What Makes an Explanation Believable?: Mechanistic and Anthropomorphic Explanations of Natural Phenomena

Jordan Schoenherr (psychophysics.lab@gmail.com)
Department of Psychology, Carleton University
1125 Colonel By Drive, Ottawa, ON K1S5B6 Canada

Robert Thomson (rthomson@connect.carleton.ca), Jim Davies (jim@jimdavies.org)
Institute of Cognitive Science, Carleton University
1125 Colonel By Drive, Ottawa, ON K1S5B6 Canada

Abstract

Many biases in decision-making and reasoning are a result of ignoring logical rules and relevant information while focusing on irrelevant cues present within an argument. In the present study we examine explanatory schemata – a set of interrelated concepts - that are deemed relevant to participants. Participants were first trained in a syllogistic reasoning task and were then presented descriptions of natural phenomena and explanations. An instructional manipulation varied the source of the explanations (scientists or people) as well as the animacy of the natural phenomena (living or nonliving). Explanations used either mechanistic (e.g., force) or anthropomorphic (e.g., wants) terms. We found that participants were more accurate when assessing mechanistic explanations.

Keywords: anthropomorphism; heuristics; syllogistic reasoning; belief bias

Introduction

Each day, the media presents information to the public from purportedly credible sources. People must then generate beliefs based on these explanations. This is especially true of scientific discoveries. Weisberg, Keil, Goodstein, Rawson and Gray (2008) found that explanations from the psychological sciences were seen as more satisfying when accompanied by irrelevant neuroscientific information. This result is also supported by prior research, which has identified that the kind of explanation (Cosmides & Tooby, 1992) and prior knowledge (Kahneman & Tversky, 1982) affects the accuracy of people's judgments. In the present study we examined the kinds of prior knowledge that can lead to inaccurate assessments of explanations of natural phenomena by manipulating the source of the explanation, the properties of the natural phenomena and the mode of explanation. We additionally used subjective confidence reports to determine whether participants were aware of the factors influencing their performance.

Biases from Domain-Specific Knowledge

Subjective biases are generally attributed to a variety of decision-making rules and heuristics (for a review see Kahneman, 2003; Kahneman & Tversky 1982). Heuristic-related biases are also observed in the context of rule-based syllogistic reasoning tasks (e.g., Evans, Barston & Pollard, 1983; Sa et al., 1999). In a typical syllogistic reasoning task, participants are given premises and are required to indicate whether a conclusion logically follows from them. To examine the effects of prior beliefs on successful task completion, Evans et al. (1983) varied both the validity of the argument and believability of the conclusion. Validity follows from rules of formal logic whereas believability stems from how closely the conclusion conforms to one's prior beliefs. When the argument is invalid but the conclusion is believable, participants should disregard their prior belief and focus on the invalidity of the argument. A belief bias is observed when participants accept these invalid yet believable arguments.

Knowledge Effects in Decision-Making

Failures of decision-making have also been observed in the Wason Selection Task (Wason, 1966). In this task, participants are presented with four cards in order to identify whether a rule is false (e.g., If a card has a vowel on the obverse it will have an even number on the reverse). To successfully complete the task, participants should select a card that would disconfirm the rule (an odd number) and one that confirms the statement (a vowel). Over a wide range of subject categories, average responses rarely exceed 25% accuracy (Cosmides & Tooby, 1992; Stanovich, 2004).

Performance in the Wason Selection Task can be improved when domain-specific knowledge facilitates the selection of an accurate response (Cosmides & Tooby, 1992; Gigerenzer & Selten, 2000). To account for this evidence, Cosmides (1985) speculated that failures in reasoning tasks could be attributable to a mismatch between the domains considered in the task and domain-specific cognitive modules created through natural selection. If these tasks reflected verification of violations of social contracts – served by an innate cognitive module in her account – then participants' performance should improve (Cosmides 1989; Cosmides & Tooby, 1992). For instance, participants could be presented with the task of verifying that customers of a pub are of an appropriate age for entry. Studies that have employed such methodologies have observed performance at or exceeding 75% accuracy (e.g., Gigerenzer & Hug, 1992; Griggs & Cox, 1982). Thus, if prior knowledge is available, the extent to which it overlaps with task demands should determine performance (cf. Liberman & Klar, 1996). Given that people possess both naïve psychological and physical theories about the world – whether learned or innate – it is surprising that prior research has not examined whether these naïve theories could be used to draw analogies with other domains, facilitate the comprehension of logical arguments as well as determining the extent to which some naïve theories serve this function better than others.

Mechanistic and Intentional Reasoning Strategies

One possibility proposed by Dennett (1987) was that individuals could draw analogies from naïve psychological theories concerning intentionality to facilitate comprehension of natural phenomena. As originally conceived by Dennett (1987), an *intentional stance* is a generative explanatory theory that individuals use to impute intentionality to objects and entities. Dennett's basic proposal requires that we regard an entity or object as a rational goal-directed agent. This approach, according to Dennett, reduces the burden of constructing more complex theories based on physical forces (a *physical stance*) or function (a *design stance*). Consequently, an anthropomorphic analogy could be used to more readily encode and decode the relations of the parts within a system. Although there is considerable evidence for the early development of anthropomorphic reasoning heuristics (Wellman, Cross & Watson, 2001), it is unclear whether there is a comparative advantage that persists into adulthood (cf., Miller & Aloise, 1989).

Adults, however, also possess heuristics based on naïve physical theories about the world (e.g., McCloskey, 1983). Moreover, in our society people are taught to conceive of the world in terms of cause and effect with objects interacting with one another via abstract forces (e.g., Nisbett, 2003). It is possible that even if humans are predisposed to apply an intentional stance, explanations that draw on mechanistic explanations might be perceived as more familiar, authoritative, and as a result more likely to be correct. In a recent study conducted by Weisberg et al. (2008), both novices and experts were given explanations that were either 'good' or 'bad'. The explanations also differed in terms of whether they included irrelevant neuroscientific evidence or contained no evidence. Their results indicated that, in the novice groups, participants were more likely to accept 'good' or 'bad' explanations when neuroscientific evidence was provided. Thus, irrelevant evidence is used as a cue in determining the quality of an argument.

Present Research: Anthropomorphic Reasoning

The results of Weisberg et al.'s (2008) study provide a reasonable extension of previous research on reasoning abilities. Their results, however, leave several open questions. For instance, beliefs rarely consist of a single proposition. Instead, most beliefs comprise a complicated set of interrelated concepts and propositions. Thus, Weisberg et al.'s results might have been caused by several possible properties of the stimuli.

One possibility is that individuals are rarely exposed to neuroscientific evidence and as a result, may be unable to adequately judge its relevance. Specifically, participants might place a premium on explanations at the neuron-level because of a naïve theory that they have a strong causal role in human behaviour. For this reason, participants' beliefs about neurons may be influenced by the fact that most entities with neurons have some form of intentionality thereby supporting psychological explanations. Alternatively, naïve theories about neurons might contain a belief that neurons are subject to chemical and physical forces that are best explained at a mechanistic level of which they are not aware. By examining the broader domain of 'natural phenomena' we can examine whether anthropomorphic or mechanistic explanations are taken as more believable.

A second consideration is whether the source of the explanation (e.g., scientists) also influenced participants' decisions. Namely, one cannot have neuroscientific evidence without scientists but one can be provided with a mechanistic explanation by a layperson. It might be the case that the *source* of the information is the principle element of the explanation and not the evidence *per se*. Thus, by controlling for the source of the explanation (e.g., 'scientists' or 'people'), the saliency of this bias can be manipulated.

Finally, natural phenomena exist on a continuum of animacy. When invoking neuroscientific evidence, one necessarily implies that the phenomena under consideration are alive in some sense. Thus, participants might be primed to consider an explanation in a qualitatively different manner. It might be the case that there is congruence between the source of the explanations and the type of explanation given. For instance, although both ants and molecules move, they do so for very different reasons (i.e., one is alive (animate) whereas the other is only subject to physical forces). Thus, by varying the animacy of the natural phenomena under consideration, we can control for congruence between explanation type (i.e., mechanistic or anthropomorphic) and the source of the explanation (i.e., people or scientists).

It is also unclear whether participants are explicitly aware of the reasoning strategy they have adopted when evaluating the validity of arguments. Elsewhere, subjective measures of awareness such as confidence reports have been used to differentiate between sources of implicit and explicit knowledge (e.g., Dienes & Berry, 1997). In the present study, participants were required to report confidence in their responses in order to determine whether they were using an explicit reasoning strategy or whether biases were the results of an implicit reasoning strategy. Confidents reports were compared to mean proportion correct to obtain a measure of participants' awareness of their performance (e.g., Baranski & Petrusic, 1994; Keren, 1991). Significant deviation between participants' perception of their performance and their obtained performance would suggest that they were unaware of the rules they used to assess the validity of the explanations.

Experiment

Although the results of Weisberg et al. (2008) complement many findings in the belief-bias literature, they failed to control for several factors. To examine whether these factors affect reasoning more generally and whether adoption of a heuristic akin to the intentional stance (Dennett, 1987) can aid in the assessment of an argument's validity, we manipulated the source of the explanation, the animacy of the phenomenon, and the type of explanation while also controlling for the source of the explanation.

Four types of syllogisms were used; half were valid and the other half were invalid. The source of the explanation ('scientists' or 'people') and the type of explanation (mechanistic or anthropomorphic) were varied in a pure-block design with the order of block presentation counterbalanced. The animacy of the phenomena ('living' or 'nonliving') was included as a between-subjects variable.

Method

Participants

Sixty undergraduates participated in the experiment receiving 1% toward their final grade in an introductory psychology class. Two participants were excluded due to an experimental error.

Stimuli

Four types of logical syllogisms were used as the basis of the stimuli: Modus Ponens (MP: If P then Q; P; therefore Q), Modus Tollens (MT: If P then Q; not Q; therefore not P), hypothetical (HS: If P then Q; If Q then R; therefore, if P then R) and disjunctive (DS: Either P or Q; Not P; therefore, Q). To avoid associations with any prior knowledge for specific entities, entity names used in the syllogisms consisted of four-letter pronounceable non-words (e.g., Lozu, Baje, Yulo).

In the training set, participants were presented with a standard syllogism that included non-words in the premises and conclusions. In the experimental set, syllogisms were modified such that there was a description and an explanation. Descriptions of phenomena contained the first premise in the syllogism (e.g., If P then Q). Explanations contained second premise and the conclusion (e.g., P; therefore Q) as well as an irrelevant explanatory element that was either anthropomorphic (e.g., P likes Q) or mechanistic (P is drawn to Q by a force). The explanatory element was positioned between the second premise and the conclusion. Explanation validity was also varied.

Procedure

Participants were first presented with a short training phase of 16 trials consisting of each type of syllogism (modus ponens, modus tollens, hypothetical and disjunctive) and validity conditions (valid and invalid). Instructions were presented prior to each experimental block. In the training phase, participants were merely instructed that they would be presented with logical syllogisms and were required to indicate the validity of the statement, that is, whether the explanation followed logically from the description. In the experimental phase, participants were presented with 32 trials. Sixteen trials consisted of a syllogism from each condition (valid or invalid; anthropomorphic or mechanistic; MP, MT, HS or DS). Another sixteen trials were presented with the same syllogisms but changing the nonsense words so that there would be no bias from any associations from previous trials.

Table 1.
Samples of Syllogisms Modified with Mechanistic and Anthropomorphic Explanations. Major and minor premises are denoted by P1 and P2, respectively, and the conclusion is denoted by C. The irrelevant explanation is denoted by E.

Mech. Valid	**Description**: If [a Baje moves toward a Yulo]^{P1} then [they will stick together]^{P2} **Explanations**: [A Baje moves toward a Yulo] because [Bajes and Yulos are bound by a force]^{E} that [attracts them]^{C}.
Anthro. Valid	**Description**: If [a Lozu moves toward a Hexi] ^{P1} then [they will stick together] **Explanations**: [A Lozu moves toward a Hexi] because [Lozus and Hexis like one another]^{E} so they [are drawn together] ^{C}.
Mech. Invalid	**Description**: If [a Dafe moves toward a Noha] ^{P1} then [they will stick together] ^{P2} **Explanations**: [A Dafe moves toward a Noha] because [Dafes and Nohas are bound by a force]^{E} that [repels them] ^{C}.
Anthro. Invalid	**Description**: If [a Vipo moves toward a Pova] ^{P1} then [they will stick together] ^{P2} **Explanations**: [A Vipo moves toward a Pova]^{P1} because [Vipos and Povas dislike one another]^{E} so they [are driven apart]^{C}.

Participants completed two blocks of trials. In one block, they were told that the observations and explanations were created by scientists. In the other block, they were told that people created the observations and explanations. The same description-explanations sets were used in both blocks of trials. Block order was counterbalanced between participants.

Finally, half the participants were informed that the syllogisms involved living phenomena whereas the other half were informed that the phenomena were nonliving. After participants indicated the validity in the statement they were required to rate their subjective confidence they had provided the correct answer using values 6-point scale with values 50% (guessing) through 100% (certain).

Results

A 2 (validity: valid, invalid) x 2 (explanation type: anthropomorphic, mechanistic) x 2 (explanation source: person, scientist) x 2 (animacy: animate, inanimate) mixed repeated measures ANOVA was conducted. Given that training stimuli were used to familiarize participants with logical syllogisms and that their order was fixed, these stimuli were not included in the analysis, however their means are presented in Table 2 for comparison purposes.

Accuracy

Replicating the belief-bias effect, the interaction between validity and explanation type was significant, *F*(1, 58) = 95.607, *MSE* = .037, *p* < .001, η_p^2 = .622. The main effect of explanation type was also significant, *F*(1, 58) = 11.872, *MSE* = .037, *p* < .005, η_p^2 = .170. Importantly, the same general pattern was observed for both mechanistic explanations and the training syllogisms for both valid and invalid explanations. Given that both the training syllogisms and mechanistic statements show comparable patterns, this suggests that the reasoning process proceeded in a similar manner. By contrast, the reverse pattern of results was evidenced for anthropomorphic explanations.

Table 2
Proportion correct and decision response time (s) for explanation type and validity. Standard error is reported in parentheses.

		P(COR)	DRT	Cal.
Valid	**Train.**	.82 (.03)	16.5s (0.9)	.08 (.01)
	Mech.	.73 (.01)	17.2s (0.6)	.12 (.01)
	Anthro.	.59 (.02)	17.9s (0.8)	.20 (.02)
Invalid	**Train.**	.52 (.01)	16.5s (0.9)	.28 (.03)
	Mech.	.54 (.02)	16.2s (0.7)	.22 (.01)
	Anthro.	.75 (.02)	18.7s (0.6)	.12 (.01)

An interaction of explanation source, validity and animacy was also found to be significant, *F*(1, 58) = 5.656, *MSE* = .017, *p* < .05, η_p^2 = .089. As is clear from Figure 1, participants were more accurate when judging valid statements when they thought that the phenomena were living and the explanations were offered by people and when they thought that the phenomena were non-living and the explanations were offered by scientists.

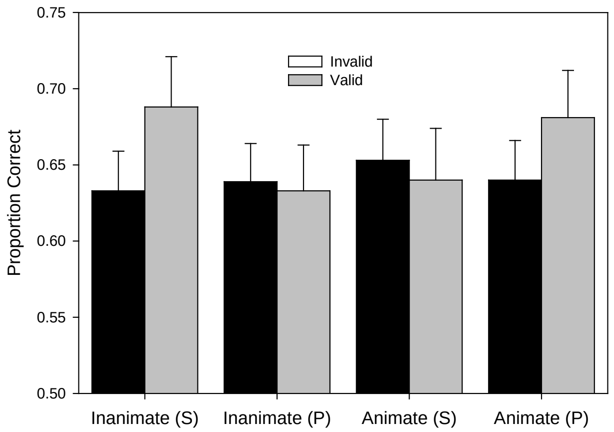

Figure 1. The effect of animacy, explanation source and validity on proportion correct. Explanation sources were either scientists (S) or people (P). Error bars are given in Standard Error (SE = 2).

Decision Response Time

An ANOVA was conducted on decision response time in seconds. Decision response time was affected by explanation type, *F*(1, 58) = 19.578, *MSE* = 25594, *p* < .001, η_p^2 = .252. A marginal interaction of explanation type and validity was also observed, *F*(1, 58) = 3.647, *MSE* = 22996, *p* = .061, η_p^2 = .059. Overall, it took participants longer to assess anthropomorphic syllogisms than mechanistic syllogisms with the marginal interaction of validity indicating that participants took considerably longer when responding to invalid anthropomorphic case.

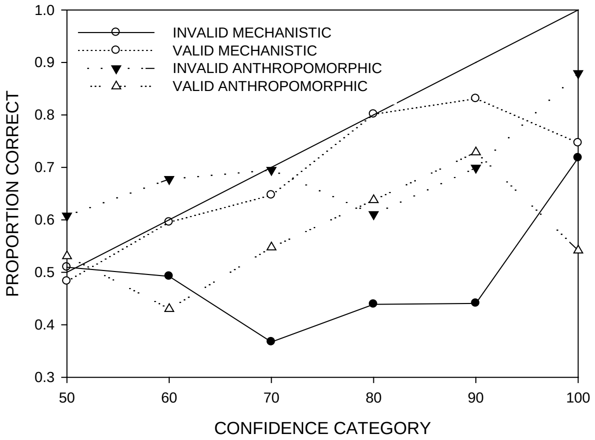

Figure 2. The effect of explanation type and validity on subjective confidence calibration.

Subjective Confidence Calibration

Confidence calibration is defined as the extent to which participants' use of a confidence category deviates from obtained proportion correct (for reviews and discussions see Baranski & Petrusic, 1994). The ANOVA conducted on confidence calibration revealed a significant interaction of explanation type and validity, *F*(1, 58) = 61.2, *MSE* = .014, *p* < .001, η_p^2 = .518. Figure 2 contains the overall results for calibration analysis and Table 2 contains the respective means. The extent to which the calibration curves deviate from perfect calibration (denoted by the diagonal line) indicates that participants were reasonably well calibrated for valid mechanistic and invalid anthropomorphic explanations but poorly calibrated in all other conditions.

Discussion

The results of the experiment support earlier investigation wherein individuals were observed to have response biases induced by task-irrelevant information and prior knowledge. In the present study, the results of analyses of proportion correct revealed that mechanistic explanations were generally perceived to be more valid than anthropomorphic explanations. Such a finding can be taken as support for the belief bias effect (Evans et al., 1983).

It is important to note the nature of the belief bias effect here. As can be seen in Table 2, the pattern of results for mechanistic explanations is comparable to those of the training phase. This suggests that participants are likely making judgments about validity in the same manner for each. Consequently, it at first seems that there may be a bias *against anthropomorphic* explanations rather than *for mechanistic* explanations. Such a finding could be the result of a shift in intentional attribution in early stages of development (Miller & Aloise, 1989; Smith, 1978). In a study where children were required to indicate whether an act was voluntary or involuntary, Smith (1978) found that young children were more likely to impute intentionality to an act than older children. In our adult population, it might be the case that this general decline in attribution of a theory of mind is further suppressed as a consequence of being presented alongside mechanistic explanations.

Interestingly, our data suggest the potential utility of anthropomorphic explanations as a means to communicate information. In Table 2, although a complimentary pattern of performance is found for anthropomorphic and mechanistic explanations, the patterns within these conditions are not the inverse of one another. Participants are more accurate in the valid anthropomorphic condition than in the invalid mechanistic condition. Moreover, in the invalid condition participants exhibit the best calibration for anthropomorphic explanations suggesting that they can more adequately judge their performance in this condition. This pattern of results suggests that although these explanations do not appear to be valid, syllogistic reasoning is facilitated with anthropomorphic explanations. This supports suggestions that thinking about phenomena in terms of intentionality is an effective heuristic for humans (Dennett, 1987; Griffin & Bar-Cohen, 2002), and studies performed in the science education literature that suggest these analogies facilitate the comprehension of relationships between entities (e.g., Bartov, 1981).

Interestingly, we also observed another, more complex, belief-bias effect: participants apparently believed that people were more likely to offer valid explanations of living things whereas scientists were more likely to offer valid explanations of non-living things. A straightforward explanation of this finding is that participants believe that explanation validity is dependent on both the source and the familiarity they have with a domain. Apparently, for our participants, scientists are not as equipped as people to offer adequate explanations of things that are alive!

One expected result that we did not observe was the interaction of source and animacy with explanation type and validity. This is surprising as it seems likely that certain kinds of explanations would be offered by certain agents and not others, and that if something was alive, it would be more likely to be described by an anthropomorphic explanation than if it were nonliving. Although further study of this is required, this might be related to the developmental literature on the emergence of different domain knowledge. For instance, it has been suggested that animacy and an understanding of human intentionality develop at different stages. Our results suggest that this distinction might persist into adulthood.

In general, our findings that belief biases can facilitate performance support the literature that shows framing problems in terms of human interaction alters performance (Cosmides & Tooby, 1992; Gigerenzer & Hug, 1992; Griggs & Cox, 1982). However, prior to adopting anthropomorphic terms as a means to communicate information, it is critical to note that the present study demonstrates that performance is nevertheless suboptimal and that there is a trade-off with the adoption of any heuristic. As we have demonstrated here, neither mechanistic nor anthropomorphic explanations are uniformly better at facilitating reasoning.

A final result that should be considered here is the relationship between accuracy and subjective confidence reports. Previous studies suggest that participants might not have access to the knowledge that is used to successfully complete a syllogistic reasoning task. This implies no correlation between confidence and accuracy (e.g., Shykaruk & Thompson, 2006). As can be seen from Figure 2, a positive relationship was observed in the present study, suggesting that participants are reasonably well-calibrated in some conditions, i.e., valid mechanistic and invalid anthropomorphic explanations. Although one difficulty between comparing the present study and that reported by Shykaruk and Thompson (2006) is that they used a scale inappropriate for confidence calibration (e.g., ratings were given on a scale ranging from 1 to 7 for low and high confidence, respectively). The striking correspondence of confidence to accuracy with our materials suggests that there might be a fundamental difference in the domains examined in their study of belief-bias and those used here. Regardless of this difference, it is clear that participants *were* explicitly aware of their performance in some conditions.

Conclusions

At the most general level, our study is in line with a large literature showing that prior beliefs affect performance in decision-making tasks. The relationships between accuracy, response latency and subjective awareness observed here has broader implications for models of decision-making heuristics. Our study might provide support for a model presented by Glöckner and Betsch (2008). In their framework (see also Kahneman & Tversky, 2002), dissociable processes perform the search strategy and implement the decision rule. An automatic system (System 1) integrates information and executes motor responses whereas an effortful system (System 2) is responsible for information search, production and manipulation of information to enable the automatic system to perform (e.g., Stanovich, 2004). In our study, subjective calibration suggests that participants are explicitly aware of their performance and, in some conditions, this provides a well-calibrated assessment of that performance. Moreover, given that response latencies are slower for anthropomorphic explanations, this suggests a more effortful decision-making process implicating System 2.

If System 2 is involved in an effortful, decision-making heuristic, a working memory task should interfere with performance in the above task as has been demonstrated in other studies of reasoning ability (e.g., Kyllonen & Christal, 1990). In this case, failures of executive function could presumably make it more likely that a participant would use a default schema. Although not investigated here, this possibility is currently being examined by Schoenherr and Thomson (in preparation). In their study, stimuli identical to those used here were presented along with a concurrent working memory load. Rather than showing a reversal toward the acceptance of anthropomorphic statements, participants were more inclined to reject them than the present study. Thus, whereas executive function does appear to be involved, it seems more likely to function as an effortful information search strategy (e.g., Glöckner & Betsch, 2008).

More generally, our research adds to recent studies that pertain to how extraneous information can influence judgments about the validity of scientific arguments. It is clear from the present study that the effect observed by Weisberg et al (2008) is not limited to neuroscientific explanations, and that *many* more factors need to be controlled when examining such biases. Given the ostensibly independent belief-biases associated with the source of information and animacy of the phenomena and those associated with the type of explanation, further studies should examine the set of factors that suggest valid arguments to laypersons and methods for more effectively communicating scientific arguments.

References

Baranski, J. V., & Petrusic, W. M. (1994). The Calibration and resolution of confidence in perceptual judgements. *Perception & Psychophysics*, 55, 412-428.

Bartov, H. (1981). Teaching students to understand the advantages and disadvantages of teleological and anthropomorphic statements in biology. *Journal of Research in Science Teaching*, 18, 79-86.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*. 31**,** 187-276.

Cosmides, L. & Tooby, J. (1992). Cognitive adaptations for social exchange. In *The Adapted mind: Evolutionary Psychology and the Generation of Culture*, ed. J. Barkow, L. Cosmides & J. Tooby, pp. 163-224. New York: Oxford University Press.

Dennett, D. (1987). *The Intentional Stance*. Cambridge: MIT Press.

Dienes, Z., & Berry, D. (1997). Implicit learning: Below the subjective threshold. *Psychonomic Bulletin & Review*, 4, 3-23.

Evans, J. St. B. T., Barston, J., Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition*, 11, 295-306.

Gigerenzer, G. & Hug, K. (1992). Domain-specific reasoning, social contracts, cheating, and perspective change. *Cognition*, 43, 127-171.

Gigerenzer, G. & Selten, R. (2000). *Bounded Rationality: The Adaptive Toolbox*. Cambridge: MIT Press.`

Glöckner, A. & Betsch, T. (2008). Modeling option and strategy choices with connectionist networks: Towards an integrative model of automatic and deliberate decision making. *Judgment and Decision Making*, 3, 215–228.

Griffin, R. & Bar-Cohen, S. (2002). *The intentional stance: developmental and neurocognitive perspectives*. In Brook, A. & Ross, D. (eds.) *Daniel Dennett*. Cambridge Univeristy Press: Cambridge.

Griggs, R. A., & Cox, J. R. (1982). The elusive thematic-materials effect in Wason's selection task. *British Journal of Psychology*, 73, 407–420.

Hirschfield, L. A. & Gelman, S. A. (1994). *Mapping the mind: Domain specificity in cognition and culture*. New York, Cambridge University Press.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, 58, 697-720.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3,* 430–454.

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica, 77,* 217–273.

Kyllonen, P. C. & Christal, R. E. (1990). Reasoning ability is little more working-memory capacity?! *Intelligence*, 14, 389-433.

Liberman, N. & Klar, Y. (1996). Hypothesis testing in Wason's selection task: Social exchange cheating detection or task understanding. *Cognition*, 58, 127-156.

McCloskey, M. (1983). Naïve theories of motion. In *Mental Models*, ed. D. Gentner, A. Stevens, pp. 299-324. Hillsdale, NJ: Erlbaum.

Miller, P. H. & Aloise, P. A. (1989). Young children's understanding of the psychological causes of behaviour. *Child Development*, 60, 257-285.

Mitchell, P., Robinson, E. J., Isaacs, J. E. & Nye, R. M. (1996). Contamination in reasoning about false belief: An instance of realist bias in adults but not children. *Cognition*, 59, 1-21.

Newstead S.E., Evans J. B. T. (1993). Mental models as an explanation of belief bias effects in syllogistic reasoning. *Cognition* 46, 93-97.

Nisbett, R. E. (2003). *The Geography of Thought: How Asians and Westerners Think Differently… and Why*. New York: Free Press.

Sa, W., West, R. F., & Stanovich, K. E. (1999). The domain specificity and generality of belief bias: Searching for generalizable critical thinking skills. Journal of Educational Psychology, 91, 497-510.

Schoenherr, J. R. & Thomson R. (in preparation). *Inducing a bias in reasoning with anthropomorphic and mechanistic explanations by means of a working memory task.*

Smith, M. (1978). Cognizing the behavior stream: The recognition of intentional action. *Child Development*, 49, 736-743.

Shynkaruk, J., M. & Thompson, V. A. (2006). Confidence and accuracy in deductive reasoning. *Memory & Cognition*, 34, 619-632.

Stanovich, K. E. (2004). *The Robot's Rebellion: Finding Meaning in the Age of Darwin*. University of Chicago Press.

Wason, P. C. (1966). Reasoning. In *New Horizons in Psychology*, ed. B. Foss, 135-151. Harmondsworth: Penguin.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, 72, 655-684.

Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E. & Gray, J. R. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience*, 20, 470–477.