Title:
Morgan's Canon, Meet Hume's Dictum: Avoiding Anthropofabulation in Cross-Species Comparisons

Cameron Buckner
Visiting Assistant Professor
University of Houston
513 Agnes Arnold Hall
Houston, TX 77204-3004
Phone: 713-743-3010
Fax: 713-743-5162

Humboldt Postdoctoral Fellow
Ruhr-University Bochum

cjbuckner@uh.edu

Abstract: How should we determine the distribution of psychological traits—such as Theory of Mind, episodic memory, and metacognition—throughout the Animal kingdom? Researchers have long worried about the distorting effects of anthropomorphic bias on this comparative project. A purported corrective against this bias was offered as a cornerstone of comparative psychology by C. Lloyd Morgan in his famous "Canon". Also dangerous, however, is a distinct bias that loads the deck against animal mentality: our tendency to tie the competence criteria for cognitive capacities to an exaggerated sense of typical human performance. I dub this error "anthropofabulation", since it combines anthropocentrism with confabulation about our own prowess. Anthropofabulation has long distorted the debate about animal minds, but it is a bias that has been little discussed and against which the Canon provides no protection. Luckily, there is a venerable corrective against anthropofabulation: a principle offered long ago by David Hume, which I call "Hume's Dictum". In this paper, I argue that Hume's Dictum deserves a privileged place next to Morgan's Canon in the methodology of comparative psychology, illustrating my point through a discussion of the debate over Theory of Mind in nonhuman animals.

I. Introduction

C. Lloyd Morgan, one of the founders of comparative psychology, worried that his new science would be vulnerable to anthropomorphic bias, here taken as the tendency to attribute human characteristics to non-human entities on the basis of insufficient or superficial evidence. To mitigate this bias, he recommended a corrective, which has come to be known as "Morgan's Canon". The Canon states that "in no case may we interpret an action as the outcome of the exercise of higher psychological processes, if it can be fairly interpreted in terms of processes which stand lower in the scale of psychological evolution and development" (Morgan, 1903). Since its publication, nearly every aspect of the Canon has been the subject of misinterpretation (Radick, 2000; Thomas, 2006), but its proper interpretation and justification need not concern us here (for that see Sober, 1998, 2005). The important point for present purposes is that the Canon—or, at least, a suitable replacement principle, should we decide that the Canon has been so misapplied as to be beyond salvage (Fitzpatrick, 2008)—provides useful guidance in countering anthropomorphic bias with careful experimental design. In particular, it has reminded comparative cognition researchers to rule out alternative explanations of data—especially reflexes, innate-releasing mechanisms, and basic forms of classical and operant conditioning—by performing adequate controls and devising problems or tasks that could not be solved by these ubiquitous contrast classes. Where this methodology succeeds, one avoids anthropomorphic bias by eschewing the use of our folk interpretive tendencies to decide whether a "higher" explanation of some data is legitimate. So long as experiments have been properly designed to arbitrate between competing hypotheses, as Sober (2005, 97) puts it, "the only prophylactic we need is empiricism."[1]

[1] In practice, many comparative psychologists read the Canon in a stronger way, as recommending the default position that animal behavior is driven by lower processes unless evidence can be provided that only explanation in terms of higher processes is adequate. On the weaker reading adopted throughout this article, we should instead

There is, however, another serious problem plaguing comparative research against which Morgan's Canon provides no protection. The issue is that the psychological abilities of animals largely fall somewhere in-between the "lowest" forms of reflexes and associative learning and the "highest" forms of cognition, and we currently possess only the rudiments of a psychological taxonomy adequate to characterize the range of similarities and differences. As such, comparative psychologists have set off to investigate whether animals possess capacities such as Theory of Mind (ToM), episodic memory, and metacognition, before knowing precisely what types or degrees of similarity would be relevant. These concepts are not even done baking in human psychology, and so deciding how to apply them to animals is especially fraught (Emery & Clayton, 2009; Shettleworth, 2009). Since a decision on what these terms should mean is a decision about what abilities comparative psychologists should investigate, steps here should be taken carefully.

In approaching these challenges, the debate over whether animals have a ToM serves as an excellent case study, for three decades of empirical research into the question have inspired vigorous debate but little consensus. Proponents of animal ToM argue that recent findings provide evidence that some animals can represent the perceptual states of conspecifics (Flombaum & Santos, 2005; Bugnyar, 2007; Call & Tomasello, 2008; Emery & Clayton, 2009), while others—most notably Penn & Povinelli (2007), but also more recently Lurz (2009, 2011)—are more skeptical. They argue that these results can more parsimoniously be explained in terms of "behavior-reading" capacities—that is, capacities sensitive only to contingencies between observable cues (such as gaze direction and body orientation) and their behavioral outcomes. A puzzling feature of this debate (as well as similar comparative debates over other capacities such as episodic memory and metacognition) is that the controversy has only deepened as experimental results have accumulated.

This failure to converge on consensus suggests deeper disagreements over methodology, evidence, and interpretation. Indeed, several methodological and epistemic components of this dispute have been well-explored by philosophers and psychologists, including: the dangers of anthropomorphic bias in the interpretation of experimental results (Keeley, 2004; Wynne, 2007), the issue of which hypotheses are more parsimonious (Fitzpatrick, 2009; Heyes, 1998), the falsifiability of behavior-reading or ToM-based hypotheses (Fletcher &

---

remain agnostic about results that could be explained by either higher or lower causes. I do not here explore the question as to which interpretation was intended by Morgan (for that see Richards (1989) and Radick (2000), who reach somewhat different conclusions).

Carruthers, 2012), the ecological validity of experiments (Povinelli & Vonk, 2003; Tomasello, Call, & Hare, 2003), and whether experimenters are even performing the right kind of task to elicit ToM (Andrews, 2012; Andrews, 2005). In each case, promising methodological correctives have been suggested, though at present the controversy endures.

By contrast, the semantic dimensions of these debates have largely been neglected. This is unfortunate, for whether animals possess ToM, episodic memory, or metacognition surely depends upon what we mean by 'ToM', 'episodic memory', and 'metacognition'. However, we simply do not yet know precisely what these terms of art mean; they are vague in the sense that they have a large space of borderline cases for which there is no consensus as to whether they apply.[2] Moreover, a quick literature scan suggests that 'ToM' has been interpreted in significantly different ways by proponents and skeptics, and as a result applied to the same data in different ways. For example, both sides of the dispute over ToM agree that chimpanzees have demonstrated some success in responding appropriately to the perceptual states of conspecifics in competitive situations but have largely failed to respond appropriately in cooperative situations. Proponents tend to favor graded notions of 'ToM' according to which success in competitive contexts alone could be sufficient; Bugnyar (2007, 15) suggests that "'full-blown' ToM does not represent one single mechanism but is composed of a set of skills", Call & Tomasello (2008) recommend a "broad construal of the phrase 'theory of mind'" in which chimpanzees might display ToM in ecologically-valid contexts but not in others, and Santos, Flombaum, & Phillips (2007, 445) write that their subjects' ability to "[finely-tune] to exactly the variables, postures, and behaviors in their environment that are relevant to problems of social reasoning [in competitive situations alone]…in essence boils down to a ToM system."[3] Against these more permissive interpretations, skeptics such as Penn & Povinelli (Penn & Povinelli 2012, 16) assert that "the essence of

[2] This vagueness could be either epistemic or metaphysical. If the disputed terms are natural kind terms, for instance, then they may really have sharp boundaries, and borderline cases are a temporary product of our ignorance of these kinds' underlying natures. Alternatively, if defining a term requires ineliminable appeal to evaluatives—like "reliable" or "intelligent"—then vagueness may be a permanent feature of its associated concept.

[3] Listing these proponents together is not meant to suggest that there are not subtle differences amongst their positions—only that they share a common thread, denied by Povinelli and colleagues, that existing data from experiments on competitive situations alone provides good evidence that some non-human primates have a perceptual ToM.

[a] ToM …is the ability to explicitly represent (i.e., predicate) and reason about the causal role played by a given mental state *across disparate behavioral contexts*" (my emphasis), and thus that none of these context-bound analogues or precursors are worthy of the name.

The skeptical claim that these islands of social understanding do not count as ToM is not at bottom empirical, but rather semantic. And worse, psychologists have no established methods to responsibly evaluate this kind of semantic question. As a result, once a skeptic stomps around claiming that some pinnacle of human cognitive achievement is essential to a psychological capacity, the comparative waters around that capacity are muddied for everyone. Of course, this is not to suggest that the debate is *merely* semantic—as though the issue could be settled by conceptual analysis. As we will see, the semantic question is closely intertwined with the other methodological issues mentioned above, and attempts to precisify the disputed terms must be evaluated at least in part by their empirical adequacy—by how well precisifications can support fruitful empirical inquiries. Neither, however, can these debates be resolved solely by conducting more or better experiments, without also reaching consensus on what ought to count as "genuine" ToM, episodic memory, and metacognition.

Despair might set in with the realization that different research groups are talking past one another in these debates, prompting some broad points of consensus. Both sides should move away from pass/fail tests and sweeping claims, rely more on predictions generated from precise computational models of capacities rather than informal intuitions, and be more concerned with what each species can do across a range of problems and contexts rather than whether they perform as well as humans on a few artificial tasks (Shettleworth, 2009). Perhaps that ought to be the end of the discussion; but to retreat from difficult semantic questions entirely is to simply give up on comparative psychology, replacing it with human psychology, chimpanzee psychology, raven psychology, and so on. This would come with costs; we would never know which capacities animals share with us (and each other), and how alike are the roles they play in our and their lives. So—and here comes a big conditional claim—if we find these comparative questions interesting enough to merit empirical investigation, then we ought to dig in our heels somewhere and develop ground rules for drawing those lines in a responsible and useful way.

In this paper, I propose codifying one such principle, which I dub "Hume's Dictum" after its author, the philosopher David Hume (1711-1776). Analogues of Hume's Dictum have occasionally popped up in these debates, but it has not yet been formally recognized as a principle of equal importance with other well-known principles such as Morgan's Canon. As a result, researchers who would otherwise endorse the Dictum have sometimes subtly

violated it. While I do not propose that the Dictum can settle all unclarity surrounding 'ToM' or the other capacities mentioned above, it does provide some much-needed semantic guidance in what is otherwise a methodological void.

In Section II, I discuss a type of semantic bias that I call "anthropofabulation", placing it with respect to two other comparative biases (anthropomorphism and anthropocentrism) and reviewing the disadvantages of succumbing to the bias. In Section III, I introduce Hume's Dictum and explain how it can help us avoid anthropofabulation, working in conjunction with other principles such as Morgan's Canon to point towards a more fruitful direction for future comparative research.

II. Anthropocentrism and Anthropofabulation

To explore our primary case study, let us return to the notion of mental state representation that skeptics suppose essential for genuine ToM. The relevant sort of representation has proven difficult to characterize, but the idea is derived from the original suggestion of Premack and Woodruff (1978) that ToM essentially involves the attribution of mental states that are not directly observable. Penn & Povinelli (2007) attempt to make this notion precise by offering a semi-formal notation in which the critical feature of ToM is the ability to represent information about another's mental states, with such higher-order representations denoted in their formalism as '*ms*' states. While they decline to provide a general theory of when one state represents information about another (though gesturing at Dretske, 1988), they propose a "stopgap" answer that an *ms* candidate does so if and only if "the state of the *ms* variable co-varies with the state of the other cognitive state in a generally reliable manner" (Penn & Povinelli, 2007, 733). Barring telepathy, however, animals lack direct perceptual access to the cognitive states of other organisms, and so the best they could do is derive another's mental state indirectly from perceivable evidence by way of a mediating theory. Animals, skeptics claim, have shown no ability to represent mental states in this way—because previous evidence can be better explained in terms of an (admittedly cognitive and sophisticated) ability to group perceptual situations into abstract classes such as *threat-posture*, *eye-* or *face-direction*, *body-position*, and *direct-line-of-gaze*.

A concern with this '*ms*'-notation is that its semantics reproduces the vagueness of the original concept of ToM that it was meant to clarify. In particular, by declining to confront the philosophical question of when one state represents another, Penn & Povinelli do not specify how reliably or across how many different contexts a

representational state must covary with an unobservable mental state to count as representing it.[4] Clearly, we cannot require perfect covariation; philosophical consensus holds that the possibility of misrepresentation is a necessary condition for representation (Dretske, 1986). Furthermore, signal detection theory illustrates that the correlation between the state of a representation and the state of its referent could be reliable enough for adaptive purposes at low levels of covariation if the tradeoff between hits, misses, false alarms, and correct rejections worked out in the right way (Macmillan, 2002, and see Table 1). For example, if missing signs of anger in a dominant can result in a subordinate's death, it will pay off for it to recruit a low-reliability anger-indicator to control fleeing movements, given that the imbalance between the costs of unnecessarily running away vs. death is large. Doing so might not reflect a deficiency of representational power in the animal in question, but rather a rational allocation of epistemic resources.

	***anger* present**	***anger absent***
candidate *ms* active	Hit	False Alarm
candidate *ms* inactive	Miss	Correct Rejection

Figure 1. Possible outcomes in Signal Detection Theory between a candidate *ms* representation and the mental state of another organism, *anger*.

The important point for present purposes is that if perfect reliability is not in the cards, we must decide how reliably and across what range of contexts one mental state must track the mental state of another to count as representing it. Humans also represent the mental states of conspecifics imperfectly, but Povinelli and other skeptics are always satisfied that at least the upper range of human-level performance is sufficient for ToM. In short, the skeptical arguments of Povinelli and colleagues all depend upon the assumption that ToM allows agents to track the mental states of others to a degree ostensively illustrated by a few experiments (many of which have never been

[4] As degree of representational reliability is a robust source of vagueness, it is likely that the treatment of this section could be extended to debates over the possession by animals of other capacities characterized in terms of representational contents—including at least metacognition (similarly described as the ability to represent one's own mental states) and episodic memory (often characterized in terms of the "what-when-where" information recorded in an episodic memory).

performed) that at least human children by a certain age are expected to pass (around 5 years of age is offered as the likely threshold), and other animals to fail (Povinelli & Vonk 2003; Penn & Povinelli 2007).

This emphasis on a distinctively human degree of representational reliability as criterial of ToM reflects an additional bias that has been called "anthropocentrism" (Emery & Clayton, 2009). Anthropocentric bias has received less theoretical attention than anthropomorphic bias, and as a result is less well-understood. Povinelli, who himself inveighs against the bias (Povinelli, 2004, 29), describes it as "[holding] the human mind [to be] the gold standard against which other minds must be judged." So stated, it is hard to see how he is not himself guilty of committing it (though see Povinelli & Vonk, 2004), but we should charitably note that anthropocentric bias can come in at least three distinct forms: methodological, evaluative, and semantic. Methodological anthropocentrism is a bias in the selection of experimental tasks on which to evaluate animals' psychological skills; it occurs when we tend to test animals on tasks at which humans excel or that are taken directly from human psychology, without consideration of the animal's own distinctive abilities or ecological niche. Evaluative anthropocentrism occurs when we tend to hold that animals are "intelligent", "interesting", or otherwise valuable only if they behave just like us. It is clear from context that it is methodological and evaluative anthropocentrism that Povinelli (2004) condemns.

Semantic anthropocentrism, however, is the form that I will focus on below, and involves precisifying vaguely-defined psychological terms to human-level ability. It is thus a form of semantic bias, similar to the tendency to demand a smaller number of hairs for baldness if we have just been primed by "Patrick Stewart" than if we had just been primed with "Cher". To add this idea to Penn and Povinelli's notation, let us denote a category precisified to require human-level ability with a "+" sign, and thus a state exhibiting human-level reliability in tracking a mental state of another agent by "*ms*+". It is this third, semantic form of anthropocentrism that Povinelli and other skeptics routinely and repeatedly commit.

Since Povinelli has denied that his arguments depend upon a controversially anthropocentric interpretation of 'ToM' (Povinelli & Vonk 2004), it will take some care to explain this charge. In their rebuttal to an earlier criticism, Tomasello et al. (2003) claim that Povinelli and his colleagues commit an anthropocentric error by adopting a "black and white" picture of ToM rather than the graded notion they prefer. Povinelli & Vonk (2004, 17-18) respond that their research program has always endorsed a graded notion of ToM, in the sense that different species might be able to represent different mental states at different levels of development, and that their ability to

do so might be composed of a variety of different skills or components. This response, however, conflates two notions of gradation, and thus two distinct kinds of semantic bias. The first notion of gradation holds that the ToM system might be decomposed into parts responsible for representing distinct mental states, parts that might emerge at different points in evolution and development; and the ability to robustly represent any mental state could be counted sufficient for ToM. Povinelli and his colleagues are rightly declared innocent of committing this form of semantic anthropocentrism, for they are clear that they would be willing to accept a robust ability to represent perceptual states alone as evidence for ToM. The second understanding of gradation, however, pertains to the varying degrees of reliability (e.g. number of different cues and their perceptual disparity) and varying numbers of contexts (e.g. cooperative vs. competitive) across which an *ms* candidate must covary with a target mental state to count as ToM. Povinelli and colleagues do regularly commit this latter form of semantic anthropocentrism, for while conceding that both humans and animals only approximate the necessary representational power to different degrees (Penn et al. 2008, 161), the degree of abstraction and domain generality they require for 'genuine ToM' has always been none-too-subtly set to the highest levels of human performance.

Indeed, a curious feature of Povinelli's position is his repeated concession that even humans rarely engage in the abstract, domain-general, relational form of mental state representation that he supposes essential for ToM. In most social interactions, such abilities would be redundant with information that can be obtained more directly from our embodied and enactive perceptual engagement with others. While insisting that the relevant ability to abstractly represent mental states is "manifestly obvious in human behavior", he has recently conceded that not even humans rely upon the relevant ability "most of the time in regard to our everyday interactions with others" (Gallagher & Povinelli, 2012, 151), that we routinely overestimate the degree to which we engage in such mental state representation (Penn & Povinelli, forthcoming), and even that adult humans approach social situations "in more chimp-like ways than young children" (Povinelli, 2011, 292). Nevertheless, Povinelli has continued to invoke this extraordinary degree of reliability as criterial of ToM, while admitting that ToM so precisified will not explain the forms of everyday human social cognition that we probably share with animals.

In other words, Povinelli has compounded his semantic anthropocentrism by tying it to an inflated account of human cognitive abilities. Such inflation is tempting, for our tendency to exaggerate our own intelligence, rationality, and reflective prowess is a feature of human psychology as well-established as our tendency to anthropomorphize. Psychologists have repeatedly demonstrated that we ascribe carefully reasoned justifications for

actions that were performed due to whims, heuristics, or situational factors; we confabulate memories of forgotten details and even of events that never happened; and we are routinely overconfident in our own abilities and disregard or misinterpret evidence to the contrary (Ariely, 2009, 2012; Bermúdez, 2003; Gilovich, Griffin, & Kahneman, 2002; Nisbett & Ross, 1980, 1991; Nisbett & Wilson, 1977; Tversky & Kahneman, 1974). While this literature has sometimes been over-interpreted, Malle, Knobe, and Nelson (Malle, Knobe, & Nelson, 2007; and see also Malle 2011) found in a careful meta-analysis of dozens of studies that people are more likely to explain their own behavior in terms of representational states like beliefs and desires and the behavior of others in terms of causal-historical factors such as cultural background, personality, and contextual cues. In short, insofar as the skeptical challenge sets up animals as the "others" to be evaluated against "our" human-level performance, it may be exploiting our tendency to overestimate our own cognitive sophistication while construing others as automata whose behaviors are determined by situational influences, making inflated approaches to vague psychological terms seem more plausible than they otherwise would.

Indeed, this attitude can be found lurking behind skeptical arguments against the possession by animals of other cognitive capacities. Skeptics of animal metacognition have endorsed a similarly complex notion of "strong" metacognition that could not be implemented by first-order mechanisms sensitive only to evidence like the relative strengths of beliefs (Carruthers, 2008), while conceding that most evidence for human metacognition can also be explained by such mechanisms and that people will tend to overinterpret their own performance (Carruthers, 2009, 130). Against animal episodic memory, skeptics have claimed that our purported ability to mentally replay conscious experiences of past events is an essential feature of episodic memory (Suddendorf & Corballis, 2008; Tulving, 1985), when human remembering is largely constructive and frequently confabulatory (Buckner & Carroll, 2007; Hippel & Trivers, 2011; Schacter, 1999). In each case, the skeptical position rests on a tenuous semantic premise that we must learn how to responsibly evaluate to determine whether the critique of a comparative claim is sound.[5]

[5] Note that we should distinguish skepticism about whether a given experiment is powerful enough to assess some criterion from skepticism about whether animals possess the psychological ability the criterion was elected to assess. For example, Carruthers may be right that existing animal metacognition experiments can be explained in terms of first-order mechanisms, but wrong that this prevents these experiments from providing evidence that animals possess "genuine" metacognition (because this assumption rests on an inflated criterion for metacognition).

These semantic claims are all in danger of committing an error that I call 'anthropofabulation'. Anthropofabulation arises from the combination of two biases: semantic anthropocentrism and exaggeration about typical human cognitive ability.[6] Together, they can lead one to implicitly add a "++" to some psychological term, claiming it obvious that only "++"-level ability is worthy of the name.[7] Given that some of these terms originated in human psychology, one might mount a principled defense of "single-+" semantic anthropocentrism about them—but such defenses are more problematic when combined with an exaggerated account of typical human performance. To be clear, the error does not arise result merely from drawing such distinctions—exceptional human performance, like exceptional performance in any animal, is interesting and should be studied—but rather in assuming that only the incremented levels of ability reflect a "genuine" form of the capacity in question, the only form relevant to comparative psychology, and especially the form that other researchers have asserted that animals possess.

[6] Though I will not explore the implications here, semantic anthropocentrism and confabulation about human performance could also be combined with anthropomorphism. This triumvirate of biases would result in the worst outcome yet: a comparative methodology which both presumed an inflated account of some psychological capacity *and* attributed that inflated capacity to animals on the basis of insufficient evidence. While I suspect this kind of triple error is the mistake that Penn & Povinelli take their critical targets to have committed, this case falls apart if proponents do not share in their semantic anthropocentrism.

[7] The sense in which anthropofabulation is implicit or automatic requires some elaboration. Often, both errors that comprise anthropofabulation will be tacit—anthropofabulists will implicitly confabulate about the complexity of their own performance *and* implicitly presume that only similarly complex performance is worthy of the name. Penn & Povinelli, however, repeatedly complain that our understanding of ToM has relied too much on what they call "our species' inveterate intuitions about how our own ToM works" (2007, 732), so they cannot be accused of committing the former error unawares. However, there is little evidence that they have critically evaluated the latter semantic assumption that "genuine" ToM requires the kind of domain-general competence presumed by this inflated folk psychology.

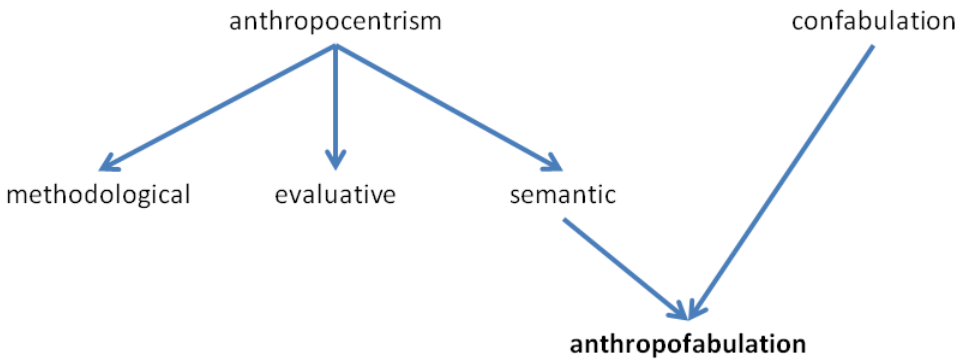

Figure 2. A taxonomy showing the descent of anthropofabulation, the tendency to set the criteria for psychological capacities to an artificially-inflated sense of what humans can or routinely do.

Given these concessions, together with the assumption that 'ToM' is vague and in need of clarification, it is difficult to see why the term must be explicated to exceptional, domain-general human performance. Against this inflation, it is clearly not the sense of 'ToM' intended by animal ToM proponents. Moreover, even if Penn et al. (2008) are right that this extraordinary level of ability enables many uniquely human achievements, such as our abilities to learn language, pass on a complex cultural heritage, and engage in political intrigue, it does not follow that this is the only empirically interesting place to draw the line. There are other coherent interpretations of 'ToM' that would emphasize different comparisons between human and animal performance and more charitably interpret proponents' claims (Buckner forthcoming), and which similarities or differences are theoretically important depends upon our explanatory purposes. While the skeptics' precisifications might better articulate the anthropological borderline, other scientists come at the issue with different interests. For example, we might want to study animal social cognition to get a clearer picture of the wide range of tacit abilities that do the lion's share of work in human social interactions; to map out the ecological pressures (such as large dominance hierarchies or domestication) that favor increased levels of social understanding; or to discover model organisms in order to understand and treat human pathologies. These interests—notably, the ones more commonly invoked by animal ToM proponents—are at least as legitimate as the goal of learning what makes humans unique, and each would speak in favor of less restrictive precisifications of 'ToM'.

To avoid privileging one of these theoretical interests at the expense of others without adequate justification, such semantic questions should not be addressed by appealing to intuitions about what is a "genuine" or "strong" form of a capacity, but rather by assessing considerations of comparative cognitive taxonomy. Typically, however, no systematic analysis of considerations like scope and power is offered by skeptics in defense

of their explications of key terms; instead, they quickly move on from the assertion that their interpretation of the term is the only faithful, valuable, or scientifically-rigorous choice. In the skeptics' defense, proponents are rarely explicit about their own semantic assumptions, making a comparison of competing interpretations difficult; however, while proponents can often be fairly accused of vagueness, this does not license skeptics to uncharitably place their own strong interpretations on proponents' claims when other options are available.

Broadly, different interpretations should be evaluated in terms of the trade-off between the **scope** and **power** of the competing taxonomic schemes they support—scope indicating the number of cases to which a precisified term applies (i.e. "how many things does it tell us about?"), and power indicating the number and explanatory power of the generalizations derived from the term's application to those cases (i.e. "how much does it tell us about them?").[8] Should an explication cleanly separate human and animal performance, this indeed increases the term's explanatory power; but should those generalizations apply only in rare cases even for humans, this increase in power is purchased only at the high price of greatly diminished scope. In cases where there are sensible alternative interpretations with more attractive scope/power tradeoffs, such exclusionary explications are at a clear disadvantage. Worse, such explications *ex hypothesi* tell us little about how animals (and humans, most of the time) actually do approach tasks, and so risk taking terms with productive research programs behind them out of the comparative discussion without replacing them by anything more precise.

Skeptics may protest at this point, contending that the worry that the experiments of proponents are not powerful enough in principle to distinguish genuine from "as-if" forms of an ability are precisely indictments of the explanatory power of the proponents' psychological taxonomies. Crucially, however, skeptics must first be granted their interpretations of psychological terms for these methodological critiques to succeed, so such interpretations of experiments cannot count as independent evidence in favor of their inflationary explications of key terms. For example, if the more limited degrees of reliability and numbers of contexts across which chimpanzees have demonstrated an ability to respond to the mental states of others are adequate to count as evidence for ToM on ecumenical construals such as Whiten's "intervening variable" approach (1996)—as proponents have repeatedly

[8] For a related discussion of comparisons between taxonomies, see Griffiths (1999, 216-219); and for an in depth discussion of explanatory power, see Ylikoski & Kuorikoski (2010).

contended (Tomasello & Call 2006)[9]—then these experiments instead enhance the scope and power of proponents' alternative taxonomic schemes. Moreover, anthropofabulous taxonomies certainly have more scope/power disadvantages than skeptics admit—for their anthropofabulous interpretation of 'ToM' has forced skeptics to shoehorn varied animal performance into poorly-defined categories such as "behavior-reading" that might be used post-hoc to explain nearly any experimental outcome (Fletcher & Carruthers 2012). In short, the central argument skeptics have offered against the explanatory power of proponent's taxonomic schemes only succeeds if they have already been granted their interpretations of key terms, and so cannot be counted as independent evidence in favor of those interpretations. Such arguments at best show that the skeptics' interpretations of proponents' experiments are consistent with their own skepticism.

In summary, the problem with anthropofabulation is not that it *always* favors false verdicts; like anthropomorphic bias, anthropofabulous reasoning can sometimes happen upon the right answer. The problem is that it biases us to accept inflationary answers to semantic questions on the basis of insufficient evidence. Unchecked, the bias can lead us to underestimate the psychological abilities of animals, further impoverish our taxonomies of the intermediate levels of cognition, and focus comparative discussion on rarified human abilities without adequate theoretical justification.

[9] Whiten's "intervening variable" notion does require the integration of information across some range of perceptually disparate situations to justify the appeal to a hidden variable. However, Call & Tomasello (2006) contend that existing experiments already demonstrate that chimpanzees respond appropriately to the mental states of others across a variety of perceptually disparate situations by integrating a complex combination of cues, including eye-directions, body orientations, whether the line of gaze does or does not terminate in a plausible target, presence or absence of occluders, and the type of occluders involved.

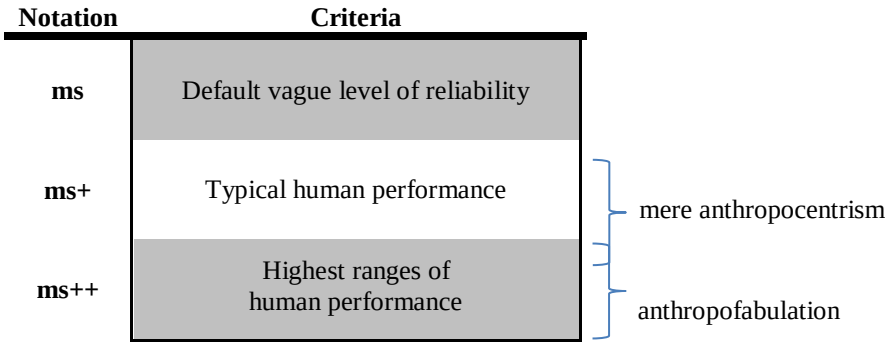

Table 2. Explanation of "+"modification of Penn & Povinelli (2007)'s quasi-formalism, in terms of degree of reliability required to demonstrate competence in representing mental states of others.

III. Hume's Dictum

If our tendency to anthropofabulate is automatic and powerful, how might we guard against this form of semantic bias? I suggest that just as Morgan's Canon has served as a corrective against anthropomorphic bias, a similar rule of thumb can help us limit anthropofabulation. Luckily, anthropofabulation (like many other sins) was not invented in the 20th Century, and so we do not have to start from scratch. In particular, Hume discerned the bias centuries ago in the doctrines of rationalists like Descartes (who attributed elaborate rational faculties to humans while infamously holding that animals are mere mechanical automata); and to guard against it, he recommended a corrective, which I will refer to as "Hume's Dictum" (1739/2000):

> When any hypothesis . . . is advanc'd to explain a mental operation, which is common to men and beasts, we must apply the same hypothesis to both; and as every true hypothesis will abide this trial, so I may venture to affirm, that no false one will ever be able to endure it. The common defect of those systems, which philosophers have employ'd to account for the actions of the mind, is, that they suppose such a subtility and refinement of thought, as not only exceeds the capacity of mere animals but even of children and the common people in our own species." (T1.3.16.3; SBN 177)

Let us proceed to unpack the Dictum carefully.

The first interpretive issue with the Dictum is that its application in the present context would appear to beg the question against skeptics. Precisely what is disputed is whether capacities like ToM are common to humans and animals, so it is inappropriate to begin from the assumption that they are shared. To apply the case to the present series of interpretive questions, we should here insert a "within bounds of rational possibility" before "common to men and beasts". In other words, suppose that the vague disputed terms have a consensus extension agreed upon by

all parties, and then a disputed penumbra. All precisifications of the term must satisfy what we might call a term's "consensus stereotype", or the set of descriptions associated with the term to which we would appeal to decide whether someone understood it. For example, perhaps it is uncontroversial that ToM is the capacity that:

1. Enables its possessor to predict and/or explain the behavior of others,
2. Humans distinctively excel at,
3. Is typically impaired in autistic individuals, and
4. Exhibits these other properties by representing another's mental states.

This consensus stereotype sets the bounds for rational debate (disagreements arise only later, when researchers attempt to precisify a vague stereotype). If one recommends a precisification of a vague concept that grossly violates its consensus stereotype, it is like calling (a non-shorn) Cher 'bald'; we must conclude either that the speaker means the word in a non-literal sense or does not understand its meaning.[10] Thus, the Dictum is relevant to terms for which some precisifications within the bounds of its stereotype would plausibly apply to the abilities of both humans and animals. This interpretive point will be of little practical consequence, however, for as we shall see applying the Dictum even more liberally would not prevent us from favoring criteria that humans would typically satisfy but animals would fail.

Secondly, the Dictum holds that when assessing whether some psychological capacity is shared between humans and animals, we should adopt competence criteria that can be fairly applied to both. "Competence criteria" here indicates not merely operational criteria chosen to assess the presence of a capacity in some particular experiment, but rather the general abilities or dispositions (derived by precisifying a term's stereotype) that a subject should by definition possess if it is endowed with the psychological capacity in question. There is, of course, a close relationship between a capacity's general competence criteria and the specific operational criteria selected to assess its presence in any particular experiment. For example, the skeptics' competence criteria for perceptual ToM might be "a domain-general ability to represent and respond appropriately to the perceptual states of others", and an operational criterion derived from it for a cooperative food-begging experiment might be "begs for food from the collaborator that can see." Moreover, a competence criterion will typically only be exposed as problematic when it in practice limits experimentalists to operational criteria that only exceptional human performance can satisfy.

[10] This is not to say that consensus stereotypes cannot change over time. For example, if we learned that the association between autism and ToM-deficits were a statistical artifact, or that 'autism' conflates distinct disorders with different etiologies, we might revise our stereotype to modify or exclude property 3.

Nevertheless, as understood here the pure form of anthropofabulation arises at the definitional stage, by precisifying a term's meaning by appealing to inflationary competence criteria, rather than by deriving an inflationary operational criterion from an adequate definition.[11]

The stipulation that we adopt only competence criteria that fairly apply to both humans and animals is also non-controversial, but carries some subtle consequences. For example, this proscription implies that we should do our best to avoid criteria that could not be deployed on animals, such as verbal descriptions and probes, preferring those that can, such as non-verbal forms of response assessment (e.g. comparisons of looking-times).[12] It also requires that we do our best to provide animals with learning histories and cultural scaffolding comparable to those enjoyed by the human subjects purported to satisfy the criteria. At the very least, it demands that we skeptically regard human subjects' verbal reports as behavioral data that may be misleading, rather than as necessarily reliable reports of subjects' cognitive activity. Anthropofabulation renders such challenges more difficult to overcome, for it directs our attention towards increasingly complex, artificial, and anthropocentric tasks as the only ones powerful enough to distinguish "genuine" from "as-if" forms of ability.

To consider some examples, Boesch (2007) comprehensively reviews the comparative literature on ToM, pointing out how nearly every experiment violates these ideals of fairness by pitting captive chimpanzees against free-ranging humans, humans working with conspecifics against chimpanzees working with heterospecifics, humans with parents nearby against apes without parents nearby, or humans on familiar materials against apes on unfamiliar materials. Anthropofabulation unduly minimizes the importance of these disanalogies, for the presumption that humans routinely deploy the idealized, domain-general forms of ToM can lead us to assume that the success of the human subjects in these experiments does not depend upon this additional developmental and environmental scaffolding. To consider another example, episodic memory researchers have, following Tulving (1985), long relied on asking subjects whether they really "remember" or merely "know" some fact to decide whether the autonoietic consciousness of previously experienced events supposed essential for episodic memory has been activated in an experiment. While there is no doubt some mechanism producing subjects' responses to this probe, we should regard

[11] While the former, definitional form of anthropofabulation is more methodologically troublesome (because it semantically institutionalizes bias), the latter is also a serious and closely-related form of error that can similarly be avoided by attending to Hume's Dictum.

[12] Unless, of course, verbal acuity is an essential component of the capacity being assessed.

with skepticism the tacit assumption that their reports necessarily derive from a conscious replay of past experience. Indeed, it has been shown repeatedly that confabulators report the subjective sense of remembering events that never happened (Schnider, 2008), and people generally overestimate the reliability of human memory, creating notorious problems with eyewitness testimony in court cases (Simons & Chabris, 2011).

For both capacities, emphasis on anthropofabulous criteria have resulted in biased comparisons, whereas fairer comparisons would more clearly expose the limited scope of such inflationary criteria in even human psychology. In short, the Dictum's second component can be read as an "equal skepticism" clause; whatever level of skepticism we apply in selecting and applying the competence criteria for evidence of animal abilities, so should we deploy in the case of humans—while noting that we are especially prone to exaggerate our own prowess and underplay the importance of our own background, situation, and culture.

The final component of the Dictum is the demand that we set competence criteria for vaguely-defined capacities not to the highest ranks of human performance, but rather only to the typical performance of children and the folk. The Dictum acquires some significant bite in current debates with this restriction, for by requiring that competence criteria be satisfiable by children and typical adult performance, it makes it more difficult to add a "++" (or above) to every psychological capacity.

The stickiest question in applying the Dictum, however, regards this mention of children. For example, ToM in humans is composed of a variety of skills and levels of ability that emerge at different developmental stages. Should we set the age limit for children to six months of age, we have virtually guaranteed that ToM will be common to both humans and animals, since six-month-olds are unlikely to have developed much reliability in representing any sorts of mental states. Similarly, if we set the bar to twenty years of age, we will guarantee that ToM will be uniquely human, as average humans at this point will have achieved a level of reliability that no animal could ever reach. The same goes for the other capacities; for example, stability in the "remembers" vs. "knows" probe often used in episodic memory research does not emerge until late adolescence (Piolino et al., 2007). Importantly, such age ranges are likely variable even for humans; Boesch (2007, 231) argues that the age at which explicit false-belief understanding (often regarded as a criterial milestone in the ontogeny of human ToM) emerges in humans varies widely and is heavily influenced by developmental and cultural factors. Likely there can be no general answer to the "how old" question here; age-ranges must rather be assessed on a case-by-case basis, depending upon the capacity and comparative claims in question. As a heuristic, we might limit our attention to age

ranges before which humans of the relevant comparison culture could be expected to have enjoyed a relevant learning history and cultural context difficult or impossible to reproduce in animal subjects. This heuristic is justified by the idea that the later we look in human development, the more likely it becomes that superior performance is due to a more extensive learning history and/or cultural scaffolding rather than to some uniquely human cognitive mechanism (Barrett, 2008; McGonigle & Chalmers, 2008). While modest, this heuristic provides useful guidance in evaluating comparative claims. For example, while Povinelli & Eddy (1996) famously reported that their chimpanzees failed a range of ToM tasks, chimpanzees tested by Bulloch et al. (2008) are reported as having passed some of the same tasks—the latter speculating that their chimps were more successful due to more extensive (and thus more human-like) experience with human collaborators.

I believe that skeptics would generally endorse these components of Hume's Dictum, albeit while attempting to push their limits. As noted above, setting the bar at *ms*+ is indeed semantic anthropocentrism; but whether "single-+"anthropocentrism is an error is at least partly an empirical question. However, the Dictum does make it much more difficult to interpret vague psychological terms in an anthropofabulous ('++') way by requiring that precisifications possess at least a modicum of scope, covering at least typical, fairly-assessed human performance.

Skeptics may worry at this point that Hume's Dictum has become too strong, in that would prevent us in principle from discussing exceptional human abilities in a comparative context. It bears emphasizing that Hume's Dictum as interpreted here does no such thing, for it applies only to vague terms whose consensus stereotype covers at least some animal performance. It does not prevent us from coining new terms or developing conventions for modifying old ones to clarify features supposedly manifest only in exceptional human performance. Moreover, if anthropofabulation commonly derives from overestimation of the degree of reliability and domain-generality of human representational prowess, then one can explicitly characterize the differences between typical and exceptional human ability as differences of degree in some more vaguely-characterized type of capacity—as I have done with the "++" modification of Povinelli's '*ms'*-formalism above. Introducing such compositional modifiers allows the accommodation of a more coarse-grained comparative taxonomy to the finer-grained differences in human and animal performance, in the way that the periodic table of elements can be accommodated to diversity in

the nuclear configurations of different chemical compounds, ions, and isotopes without having to decide which forms of a substance are "genuine" and which merely "as-if".[13]

A final question about Hume's Dictum is whether, and in what ways, it can be combined with Morgan's Canon. It might initially seem that the two principles are in direct conflict; for Morgan's Canon, at least on a strong reading, recommends the conclusion that animals lack "higher" processes if experiments fail to establish them, whereas Hume's Dictum should lead us to worry that experiments only failed to establish higher processes because those experiments had been designed to assess inflated criteria. However, on a more moderate reading of the Canon (cf. footnote 1), the two principles are actually complementary, for each is designed to counter a distinct bias that arises at a distinct phase of comparative research.[14] Anthropofabulation biases our interpretation of vague psychological terms, leading us to design experiments to assess artificially-inflated competence criteria; anthropomorphism biases our interpretation of experimental results, leading us to accept that some data satisfy some competence criteria on the basis of insufficient evidence. Moving beyond oversimplified "pro-" or "anti-"animal thinking, a healthy comparative project must avoid both biases—by electing fair criteria for experiments to assess, and then relying on objective, empirical assessment of whether those criteria have been satisfied.

IV. Conclusion

> *In a Darwinian framework, there is no good reason to avoid concepts merely because they derive from the behaviors of the species to which we belong. Application of these concepts to animals not only enriches the range of hypotheses to be considered, but it also changes the view of ourselves: the more human-like we permit animals to become, the more animal-like we become in the process.* (De Waal, 2000, 272, quoted in Keeley, 2004)

---

[13] For a brief discussion on how compositional operators can support the accommodation of abstract taxonomies to underlying diversity, see Boyd (1999, 157-158). Note that differences in degree captured by such modifiers may or may not indicate qualitative differences; for example, the two architecturally distinct systems for human ToM postulated by Apperly & Butterfill (2009) might be mapped to *ms*+ and *ms*++, respectively.

[14] Indeed, a reviewer points out that Morgan himself would likely have been sympathetic to Hume's position here. Morgan was well-aware that humans often overestimate their own prowess and that this can interfere with cross-species comparisons—warning that "to interpret animal behavior one must learn also to see one's own mentality at levels of development much lower than one's top-level of reflective self-consciousness. It is not easy, and savors somewhat of paradox" (Morgan 1930, 250).

Adhering to Hume's Dictum improves not only our understanding of animal minds, but of human minds as well. Povinelli implores us to avoid methodological and evaluative anthropocentrism so that we may study chimpanzee psychology for its own sake, without seeing them as "smaller, duller, less talkative versions" (2004, 29) of humans. Similarly eschewing anthropofabulation, however, can help us take a long, hard look in the mirror and realize that we are also typically smaller, duller, and less talkative than we tend to suppose. This is not to denigrate the import or interestingness of the resultant human or comparative psychology; on the contrary, it is notable that it took researchers only months after the first digital computer was built to create programs that could manipulate higher-order relations on predigested formal problems well beyond the ability of any human, whereas we still struggle to build machines that can solve problems of social coordination or learn novel causal relationships as reliably and flexibly as humans and animals in real-time perceptual circumstances. While the former are indeed distinctively human cognitive achievements, they play a much smaller role in our everyday lives than the other capacities—the tacit, subpersonal, interactive, and heuristic—that still await precise characterization and machine reproduction. Much of the challenge and wonder in studying nonhuman animals is that doing so can help us develop a conceptual taxonomy adequate to describe the vast underground foundation of shared abilities supporting the "heights" of human achievement. In short, avoiding anthropofabulation not only helps us better understand the minds of animals; it can also better acquaint us with our own.

Acknowledgments: I am grateful to Colin Allen, Louise Barrett, Melinda Fagan, Anika Fiebich, Daniel Povinelli, and two anonymous reviewers for helpful comments and discussion. The essay also benefitted from discussions on an early version presented at the 2011 Winter Conference on Animal Learning & Behavior. Finally, I thank the Alexander von Humboldt Stiftung for the postdoctoral fellowship that partially supported this research.

References

Andrews, K. (2012). *Do apes read minds? Toward a new folk psychology*. Cambridge: MIT Press.

Andrews, K. (2005). Chimpanzee theory of mind: Looking in all the wrong places? *Mind & Language*, *20*(5), 521–536.

Apperly, I. & Buterfill, S. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review* 116, 953-970.

Ariely, D. (2009). *Predictably irrational: The hidden forces that shape our decisions*. New York: Harper Collins.

Ariely, D. (2012). *The (honest) truth about dishonesty: How we lie to everyone–especially ourselves.* New York: Harper Collins.

Barrett, L. (2008). Out of their heads: Turning relational reinterpretation inside out. *Behavioral and Brain Sciences*, *31*(2), 130–131.

Bermúdez, J. L. (2003). The domain of folk psychology. In A. O'Hear (Ed.), *Minds and Persons* (pp. 1–29). Cambridge: Cambridge University Press.

Boesch, C. (2007). What makes us human (Homo sapiens)? The challenge of cognitive cross-species comparison. *Journal of Comparative Psychology*, *121*(3), 227–240.

Boyd, R. (1999). Homeostasis, species, and higher taxa. In R. Wilson (Ed.), *Species: New interdisciplinary essays* (pp. 141–185). Cambridge: Cambridge University Press.

Buckner, R., & Carroll, D. (2007). Self-projection and the brain. *Trends in cognitive sciences*, *11*(2), 49–57.

Buckner, C. (Forthcoming). The semantic problem(s) with research on animal mindreading. *Mind & Language*.

Bugnyar, T. (2007). An integrative approach to the study of "theory-of-mind"-like abilities in ravens. *Japanese Journal of Animal Psychology*, *57*(1), 15–27.

Bulloch, M., Boysen, S., & Furlong, E. (2008). Visual attention and its relation to knowledge states in chimpanzees, Pan troglodytes. *Animal Behaviour*, *76*(4), 1147–1155.

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences*, *12*(5), 187–92. doi:10.1016/j.tics.2008.02.010

Carruthers, P. (2008). Meta-cognition in animals: A skeptical look. *Mind & Language*, *23*(1), 58–89.

Carruthers, Peter. (2009). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and brain sciences*, *32*(2), 121–138.

De Waal, F. (2000). Anthropomorphism and anthropodenial: consistency in our thinking about humans and other animals. *Philosophical Topics*, *27*, 255–280.

Dretske, F. (1986). Misrepresentation. In R. Bogan (Ed.), *Belief: Form, content and function* (pp. 17–36). New York: Oxford.

Dretske, Fred. (1988). *Explaining behavior: Reasons in a world of causes.* Cambridge: The MIT Press.

Emery, N. J., & Clayton, N. S. (2009). Comparative social cognition. *Annual Review of Psychology*, *60*(1), 87–113.

Fitzpatrick, S. (2008). Doing away with Morgan's Canon. *Mind & Language*, *23*(2), 224–246.

Fitzpatrick, S. (2009). The primate mindreading controversy: a case study in simplicity and methodology in animal psychology. In Robert Lurz (Ed.), *The philosophy of animal minds* (pp. 258–277). Cambridge: Cambridge University Press.

Fletcher, L., & Carruthers, P. (2012). Behavior--reading versus mentalizing in animals. In J. Metcalfe & H. Terrace (Eds.), *Agency and Joint Attention.* Oxford: Oxford University Press.

Flombaum, J. I., & Santos, L. R. (2005). Rhesus monkeys attribute perceptions to others. *Current Biology*, *15*(5), 447–452.

Gallagher, S., & Povinelli, D. (2012). Enactive and behavioral abstraction accounts of social understanding in chimpanzees, infants, and adults. *Review of Philosophy and Psychology*, *3*, 145–169.

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and biases: The psychology of intuitive judgement*. Cambridge: Cambridge University Press.

Griffiths, P. (1999). Squaring the circle: natural kinds with historical essences. In R. Wilson (Ed.), *Species: New interdisciplinary essays* (pp. 208–228). Cambridge: MIT Press.

Heyes, C. (1998). Theory of mind in nonhuman primates. *Behavioral and Brain Sciences*, *21*, 101–148.

Hippel, W. Von, & Trivers, R. (2011). The evolution and psychology of self-deception. *Behavioral and Brain Sciences*, *34*(1), 1–56.

Hume, D. (1739). *A treatise on human nature*. Oxford University Press.

Keeley, B. (2004). Anthropomorphism, primatomorphism, mammalomorphism: understanding cross-species comparisons. *Biology and Philosophy*, *19*, 521–540.

Lurz, Robert. (2009). If chimpanzees are mindreaders, could behavioral science tell? Toward a solution of the logical problem. *Philosophical Psychology*, *22*(3), 305–328. doi:10.1080/09515080902970673

Lurz, RW. (2011). *Mindreading animals: The debate over what animals know about other minds*. Cambridge: The MIT Press.

Macmillan, N. (2002). Signal detection theory. In S. Yantis (Ed.), *Stevens' handbook of experimental psychology*. Mankato: Coughlan Publishing.

Malle, B. (2011). Time to give up the dogmas of attribution: An alternative theory of behavior explanation. In M. Zanna (Ed.), *Advances in experimental social psychology*. Salt Lake City: Academic Press.

Malle, B., Knobe, J., & Nelson, S. (2007). Actor-observer asymmetries in explanations of behavior: New answers to an old question. *Journal of Personality and Social Psychology*, *93*(4), 491–514.

McGonigle, B., & Chalmers, M. (2008). Putting Descartes before the horse (again!). *Behavioral and Brain Sciences*, *32*(2), 142–143.

Morgan, C. L. (1903). *An introduction to comparative psychology* (p. 386). W. Scott.

Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs: Prentice-Hall.

Nisbett, R., & Ross, L. (1991). *The person and the situation*. *NY: McGraw Hill*. New York: McGraw-Hill.

Nisbett, R., & Wilson, T. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–257.

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: explaining the discontinuity between human and nonhuman minds. (B. Smith & D. Woodruff Smith, Eds.)*Behavioral and Brain Sciences*, *31*(2), 109–130; discussion 130–178.

Penn, D. C., & Povinelli, D. J. (2007). On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind. *Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences*, *362*(1480), 731–744.

Penn, D., & Povinelli, D. (forthcoming). The comparative delusion: the behaviorstic/mentalistic dichotomy in comparative theory of mind research. *Agency and joint attention.* H. A. Terrace and J. Metcalfe, eds. Oxford University Press.

Piolino, P., Hisland, M., Ruffeveille, I., Matuszewski, V., Jambaque, I., & Eustache, F. (2007). Do school-age children remember or know the personal past? *Consciousness and Cognition*, *16*(2007), 84–101.

Povinelli, D. (2011). *World without weight: Perspectives on an alien mind*. Oxford: Oxford University Press.

Povinelli, Daniel, & Eddy, T. (1996). What young chimpanzees know about seeing. *Monographs of the Society for Research in Child Development*, *61*(3), 1–189.

Povinelli, Daniel, & Vonk, J. (2003). Chimpanzee minds: suspiciously human? *Trends in cognitive sciences*, *7*(4), 157–160.

Povinelli, DJ. (2004). Behind the ape's appearance: Escaping anthropocentrism in the study of other minds. *Daedalus*, *133*(1), 29–41.

Povinelli, DJ, & Vonk, J. (2004). We don't need a microscope to explore the chimpanzee's mind. *Mind & Language*, *19*(1), 1–28.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(04), 515–526. doi:10.1017/S0140525X00076512

Radick, G. (2000). Morgan's canon, Garner's phonograph, and the evolutionary origins of language and reason. *The British Journal for the History of Science*, *33*(1), 3–23.

Richards, R. (1989). *Darwin and the emergence of evolutionary theories of mind and behavior*. Chicago: University of Chicago Press.

Santos, L., Flombaum, J., & Phillips, W. (2007). The evolution of human mindreading: How non-human primates can inform social cognitive neuroscience. In S. Platek, J. Keenan, & T. Shackleford (Eds.), *Evolutionary cognitive neuroscience* (pp. 433–456). Cambridge: MIT Press.

Schacter, D. (1999). The seven sins of memory: Insights from psychology and cognitive neuroscience. *American Psychologist*, *54*(3), 182–203.

Schnider, A. (2008). *The confabulating mind: How the brain creates reality*. Oxford: Oxford University Press.

Shettleworth, S. J. (2009). The evolution of comparative cognition: is the snark still a boojum? *Behavioural Processes*, *80*(3), 210–217.

Simons, D., & Chabris, C. (2011). What people believe about how memory works: A representative survey of the US population. *PLoS One*, *6*(8), e22757.

Sober, E. (1998). Morgan's canon. In D. Cummins & C. Allen (Eds.), *The Evolution of Mind* (pp. 224–242). New York: Oxford University Press.

Sober, E. (2005). Comparative psychology meets evolutionary biology: Morgan's canon and cladistic parsimony. In L. Dalston & G. Mitman (Eds.), *Thinking with animals: New perspectives on anthropomorphism*. New York: Columbia University Press.

Suddendorf, T., & Corballis, M. (2008). New evidence for animal foresight? *Animal Behaviour*, *75*(5), e1–e3. doi:10.1016/j.anbehav.2008.01.006

Thomas, R. (2006). Lloyd Morgan's Canon: A history of misrepresentation. *History & Theory of Psychology Eprint Archive*.

Tomasello, M., Call, J., & Hare, B. (2003). Chimpanzees versus humans: it's not that simple. *Trends in Cognitive Sciences*, *7*(8), 239–240.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology*, *26*(1), 1–12.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Wynne, C. (2007). What are animals? Why anthropomorphism is still not a scientific approach to behavior. *Comparative Cognition & Behavior Reviews*, *2*, 125–135.

Ylikoski, P., & Kuorikoski, J. (2010). Dissecting explanatory power. *Philosophical Studies*, *148*(2), 201–219.