R. Lele, 79-118. Chicago: Chicago University Press.

——.1992. "Did Pearson Reject the Neyman-Pearson Philosophy of Statistics?" *Synthese* 90: 233-262.

——. 1981. "In Defense of the Neyman-Pearson Theory of Confidence Intervals." *Philosophy of Science* 48: 269-280.

Mayo, Deborah & Aris Spanos. 2011. "Error Statistics," In *Handbook of the Philosophy of Science, Volume 7: Philosophy of Statistics*. Edited by Prasanta S. Bandyopadhyay and Malcolm R. Forster. 153-198. Philadelphia: Elsevier Inc.

Meketa, Irina. 2014. "A Critique of the Principle of Simplicity in Comparative Cognition" *Biology & Philosophy* doi: 10.1007/s10539-014-9429-z

Sober, Elliot. 2009. "Parsimony and Models of Animal Minds." In *The Philosophy of Animal Minds,* edited by Robert W. Lurz, 237 – 257. New York: Cambridge U.P.

———. 2005. "Comparative Psychology meets Evolutionary Biology: Morgan's Canon and Cladistic Parsimony." In *Thinking with Animals: New Perspectives on Anthropomorphism,* edited by Lorraine Datson and Gregg Mitman, 85 – 99. New York: Columbia U.P.

Wimpenny J.H, A.A.S. Weir, L. Clayton, C. Rutz, A. Kacelnik. 2009. "Cognitive Processes Associated with Sequential Tool Use in New Caledonian Crows." *PLoS ONE* 4: 6471.