Is Conflict Detection in Reasoning Domain General?

Darren Frey (darren@post.harvard.edu)
Paris Descartes University, Sorbonne Paris Cité
CNRS UMR 8240, LaPsyDÉ, Paris, France

Wim De Neys (wim.de-neys@parisdescartes.fr)
Paris Descartes University, Sorbonne Paris Cité
CNRS UMR 8240, LaPsyDÉ, Paris, France

Abstract

A great deal of reasoning research indicates that individuals are often biased by intuitive heuristics. However, contemporary results indicate that individuals seem sensitive to their biases; they seem to detect conflict with reasoning norms. One of the key remaining questions is whether this conflict sensitivity is domain general. To address this question, we administered a battery of five classical reasoning tasks to a large sample of subjects and assessed their conflict detection efficiency on each task by measuring their response confidence. Results indicate that conflict detection is, in most senses, not domain general, though there are compelling exceptions.

Keywords: conflict detection; reasoning; bias; domain generality; decision making

Introduction

That human reasoning is prone to error comes as no surprise to most everyone. Upon reflection, we discern blunders in our own reasoning in the most ordinary of circumstances: we realize we miscalculated how long a trip would take us; we come to acknowledge our snap judgment of a colleague was mistaken. Likewise, we often witness mistakes in others and throughout history, some of which aggregate in the most atrocious of ways: an innocent man is convicted, judged, tried, and sentenced to more than forty years of solitary confinement based on the shakiest possible evidence, the testimony of a single untrustworthy witness (e.g., the case of Albert Woodford, see Aviv, 2017).

Although mistakes like this seem, at first, entirely unrelated, one could argue they often issue from a uniform set of underlying tendencies. Exploring these tendencies has motivated much of the research in reasoning and decision making throughout the past four decades. One compelling and especially generative account of reasoning mistakes contrasts two types of thinking: fast, associative, heuristic thinking and slower, more demanding, rule-based reasoning (Kahneman, 2011). Returning to the example of misjudging a colleague, the fast type of reasoning (System 1) seems to account for our initially misguided impression, which is then revised upon reflection—and after gathering more evidence—by the slower, more deliberate type of thinking (System 2).

The family of dual process theories that rely on contrasting these two forms of reasoning has provided countless testable hypotheses and a diverse set of approaches and methods. While most parties agree that heuristics are useful, efficient, and often optimal means of navigating complex environments, investigators disagree about how often and in what contexts they conflict with logical and mathematical principles. Evans (2003, 2010), Kahneman (2011), and Stanovich and West (2000) insist that individuals regularly make reasoning mistakes because of unchecked heuristic inferences, while Gigerenzer (2008), Katsikopoulos (2013) and others emphasize that heuristics are generally ecologically rational and truth preserving.[1]

Until recently, one of the cardinal doctrines of the dual process account of reasoning mistakes relied on the imperceptibility of reasoning conflicts. Prominent scholars have argued that our reasoning mistakes masquerade beneath our awareness, which is, at least partially, what accounts for their ubiquity. Surely—the argument goes—if reasoners were aware of their mistakes they would correct them. However, many contemporary empirical analyses of reasoning bias suggest that individuals are often sensitive to conflicts between heuristics and normative principles even when they err (e.g., Bonner & Newell, 2010; De Neys & Glumicic, 2008; Pennycook, Fugelsang, & Koehler, 2015; Handley & Trippas, 2015).

Researchers have demonstrated this across a number of diverse reasoning tasks using many different methods. Much of the work relies on contrasting tasks that contain conflicts between intuitively cued heuristics and normative principles with structurally identical tasks containing no such conflict. Standard behavioral markers that index conflict on lower level tasks, like response times (RT) and confidence levels (Yeung & Summerfield), also indicate people are sensitive to conflict in higher level reasoning tasks (Bonner & Newell, 2010; De Neys & Glumicic, 2008; Pennycook, Fugelsang, & Koehler, 2012). Additionally, people tend to fixate visually on the conflicting elements of the tasks, as evidenced in eye and gaze tracking experiments (De Neys & Glumicic, 2008; Ball, Phillips, Wade, &

[1] Given these fairly fundamental disagreements, it should come as no surprise that what counts as a normative response in any number of contexts is hotly debated. For the sake of simplicity, terms like "normative," "correct," and "logical" will be used to indicate conclusions that are considered correct in classical logic and probability.

Quayle, 2006), and they register heightened levels of arousal on skin conductance recordings (De Neys, Moyens, & Vansteenwegen, 2010). There is neuropsychological evidence of conflict sensitivity as well, derived from both fMRI (De Neys, Vartanian, & Goel, 2008; Simon et al,. 2015) and EEG analyses (De Neys, Novitskiy, Ramautar, & Wagemans, 2010). Despite the diversity of methods and tasks supporting these effects, there are many who find the research problematic, largely because its results seem to imply that individuals have fairly immediate access to logical and probabilistic principles (Mata, Schubert, & Ferreira, 2014; Pennycook et al., 2012; Singmann, Klauer, & Kellen, 2014; Travers, Rolison, & Feeney, 2016). Even the most ardent proponents of the work acknowledge that it is still developing and in need of greater clarification (De Neys, 2012, 2014).

Although conflict detection has been explored with a variety of methods and across various tasks, no previous research examines individuals' tendencies to detect conflict across a range of tasks, giving researchers no clear sense of how or whether conflict sensitivities interact. It is unclear, for example, whether a given person's ability to detect a conflict between an intuitively cued heuristic and a reasoning rule on a particular kind of task is related to her ability to do so on different tasks. In essence, a key open question is whether conflict detection is domain general or task specific.

By further clarifying the precise nature of conflict detection, research of this sort will help characterize emergent dual process theories, especially those that explicitly rely on conflict detection mechanisms. For example, Pennycook et al's (2015) three-stage dual-process model relies on differentiating between successful conflict detection and "cognitive decoupling," which is the more resource intensive process of rejecting a conclusion at odds with reasoning rules even when it has been facilitated by a certain heuristic. Crucially, although conflict detection failures are a prominent feature of this model, it is unclear if conflict sensitivity is a stable individual difference. If conflict detection is domain general, then one would expect the prominence of these failures to extend fairly globally. Apart from further specifying the theory, such a conclusion would offer a partial account of the prevalence of cognitive biases. However, if conflict detection is task specific, then one can suppose that empirically observed detection on a given task is largely unrelated to others, and the prevalence of bias needs to be accounted for in other ways.

To address this issue, we presented a battery of the most intensively studied tasks in the field to a large number of reasoners. This enabled us to assess their conflict detection efficiency by measuring their response confidence. Examining the relationship of detection efficiency across the tasks gives us evidence with which to evaluate whether conflict detection is domain general or task specific.

Method

Participants

A total of 318 undergraduates (260 female; average age = 22.32, SD = 6.11) at Paris Descartes University completed the experiment.

Materials

The experiment consisted of adaptations of five classic reasoning tasks. For each of the five tasks, participants received two conflict items, two no-conflict items, and one abstract control, resulting in 25 items. The tasks were as follows.

Bat and Ball Items (BB) The conflict items in this set were modeled after the canonical CRT problem (Frederick, 2005): "A bat and a ball together cost $1.10. The bat costs $1 more than the ball. How much does the ball cost?" The answer that often comes to mind is 10 cents, though the correct answer is 5 cents ($0.05 + $1.05 = $1.10). Participants likely intuitively substitute the "costs $1 more than" phrase with "costs $1," so to generate no-conflict variants one simply removes this phrase (see De Neys, Rossi, & Houdé, 2013).

Ratio Bias Items (RB) Also called "denominator neglect" problems, these items consist of asking participants to choose between two trays, a small tray and a large one, containing a mixture of gray and white marbles. The participants' goal is to get a gray marble, but the marble will be drawn from the tray they select at random. In a conflict item, the absolute value of gray marbles in the large tray is greater than the absolute value of gray marbles in the small tray, but the relative value of gray marbles is greater in the small tray (e.g., 19/100 vs. 2/10, 19% vs. 20%). Since the marble is being selected at random, one should choose the tray that maximizes the relative likelihood of getting a gray marble (the small tray), but participants are often intuitively and immediately drawn to the larger tray. To generate no-conflict items one aligns the relative and absolute values in a tray, so that the tray most likely to have a gray marble—the one with the highest relative value—is also the most perceptually salient one—the one with the highest absolute value (e.g., 21/100 vs. 2/10, see Bonner & Newell, 2010).

Syllogism Items (SYL) Syllogisms are fundamental arguments in classical logic that consist of two premises and a conclusion, which necessarily follows from the premises when the argument is valid. When the conclusion is at odds with common beliefs, participants tend to deem it logically invalid even when explicitly told just to evaluate the argument's validity (Markovits & Nantel, 1989). A conflict item consists of a logically valid (or invalid) argument structure with an unbelievable (or believable) conclusion. Here is an example of an unbelievable but valid argument: All mammals can walk. Whales are mammals. ∴ Whales can walk. No-conflict items are those in which common beliefs and the argument's logical structure both cue the

same response. All problems were based on Markovits and Nantel's material (1989).

Base Rate Items (BR) Base rate items consist of statistics describing a sample from which an individual is randomly selected along with a description of the individual. Here is an example of a conflict item: "In a study 1000 people were tested. Among the participants there were 5 sixteen-year-olds and 995 forty-year-olds. Lisa is a randomly chosen participant of the study. Lisa likes to listen to techno and electronic music. She often wears tight sweaters and jeans. She loves to dance and has a small nose piercing. What is most likely? (A.) Lisa is sixteen. (B.) Lisa is forty." This item creates a conflict by calling to mind a stereotype that is at odds with the statistically most likely outcome. To generate no-conflict items, one aligns the statistics and the intuitively cued heuristic. For example to turn the above item into a no-conflict example, switch the base rates so the the sample consists of 995 sixteen-year-olds and 5 forty-year-olds. All problems were based on De Neys and Glumicic's (2008) material.

Conjunction Items (CON) Modeled on the classic Linda problem, participants received descriptions about individuals that either intuitively prompt a single statement (no-conflict) or a conjunctive statement (conflict), and they are asked to decide which statement is most likely. Since a single statement is always more likely than a conjunctive statement, subjects should always choose the single statement regardless of whether it coheres with the stereotype. Here is an example of a conflict item: "Jon is 32. He is intelligent and punctual but unimaginative and somewhat lifeless. In school, he was strong in mathematics but weak in languages and art. Which one of the following statements is most likely? (A.) Jon plays in a rock band. (B.) Jon plays in a rock band and is an accountant." Since the description generally cues an accountant stereotype, subjects often wrongly choose the less likely option, B. No-conflict items simply isolate the heuristically cued option. All problems were based on De Neys, Cromheeke, and Osman's (2011) material.

Procedure

The participants were tested in groups of no more than thirty students in a silent classroom at the beginning of a course. In addition to the conflict and no-conflict items illustrated above, participants answered one abstract neutral problem per task. These were designed to query abstract knowledge of relevant reasoning rules and were variants of the above tasks with no clear, consistent intuitive or heuristic prompts. Accuracy on the neutral control items was high (mean accuracy 81.6%, SD = 0.18). All analyses were run filtering for controls and they made no significant impact on any of the results. Thus, we will present only our unfiltered data in what follows and will not discuss the control items further.

The overall structure of the experiment, a within subject design, was manipulated in three ways: it was balanced for conflict content, task order, and conflict presentation order. The conflict and no-conflict contents were balanced across participants, such that half the participants received, for example, the conflict conjunction item above, while the other half received its no-conflict analogue, and vice versa. Additionally, the order in which a given task was presented varied, as did whether an individual first saw a conflict or no-conflict item. A partial Latin square of these factors generated 10 different experiment formats, which were distributed evenly across the participant sample.

All items were presented on their own page. At the bottom of which there was a scale where participants indicated how confident they were in their response on a range from 0% (not at all confident) to 100% (completely confident).

Results

Accuracies

Table 1 (first two rows) presents averages of accuracy levels on each of the tasks, separated by conflict status. Replicating classical findings, performance on no-conflict items was consistently higher than performance on conflict items. In all cases except for RB, contrasts between performance on conflict and no-conflict items was significant (all BB/SYL/BR/CON $t > 10.07$, $p < 0.001$; RB: $t(315) = 1.10$, $p = 0.28$).

Table 1: Accuracies and Conflict Detection Effects

Task	Conflict Accuracy (SD)	No-Conflict Accuracy (SD)	Conflict Detection (SD)
BB	51.75% (47.70)	98.10% (11.77)	23.15% (31.44)
RB	80.05% (34.20)	76.9% (36.30)	11.88% (13.00)
SYL	59.30% (38.51)	84.10% (27.36)	0.37% (19.88)
BR	34.12% (39.21)	92.15% (19.08)	10.47% (20.84)
CON	24.37% (36.56)	96.40% (14.14)	11.62% (16.73)
Avg	49.92% (23.30)	89.53% (11.10)	9.50% (13.90)

Conflict Detection

To get a sense of how widely conflict detection efficiency is distributed across tasks, it is useful to look at what proportion of the sample tended to detect conflict on the entire battery. At the aggregate level, averaged across tasks, we observe most individuals (74.70%) tend to a lowered confidence level on conflict vs. no-conflict items. Across all tasks, this difference amounts to a 9.50% diminution in confidence on incorrectly solved conflict items compared to correctly solved conflict items, $t(307) = 12.01$, $p < 0.001$. This is roughly reflected in the task by task contrasts, though it is highly variable. For example, in the case of the BB items the confidence diminution was 23.15%, while most others hovered around 10%, and, in contrast with previous findings (Stupple, Ball, Evans, Kamal-Smith, 2011), there was little difference between confidence levels on SYL items (0.37%). In all cases except for SYL, t(177)

= 0.50, p = 0.61, the task specific confidence decrease was significant: BB/RB/BR/CON: all t > 2.65, all p < 0.01.

Task Specificity and Domain Generality

With a view to evaluating whether conflict detection is task specific or domain general, we ran four primary kinds of analyses: correlations between conflict items across tasks; correlations between conflict detection effects across tasks; analyses of the distribution of conflict detection effects by individual; and regressions to predict conflict detection effects with a composite meant to uncover diffuse evidence of domain generality.

Table 2 summarizes the results of the first two analyses. The statistics above the diagonal are correlations between conflict detection effects across biased individuals on each of the tasks. The correlations below the diagonal are between accuracies on each of the conflict items across tasks for all participants (N = 318). Accuracies on conflict items were significantly correlated between all tasks (all p < 0.04), although most correlations are fairly modest, ranging from 0.12 to 0.32.

Table 2: Conflict Accuracy and Detection Correlations

	BB	RB	SYL	BR	CON
BB		0.06_{56}	-0.09_{114}	0.07_{142}	0.04_{145}
RB	0.24**		0.22_{55}	-0.02_{76}	0.15_{75}
SYL	0.26**	0.23**		0.14_{150}	0.06_{158}
BR	0.17**	0.12*	0.11*		0.16^{*}_{227}
CON	0.12**	0.13*	0.13*	0.32**	

* 0.05 < p < 0.01; ** p < 0.01. Subscripts indicate Ns.

As a reminder, a conflict detection effect is, in this context, a diminution of confidence on an incorrectly solved conflict item relative to a correctly solved no-conflict item. Given the notorious noisiness and subjectivity of confidence measures and the difficulty of interpreting conflict detection effect sizes (Frey, Johnson, & De Neys, 2017), this is measured in a binary way: one either shows the effect or one does not. In stark contrast to the pattern below the diagonal, correlations between conflict detection effects—those above the diagonal—are almost uniformly insignificant. The only exception is CON & BR, which was correlated at 0.16 (p < 0.02). The correlation between BR & SYL was marginally significant, r = 0.144, p = 0.078. Additionally, Bayes Factors for the correlations were all below 0.52, except CON & BR, which was 1.59.

The binary correlations of conflict detection effects provide no real evidence of domain generality. However, if there is a general and diffuse signal, why should that be captured by simple, pairwise correlations? Perhaps it is the case that conflict detection on a particular task is better predicted by a non-specific and global sensitivity to conflict across tasks. We ran a regression analysis to address this hypothesis, using the combined predictive power of a participant's responses across all tasks. In particular, we used logistic regressions to determine whether conflict detection on a given task was predicted by one's tendency to detect conflict on all of the other tasks. For example, to see if we can predict whether an individual shows an effect on the BB items, we tallied how often she showed an effect on all the other items (RB, SYL, BR, and CON) and used the latter as our predictor variable.

Table 3: Predicting Conflict Detection Effects

Model	Beta	Z	P	Pseudo R^2
BB	1.02	5.46	< 0.01	0.17
RB	0.37	1.34	0.18	0.02
SYL	0.07	0.40	0.69	< 0.01
BR	0.23	1.99	0.05	0.01
BR2	0.12	0.61	0.55	< 0.01
CON	0.33	2.08	0.04	0.01
CON2	0.20	1.00	0.32	< 0.01

As is clear from Table 3, the goodness of fit of these first models (BB, RB, SYL, BR, CON) was generally quite low. The pseudo R^2's (McFadden's) range from < 0.001 to 0.02, except for the BB model which was at the limit of what is considered reasonably good (0.17). The relative goodness of this model is reflected in its higher beta coefficient (1.02), which is significant (p < 0.001).

The only other models with significantly predictive coefficients were the BR and CON models. However, given the tight correlation between these items discovered in the first analysis, we wondered if this was driving the effect. Indeed, if one runs a restricted model, omitting the correlate of the predictor (leaving out CON in the BR case, and vice versa), the models (BR2 and CON2 in Table 3) have inferior goodness of fit and coefficients that are both smaller and no longer significant.[2]

Individual Differences

So far, we have uncovered little evidence of domain generality in conflict detection. However, most of the previous findings rely on averaging effects across reasoners. It might well be the case that there are individuals who show evidence of fairly generalized conflict detection. The concern we address in this section is that important differences between individuals might be lost by aggregating as we have, a concern that echoes theorists who emphasize the importance of examining individual differences in reasoning and decision making (Baron, 2010).

[2] This analysis was meant to assess whether there was a rather diffuse and non-specific conflict detection signal that predicted an individual's detection on a given task by a composite of their relative effects on other tasks. Were the data not binary, a factor analysis would perhaps be appropriate here. Essentially, this was the most liberal test we could devise to check for generality of conflict detection effects. However, it is worth noting that a more conventional test, using multiple regressions with all tasks as predictors except the one being predicted, generated the exact same pattern of results, with all models being uninformative except where BR & CON items were concerned.

To address this issue, we present a final means of characterizing the sample's overall conflict sensitivity, which is summarized in Figure 1. We scored every biased participant individually in order to get a sense of the distribution of conflict detection effects. For a given individual, the total number of tasks on which she showed a conflict detection effect was divided by the total number of tasks on which she was biased, giving us a range of detection levels spanning from 0 (showing no effect on any of the tasks on which an individual is biased) to 1 (showing an effect on all of the tasks on which an individual is biased).

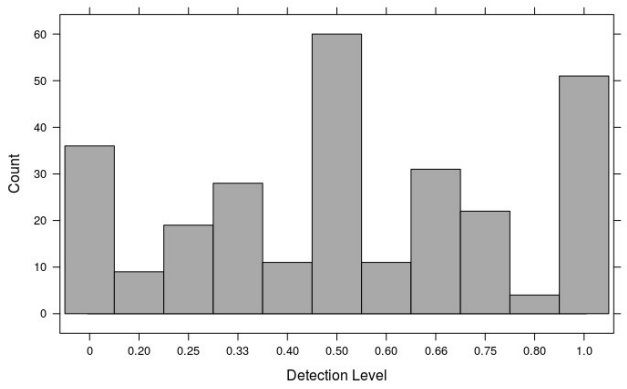

Figure 1: Frequency of Reasoners by Detection Level

There are concentrations of individuals who consistently detect (Detection Level 1: 18.09% of the sample), detect half the time (Detection Level 0.50: 21.28% of sample), and who consistently do not detect (Level 0: 12.77% of the sample), with all other participants distributed between these three groups. The observation that up to 13% of the sample shows a Detection Level 0 is in line with previous findings that suggest there are subsets of reasoners who consistently fail to detect conflict (Frey, Johnson, & De Neys, 2017, Pennycook et al., 2015). The additional observation that 18% of the sample shows perfect detection across all tasks also implies that there might be exceptions to the overall trend toward task specificity. Although this distribution is compatible with the few studies that have explored individual differences in conflict detection previously, we cannot confirm the representativeness of this kind of a distribution given our methods, as we could have arrived at it by chance.

Discussion

While performance on conflict items was consistently correlated, we found no clear indication that conflict detection is similarly correlated. Even using a more liberal measure, one that leverages the predictive power of the entire panel of tasks to anticipate conflict detection on a single task, there was only the faintest signal of generality. Nevertheless, base rate (BR) and conjunction (CON) items were correlated, and the more liberal regression models relying on them were minimally predictive, as was the model predicting the bat and ball (BB) problems. So we found, additionally, no clear evidence of hard and fast task specificity.

One might classify our findings as “domain specific,” where a domain is defined as a set of problems that share similar reasoning rules subject to comparable competing intuitive heuristics. From such a perspective, base rate and conjunction items would be considered to fall within the same domain, as they share similar underlying reasoning structures (statistics and probabilities, respectively) that are in conflict with comparable intuitively prompted heuristics (social stereotypes in both cases), and in indeed both were developed to evaluate biases resulting from the representativeness heuristic.

This hybrid outcome has a number of exciting theoretical features and practical applications. For example, Teovanović, Knežević, & Stankov (2015) argue against a single, explanatory factor underlying cognitive biases that one can easily relate to general intelligence. The account we present here is commensurate with those findings, as it seems indicative of multiple, often dissociable loci of conflict detection failures. Additionally, one of the implications of our findings is that a conflict detection failure on a given task may be largely dissociable from a conflict detection failure on a distant task. This is a hopeful conclusion, especially given the evidence that at the individual level such failures are a non-negligible source of reasoning bias (e.g., Pennycook et al., 2015). The prominence of conflict detection failures on a certain task need not paint a grim picture of reasoning globally. However, the association within what we are calling a domain indicates that at points detecting on a given task will be related to detection on a different task, a relationship that could be exploited educationally. For example, a reasonable pedagogical strategy might begin by allocating resources to the easier of two related tasks, relying on the shared conflict prompting structures to aid in instructional transfer and facilitate instruction on the second task.

These findings raise many additional questions. Since confidence measures are inherently noisy, our results are necessarily tentative. It will be important to revisit the question of the domain generality of conflict detection with additional measures, especially response times. Additionally, given that we were interested in performance across many tasks, we were only able to use a few items per task, so our findings need to be interpreted cautiously. Another particularly promising research project will be to further characterize those individuals who detect conflict in a domain general manner. For example, it would be particularly instructive to determine whether they share similar general cognitive capacities or tend have related thinking dispositions.

Acknowledgements

This research was supported by a research grant (DIAGNOR, ANR-16-CE28-0010-01) from the Agence National de la Recherche. Additionally, Darren Frey is supported by the Sorbonne Paris Cité International Grant (INSPIRE).

References

Aviv, R. (2017, January). How Albert Woodfox Survived Solitary. *The New Yorker.*

Ball, L. J., Phillips, P., Wade, C. N., & Quayle, J. D. (2006). Effects of belief and logic on syllogistic reasoning: Eye-movement evidence for selective processing models. *Experimental Psychology*, *53*(1), 77–86.

Baron, J. (2010). Looking at Individual Subjects in Research on Judgment and Decision Making (or anything). *Acta Psychologica Sinica*, *42*(1), 88–98.

Bonner, C., & Newell, B. R. (2010). In conflict with ourselves? An investigation of heuristic and analytic processes in decision making. *Memory & Cognition*, *38*(2), 186–196.

De Neys, W. (2012). Bias and Conflict: A Case for Logical Intuitions. *Perspectives on Psychological Science*, *7*(1), 28–38.

De Neys, W. (2014). Conflict detection, dual processes, and logical intuitions: Some clarifications. *Thinking & Reasoning*, *20*(2), 169–187.

De Neys, W., Cromheeke, S., & Osman, M. (2011). Biased but in Doubt: Conflict and Decision Confidence. *PLoS ONE*, *6*(1), e15954.

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition*, *106*(3), 1248–1299.

De Neys, W., Moyens, E., & Vansteenwegen, D. (2010). Feeling we're biased: Autonomic arousal and reasoning conflict. *Cognitive, Affective, & Behavioral Neuroscience*, *10*(2), 208–216.

De Neys, W., Novitskiy, N., Ramautar, J., & Wagemans, J. (2010). What makes a good reasoner?: Brain potentials and heuristic bias susceptibility. In *Proceedings of the Annual Conference of the Cognitive Science Society* (Vol. 32, pp. 1020–1025).

De Neys, W., Rossi, S., & Houdé, O. (2013). Bats, balls, and substitution sensitivity: cognitive misers are no happy fools. *Psychonomic Bulletin & Review*, *20*(2), 269–273.

De Neys, W., Vartanian, O., & Goel, V. (2008). Smarter than we think: when our brains detect that we are biased. *Psychological Science*, *19*(5), 483–489.

Evans, J. S. B. T. (2003). In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences*, *7*(10), 454–459.

Evans, J. S. B. T. (2010). Intuition and Reasoning: A Dual-Process Perspective. *Psychological Inquiry*, *21*(4), 313–326.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 25–42.

Frey, D., Johnson, E., De Neys, W. (2017, forthcoming) Individual Differences in Conflict Detection During Reasoning. *Quarterly Journal of Experimental Psychology*

Gigerenzer, G. (2008). *Gut Feelings: The Intelligence of the Unconscious* (Reprint edition). Penguin Books.

Handley, S. J., & Trippas, D. (2015). Chapter Two-Dual processes and the interplay between knowledge and structure: a new parallel processing model. *Psychology of Learning and Motivation*, *62*, 33–58.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Katsikopoulos, K. V. (2014). Bounded rationality: the two cultures. *Journal of Economic Methodology*, *21*(4), 3 61–374.

Markovits, H., & Nantel, G. (1989). The belief-bias effect in the production and evaluation of logical conclusions. *Memory & Cognition*, *17*(1), 11–17.

Mata, A., Schubert, A.-L., & B. Ferreira, M. (2014). The role of language comprehension in reasoning: How "good-enough" representations induce biases. *Cognition*, *133*(2), 457–463.

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2012). Are we good at detecting conflict during reasoning? *Cognition*, *124*(1), 101–106.

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive Psychology*, *80*, 34–72

Simon, G., Lubin, A., Houdé, O., & Neys, W. D. (2015). Anterior cingulate cortex and intuitive bias detection during number conservation. *Cognitive Neuroscience*, *6*(4), 158–168.

Singmann, H., Klauer, K. C., & Kellen, D. (2014). Intuitive logic revisited: new data and a Bayesian mixed model meta-analysis. *PloS One*, *9*(4).

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: implications for the rationality debate? *The Behavioral and Brain Sciences*, *23*(5), 645–665; discussion 665–726.

Stupple, E. J. N., Ball, L. J., Evans, J. S. B. T., & Kamal-Smith, E. (2011). When logic and belief collide: Individual differences in reasoning times support a selective processing model. *Journal of Cognitive Psychology*, *23*(8), 931–941. https://doi.org/10.1080/20445911.2011.589381

Teovanović, P., Knežević, G., & Stankov, L. (2015). Individual differences in cognitive biases: Evidence against one-factor theory of rationality. *Intelligence*, *50*, 75–86.

Travers, E., Rolison, J. J., & Feeney, A. (2016). The time course of conflict on the Cognitive Reflection Test. *Cognition*, *150*, 109–118.

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1310–1321.