Advancing fast and slow thinking theorizing : exploring the role of intuition across domains

Aikaterini Voudouri

Université Paris Cité

Ecole doctorale 261 - Cognition, Comportements, Conduites humaines

Laboratoire de Psychologie du Développement et de l'Éducation de l'enfant (UMR 8240, CNRS)

Advancing fast and slow thinking theorizing: Exploring the role of intuition across domains

Par **Aikaterini VOUDOURI**

Thèse de doctorat de Psychologie

Dirigée par **Wim DE NEYS**

Et par **Michał BIAŁEK**

Présentée et soutenue publiquement le 28/11/2023

Devant un jury composé de :

Wim DE NEYS	Directeur de recherche à l'Université Paris Cité	Directeur de thèse
Michał BIAŁEK	Professeur des universités à l'Université de Wroclaw	Co-directeur de thèse
Jérôme PRADO	Chargé de recherche à l'Université de Lyon	Rapporteur
Maria AUGUSTINOVA	Professeur des universités à l'Université de Rouen-Normandie	Rapporteur
Tamara VAN GOG	Professeur des universités à l'Université d'Utrecht	Examinateur
Eva VAN DEN BUSSCHE	Professeur des universités à l'Université de Louvain	Examinateur

Abstract

Advancing fast and slow thinking theorizing: Exploring the role of intuition across domains

Popular dual process models suggest that sound reasoning requires correction of fast, intuitive thought processes by slower, controlled deliberation. However, recent findings in logical reasoning have started to question this characterization. These studies employed classic heuristics-and-biases tasks and showed that the sound, logical response traditionally assumed to arise after deliberation is often cued by mere intuitive processing. Additionally, even when people provide biased responses, they frequently show an intuitive awareness of the problem's logical principles. The present thesis aimed to extend the generalizability of these findings beyond the field of logical reasoning, to other domains where the sound response is also traditionally believed to be cued after deliberation. Encompassing a broad range of fields, from decision-making under risk, to high-level and low-level tasks, this thesis explored whether the alleged deliberate response could also be provided intuitively and whether people possess an intuitive sensitivity to their errors. To identify the presumed intuitive response that precedes the response given after deliberation, the two-response paradigm was used across all studies. In this paradigm participants give two consecutive responses to the same problem in each trial: an initial intuitive response under time-pressure and cognitive load, followed by a final response without constraints where they can freely deliberate. In Chapter 1, I examined decision-making under risk and found that when people gave the expected-value maximizing response after deliberation, they had frequently arrived to the same response already from the initial, intuitive stage. Moreover, even when people remained loss averse, they often showed an intuitive sensitivity to expected value, as indicated by decreased confidence. In Chapter 2, I delved into high-level semantic tasks, demonstrating that while avoiding semantic illusions often requires deliberation, intuitive responding can also lead to correct answers. Additionally, people consistently displayed error sensitivity, even in the initial stage when deliberation was minimized. In Chapter 3, I focused on low-level cognitive control tasks, such as the Stroop and Flanker tasks, and found that the majority of correct responses were already provided in the initial stage,

when deliberate control was constrained. In this chapter I also explored the association between Stroop and reasoning performance. In Chapter 4, I investigated the stability of biases in heuristics-and-biases tasks and the impact of conflict detection on long-term answer change. The results indicated that both intuitive and deliberate responses remain highly stable, though not entirely, after two weeks. Critically, conflict detection was found to be a predictor of answer change; the more conflicted people felt about their responses when solving a problem, the more likely they were to change their responses over time. Across all chapters, it became evident that responses once thought to require deliberation often stemmed from mere intuitive processing and people frequently showed intuitive sensitivity to their errors. These findings establish the applicability of a recent, revised dual process framework across different domains and temporal dimensions. This thesis thereby suggests that, in general, human thinking can be better characterized as an interplay between different types of "fast" intuitions, rather than a strict dichotomy between "fast" and "slow" thinking.

Keywords: dual-process theory, conflict detection, two-response paradigm, intuition, risky decision making, semantic illusions, cognitive control

Résumé

Avancement de la théorisation de la pensée rapide et lente : Exploration du rôle de l'intuition à travers différents domaines

Les modèles de double processus suggèrent qu'un raisonnement sain nécessite la correction des processus de pensée rapides et intuitifs par une délibération plus lente et contrôlée. Toutefois, de récentes découvertes dans le domaine du raisonnement logique ont commencé à remettre en question cette caractérisation. Ces études, qui utilisent des tâches classiques d'heuristiques et de bais, ont montré que la réponse logique traditionnellement supposée découler d'une délibération est souvent déclenchée par un simple traitement intuitif. En outre, même lorsque les gens fournissent des réponses biaisées, ils font souvent déjà preuve d'une sensibilité intuitive des principes logiques du problème. La présente thèse vise à étendre la généralisation de ces résultats au-delà du domaine du raisonnement logique, et à d'autres domaines où la réponse correcte est également traditionnellement considérée comme étant déclenchée après délibération. Couvrant un large éventail de domaines (de la prise de décision en situation de risque, aux tâches de haut niveau et de bas niveau), cette thèse a cherché à déterminer si la présumée réponse délibérée pouvait également être fournie de manière intuitive, et si les individus possédaient une sensibilité intuitive à leurs erreurs. Pour identifier la réponse intuitive qui précède la réponse donnée après délibération, le paradigme à deux réponses a été utilisé. Dans ce paradigme, à chaque essai, les participants donnent deux réponses consécutives au même problème: une première réponse intuitive sous la pression du temps et de la charge cognitive, suivie d'une réponse finale, sans contrainte, où ils peuvent délibérer librement. Dans le Chapitre 1, j'ai examiné la prise de décision en situation de risque et j'ai constaté que lorsque les participants donnaient la réponse maximisant l’espérance mathématique après délibération, ils parvenaient souvent à la même réponse dès la phase initiale, intuitive. En outre, même lorsque les individus restent aversifs aux pertes, ils font souvent preuve d'une sensibilité intuitive à l’espérance, comme l'indique la diminution de la confiance. Dans le Chapitre 2, j'ai approfondi les tâches sémantiques de haut niveau, démontrant

que si éviter les illusions sémantiques nécessite souvent une délibération, la réponse intuitive peut également conduire à des réponses correctes. Les participants ont également toujours montré une sensibilité à l'erreur, même dans la phase initiale, lorsque la délibération était réduite. Dans le Chapitre 3, je me suis concentrée sur des tâches de contrôle cognitif de bas niveau, telles que les tâches de Stroop et Flanker, et j'ai constaté que la majorité des réponses correctes étaient déjà fournies pendant la phase initiale, lorsque le contrôle délibéré était limité. Dans ce chapitre, j'ai également étudié l'association entre la tâche de Stroop et les performances de raisonnement. Dans le Chapitre 4, j'ai étudié la stabilité des biais dans les tâches d'heuristiques et de biais. Les résultats indiquent que les réponses intuitives et délibérées restent en grande partie stables, mais pas entièrement, après deux semaines. La détection des conflits s'est révélée être un facteur prédictif du changement de réponse ; plus les participants se sentaient en conflit avec leurs réponses, plus ils étaient susceptibles de modifier leurs réponses au fil du temps. A travers l'ensemble de ces chapitres, il est apparu évident que les réponses que l'on croyait avoir besoin de délibération provenaient souvent d'un traitement intuitif, et que les individus montraient une sensibilité intuitive à leurs erreurs. Ces résultats démontrent l'applicabilité d'un cadre récent et révisé du double processus dans différents domaines et dimensions temporelles. Cette thèse suggère donc que la pensée humaine peut être mieux caractérisée comme une interaction entre différents types d'intuitions "rapides", plutôt que comme une stricte dichotomie entre la pensée "rapide" et la pensée "lente".

Mots clefs : théorie du double processus, détection des conflits, paradigme à deux réponses, intuition, prise de décisions risquées, illusions sémantiques, contrôle cognitif

Acknowledgements

I would like to begin by thanking my supervisor, Wim De Neys. Thank you for your invaluable mentorship, without which this thesis would not have been possible. I have learned a lot from working with you and I really appreciate your consistent support. You provided me with constant encouragement and gave me space when I needed it. I am very thankful for your tireless feedback and your commitment to being a meticulous researcher, all while allowing room for creativity and for personal growth. Last but not least, you are a very fun dinner and party companion, thank you for all the good times.

To my co-supervisor, Michał Białek, thank you for your trust in me and for your warm welcome in Wroclaw (and for teaching me how to pronounce it: Vrotswaf). Our discussions in Poland broadened my views on research and I was inspired by your ability to successfully combine so many research interests at once. Thank you for the lunches we shared–the Nepalese place will not be forgotten.

I would like to thank Esther Boissin for welcoming me into the lab, accompanying me in the initial stages of my PhD, and patiently teaching me how to use R. Your liveliness, intellect and constant support made me feel like I always had a person to rely on.

Nina Franiatte, thank you for your unconditional acceptance and open-mindedness; you are a great person. The moments we shared exploring Boston and our meaningful conversations remain among the fondest memories of my PhD journey.

Jérémie Beucler, I am very grateful for your substantial contribution to this thesis. Your intellect, diligence and creativity have undoubtedly made me become a better researcher. Thank you for your witty humour and your American and Scottish impersonations.

Matthieu Raoelison, thank you for always being willing to help. Nydia Vurdah, thank you for being my office buddy from day one; I am grateful that we shared our challenges and triumphs throughout this process. My gratitude also goes to my fellow PhD students that were there when I joined the lab and

welcomed me. Special thanks to Chiara Andreola and Iris Menu for always reassuring me during stressful moments.

To all the PhD students and postdocs who joined the lab after my arrival, thank you for making the 6th floor a vibrant and supportive place. Thank you to Prany Wantzen, Leticia Kolberg, Julia Houdayer and Elora Taieb for your calming presence and smiles. Julia Mathan for your positive and fun energy. Mélanie Maximino-Pinheiro for being such a reliable PhD representative. Gaëlle Rouvier for always genuinely asking how things were going. Nicolas Beauvais for your light-heartedness and for allowing me to make silly jokes; it was great working with you. Marine Lemaire for being open about the stressful moments and making me feel less alone.

I would like to thank Grégoire Borst for providing me with opportunities to attend conferences and summer schools, and for always leaving his office door open. Claire Pruvot thank you for your help with difficult administrative situations. Finally, thank you to all the researchers in the lab, whose diverse expertise and approachability has made Lapsydé a scientifically enriching environment.

To my mother, Xanthippi, thank you for seeing who I am and unconditionally accepting me, irrespective of my successes or setbacks, and for teaching me to not take things too seriously. Achillea, thank you for being the person I can always talk to and for understanding my thoughts before I express them; we are not the same person but it is the closest it can get. Giagia Fotini, thank you for inspiring me with your strong work ethic and your hope in life, for always keeping an eye out for me and for letting me know that you have my back.

To my father, Giannis, thank you for your reassuring words, the confidence-boost phone calls and for always picking up the phone. To Aggelos, thank you for being a great listener, for believing in me and teaching me to view difficulties with humour. To Dimitris, thank you for your constant support and for being there ever since I can remember. Fotini, thank you for always being my friend and for making me laugh so much. Olympia Kastania and the rest of the group, thank you for helping me get closer to the person I aspire to be.

Soufiane, thank you for patiently answering all my questions with understanding, for gently reassuring me that I am good enough, and for being there for me every single day. Your presence and companionship are priceless in my life.

Table of Contents

Scientific output

This thesis comprises the following peer-reviewed journal articles that have been accepted or submitted for publication:

Chapter 1: **Voudouri, A.**, Bialek, M., & De Neys, W. (under review). Fast & slow decisions under risk: Intuition rather than deliberation drives advantageous choices. *Cognition*.

Chapter 2: Beucler, J., **Voudouri, A.**, & De Neys, W. (under review). Semantic illusions, fast and slow. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Chapter 3: **Voudouri, A.**, Bago, B., Borst, G., & De Neys, W. (2023). Reasoning and cognitive control, fast and slow. *Judgment and Decision Making, 18*, E33. https://doi.org/10.1017/jdm.2023.32.

Chapter 4: **Voudouri, A.**, Bialek, M., Domurat, A., Kowal, M., & De Neys, W. (2022). Conflict detection predicts the temporal stability of intuitive and deliberate reasoning. *Thinking & Reasoning*, 1-29. https://doi.org/10.1080/13546783.2022.2077439

In addition, the following scientific output was achieved between the start and the completion of this Ph.D. and is not included in the thesis:

Boissin, E., Caparos, S., **Voudouri, A.**, & De Neys, W. (2022). Debiasing system 1: Training favours logical over stereotypical intuiting. *Judgment and Decision Making*, *17*, 646-690. https://journal.sjdm.org/22/220326/jdm220326.pdf

Symbols and Acronyms

The following list contains the symbols, acronyms, and abbreviations used in this thesis.

General

CRT	Cognitive Reflection Test
EV	Expected Value
V	Value
BB	Bat-and-ball problems
BR	Base-rate problems
SYL	Syllogistic reasoning problems
CONJ	Conjunction fallacy problems
FOR	Feeling of Rightness
s	seconds
ms	milliseconds
£	British pound sterling

Statistical parameters

P	Probability
n	Group size
N	Sample size
M	Mean
Mdn	Median
SD	Standard Deviation
SE	Standard error
Min	Minimum
Max	Maximum
Q1	First Quartile
Q3	Third Quartile
IQR	Interquartile range
CI	Confidence Interval
p	p value
t	T-Statistic

V	V-Statistic (Wilcoxon Signed-Rank Test Statistic)
F	F-statistic
W	W-statistic (Sperman's rank correlation test)
r	correlation coefficient
b	Regression coefficient
$\eta^2 g$	Partial eta squared
df	Degrees of freedom

Introduction

General Background

Thinking characterizes the human experience, constantly guiding us through a myriad of choices. Sometimes thinking requires time and effort to arrive at solutions. For instance, before selecting a mortgage plan one will presumably carefully consider the available options. Conversely, in some situations thinking can be effortless, such as grabbing the keys before leaving the house or solving basic math problems like "2 + 2". This duality in human cognition has led to the idea that there are two distinct modes of thinking: one fast, intuitive and effortless, the other slower, reflective and more effortful (Frankish & Evans, 2009). The distinction between a more fast and intuitive and a slower and deliberate thinking process has been at the center of dual-process theories of human cognition, which have popularized them as "System 1" and "System 2" respectively (Epstein, 1994; Evans, 2008; Kahneman, 2011; Sloman, 1996).

The influence of dual process theories in human thinking has been far-reaching and applied to numerous disciplines (Melnikoff & Bargh, 2018). Dual process theories have been applied, among others, to research on cognitive biases and behavioral economics (Evans, 2002; Kahneman, 2011), moral judgement (Greene & Haidt, 2002), human cooperation (Rand et al., 2012), education (e.g., (Beaulac & Kenyon, 2018), susceptibility to fake news (Pennycook & Rand, 2019), and machine intelligence (Bonnefon & Rahwan, 2020). One of the earliest areas within the cognitive sciences to popularize dual process models was the study of biases in logical reasoning (Kahneman, 2000; Kahneman & Tversky, 1972, 1973; Tversky & Kahneman, 1974; Wason & Evans, 1974). Studies in this field have shown that people often breach fundamental logico-mathematical and probabilistic principles when solving tasks which cue heuristic responses that conflict with these principles.[1] A famous example of such heuristics-and-biases tasks is the Linda problem (Tversky & Kahneman, 1983):

[1] In this thesis "logical" is used as a general term to refer to logical, probabilistic, and mathematical principles.

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.
> Which is more likely:
> a) Linda is a bank teller
> b) Linda is a bank teller and is active in the feminist movement

While the statement "Linda is a bank teller" does not fit well with the stereotypical description of Linda, the statement "Linda is active in the feminist movement" aligns strongly with the description. Thus, when presented with this scenario the majority of reasoners use the stereotype as a shortcut or "heuristic" and opt for option b (Tversky & Kahneman, 1983). However, according to probability theory, the possibility of a single event occurring is always higher than the possibility of the conjunction, so the single statement (option a) is always the correct choice. Despite the simplicity of this rule, the majority of people ignore it and choose the option that aligns with the stereotypical heuristic instead (Tversky & Kahneman, 1983). Similar findings have emerged in other heuristic-and-biases tasks, which cue heuristic responses that counteract basic logical principles (Evans, 2008; Evans & Over, 1996; Kahneman & Frederick, 2002; Kahneman & Tversky, 1973). These findings suggest that when people are presented with problems that cue both an answer based on logical rules and, at the same time, a compelling heuristic answer, they often tend to overlook the logical principles and provide biased responses based on heuristics (Kahneman, 2011).

Dual process theories offer an elegant explanation for this bias phenomenon (Evans, 2008; Kahneman, 2011). Traditionally, these theories posit that to override biased, heuristic responses and take logical principles into account, people typically need to engage in effortful deliberation (Evans, 2002, 2008; Evans & Over, 1996; Kahneman, 2011; Stanovich & West, 2000). However, human reasoners are cognitive misers who prefer not to spend extra time and resources once they have already arrived at a fast, intuitive response (Evans & Stanovich, 2013; Kahneman, 2011). As a result, they often stick to their intuitive decisions, even when these violate logical principles. Consequently, they remain biased. Only the few reasoners who have the necessary resources and motivation to engage in deliberation and overcome the biased intuitive response will manage to provide

answers that are based on logic (Evans & Stanovich, 2013; Stanovich & West, 2000).

It is important to emphasize that dual process theories do not support that intuitive thinking always leads to biased answers or that effortful deliberation will guarantee a logical response. On the contrary, dual-process theorists have opposed such simplifications (Evans, 2011; Evans & Stanovich, 2013). For instance, it is universally acknowledged that educated adults can accurately solve the problem "2 + 2" without engaging in deliberation or that even after deep reflection the average reasoner will not manage to solve a very complex mathematical problem concerning, say, nuclear physics equations. Instead, dual process theories of logical reasoning focus on specific scenarios where intuitive and deliberate processing of a problem are assumed to yield conflicting responses (Frederick, 2005). Examples of these scenarios are presented in the Cognitive Reflection Test (CRT), a group of problems whose solution can be easily computed, but which also cue compelling heuristic responses that contradict logical norms (Frederick, 2005). The most famous example of the CRT is the bat-and-ball problem: "A bat and a ball together cost $1.10. The bat costs $1 more than the ball. How much does the ball cost?". When faced with this problem, most reasoners promptly respond that the ball costs 10 cents. However, after some reflection and by solving the equation "X+Y=1.10, Y=1+X, Solve for X", it becomes clear that the ball in fact costs 5 cents and the bat, at $1 more, costs $1.05. These heuristics-and-biases tasks, such as the bat-and-ball problem and the Linda problem presented above, are designed to systematically create response conflict between heuristic and logical responses. It is within this context that overcoming biased heuristic responses is believed to necessitate effortful deliberation.

However, contrary to the assumptions of traditional dual process models, recent studies in the logical reasoning field have found that responses that were previously thought to require deliberation, can also be processed intuitively (De Neys & Pennycook, 2019). More precisely, even when deliberation is "knocked out" with experimental constraint manipulations, the alleged deliberate, logical response is still observed. Direct evidence in support of this claim comes from studies that adopted new experimental paradigms and, most notably, the two-response paradigm (De Neys & Pennycook, 2019; Thompson et al., 2011). In this paradigm participants are typically instructed to give two consecutive responses to heuristics-and-biases problems. Initially, they are asked to provide the first

response that comes to their mind as quickly as possible. Immediately afterward, they are shown the same problem again and are asked to take all the time they want to reflect on it before providing their final answer. To make sure that the initial response is provided intuitively, without the involvement of deliberation, some studies impose time pressure and/or cognitive load constraints during the initial stage of the paradigm (Bago & De Neys, 2017; Newman et al., 2017). The rationale behind this is that since deliberation requires time and cognitive resources to operate, by restricting both, people are maximally forced to respond intuitively (Bago & De Neys, 2017). Consequently, this paradigm allows researchers to separately examine the nature of more intuitive and deliberate responses (Bago & De Neys, 2017, 2020; Raoelison et al., 2020; Thompson et al., 2011).

According to traditional dual process models, the initial intuitive responses in the two-response paradigm would be expected to be biased, since reasoners would be expected to be influenced by the problem's heuristic cues. However, contrary to traditional dual process assumptions, two-response findings show that when people manage to provide a logical response in the final stage, after deliberating, they have frequently arrived at the same response already in the initial, intuitive stage (e.g., Bago & De Neys, 2017, 2019a). This suggests that in heuristics-and-biases tasks, deliberation is not always necessary to override intuitive responses and achieve logical responding, as intuitive responses might already be logical. These findings challenge the conventional views of intuition and deliberation as proposed by dual process theories, and show that the alleged deliberate response can also be intuitively cued. Interestingly, similar results have also been found beyond logical reasoning, in the domains of moral (Bago & De Neys, 2019b; Vega et al., 2021) and prosocial (Bago et al., 2021; Kessler et al., 2017) decision-making. In these domains there is also evidence that the alleged deliberate response (i.e., utilitarian moral decision or selfish prosocial choice) can often be generated with mere intuitive processing.

Further support for the intuitive processing of logical norms comes from studies that used the conflict detection paradigm (De Neys & Pennycook, 2019). This paradigm focuses on the cases where reasoners remain biased; when they provide responses that conflict with the problem's logical principles. Studies employing this paradigm typically contrast standard problems (referred to as "conflict problems"), where intuitive and deliberate processing cue conflicting

responses, with control problems (referred to as "no-conflict problems"), where both intuitive and deliberate processing are expected to generate the same response and no conflict is created between the two. For instance, the control, no-conflict version of the introductory Linda problem, would have the following response options "a. Linda is active in the feminist movement; b. Linda is active in the feminist movement and is a bank teller". Here, option a aligns both with the stereotypical description of Linda and with probabilistic principles, since it refers to a single event and not the conjunction. Therefore, even when responding intuitively, the majority of people would typically choose option a as their response.

Studies using the conflict detection paradigm have shown that biased reasoners typically show sensitivity to the fact that their responses conflict with competing logical principles. For example, they tend to report lower response confidence and longer reaction times when solving conflict problems compared to their control, no-conflict versions (e.g., Białek & De Neys, 2016; Frey et al., 2018; Gangemi et al., 2015; Mata, 2020; Šrol & De Neys, 2021; Vartanian et al., 2018; see De Neys, 2017 for a review, but also Travers et al., 2016, or Mata et al., 2017, for negative findings). Since the only difference between the two versions is the conflict that is created between logical and heuristic cues, these findings suggest that people process logical cues even when they provide heuristic responses. Put differently, if biased reasoners were completely ignoring the underlying logical cues, their performance would remain the same in both conflict and no-conflict versions of the problem.

Critically, this uncertainty about the initial response, which is also referred to as conflict detection, persists even when participants respond to the problems intuitively and deliberation is minimized with experimental load and/or time-pressure manipulations (Johnson et al., 2016; Pennycook et al., 2014; Thompson & Johnson, 2014). More specifically, even when people provide a biased response in the initial, intuitive stage of the two-response paradigm, they typically report decreased response confidence compared to their baseline confidence (e.g., Bago & De Neys, 2017, 2019b; Białek & Neys, 2017; Burič & Konrádová, 2021; Burič & Šrol, 2020). This indicates that conflict sensitivity operates rather automatically and provides further evidence that logical principles can be processed intuitively.

Moreover, conflict detection is also considered to be a mechanism that influences response change (De Neys, 2012; Pennycook et al., 2015; Purcell et

al., 2023; Thompson et al., 2011). Findings from the two-response paradigm have shown that people who experience more conflict in their initial intuitive responses, tend to be more likely revise their answers during the final deliberate stage (Bago & De Neys, 2017, 2020; Thompson & Johnson, 2014).

Based on the above findings, it becomes clear that although traditional dual process theories support that responses based on logical norms can be typically cued only after deliberation, there is lack of robust empirical evidence to support this claim (De Neys, 2022). To recap, two main findings contradict this assumption: First, responses that are traditionally believed to arise after deliberation are also provided intuitively (Bago & De Neys, 2017, 2019a; Newman et al., 2017; Raoelison & De Neys, 2019). Second, even when reasoners provide biased responses, they still show sensitivity to the problem's logical cues, and this sensitivity often operates intuitively (Bago & De Neys, 2017; Burič & Šrol, 2020; Mata, 2020; Pennycook et al., 2014; Thompson & Johnson, 2014; but see also Mata et al., 2014, and Mata & Ferreira, 2018, for negative findings). Hence, it appears that intuitive processing does not only cue heuristic responses, but also logical ones.

However, this reconceptualization of intuitive and deliberate processing does not imply that deliberation never generates logical responses or that it is never needed to correct our intuitions. On the contrary, there is evidence that when people solve heuristics-and-biases tasks, they provide slightly more logical responses after deliberating and in studies using the two-response paradigm deliberation is sometimes required to override biased intuitive responses (e.g., Bago & De Neys, 2017). The key point here is that the corrective deliberate pattern is not as frequent as previously assumed and that, more often than not, the alleged deliberate response is generated intuitively.

To account for these new findings, scholars have introduced an updated dual process model, sometimes referred to as Dual Process Theory 2.0 (De Neys, 2018). This model asserts that the response that has traditionally been considered to be cued by deliberation, can also be cued intuitively. More specifically, it proposes that when a reasoner intuitively processes a "bias" problem, they will generate multiple types of intuitions which will compete with each other (De Neys, 2022). Two primary intuitions come into play: one that cues a heuristic response (also referred to as "heuristic intuition") and one that cues a logical response (also referred to as "logical intuition"). "Heuristic intuitions" are often based on stored

semantic associations and contradict logical rules, while "logical intuitions" stem from an automatized knowledge of mathematical and probabilistic principles (De Neys, 2012, 2022; Evans, 2019; Stanovich, 2018). The stronger, most activated intuition will eventually become the selected intuitive response. When the activation levels of competing intuitions are similar in strength, the reasoner will feel more uncertain or conflicted about their response. This uncertainty might prompt further deliberation which will in turn either confirm or change the intuitive choice (Pennycook et al., 2015). If, however, one intuition clearly dominates over another in strength, the reasoner will feel certain about their intuitive choice and the dominant intuition will lead to a response without further deliberation.

The idea is that "logical intuitions" stem from a learning and practice process (Bago & De Neys, 2019a; De Neys, 2012; Evans, 2019; Raoelison et al., 2021; Stanovich, 2018). More specifically, many of the logico-mathematical principles used in heuristics-and-biases tasks are taught during schooling. As a result, people who are exposed to these principles over the years develop the ability to practice them to automaticity (De Neys, 2012; Purcell et al., 2021; Stanovich, 2018). This parallels the way experts find complex problems easier to solve compared to novices; their extensive experience enables them to immediately recognize familiar patterns. In the same way, lay people can develop a familiarity with fundamental mathematical and probabilistic concepts.

As already mentioned, the evidence for this novel characterization of intuitive reasoning primarily comes from classic heuristics-and-biases tasks, like the bat-and-ball problem (e.g., Raoelison & De Neys, 2019). Yet, it is important to explore whether the current findings extend beyond logical reasoning tasks, a concern recently raised by scholars in the field (March et al., 2023). Indeed, although "fast-and-slow" dual process models may have primarily gained popularity for explaining findings in the heuristics-and-biases field, their fundamental ideas have been applied in various domains (Melnikoff & Bargh, 2018). If the new, revised, core assumptions are to provide anything akin to a general theory of cognition (Reber & Allen, 2022), it is evidently crucial to test the generalizability of the central findings across different fields. Extending the current findings beyond logical reasoning also holds methodological implications. More specifically, many cognitive tasks have been used as predictors of deliberation abilities (Frederick, 2005; Sirota et al., 2021). However, if these tasks yield evidence suggesting that the alleged deliberate response can also be prompted

intuitively, this would undermine their use as predictors. Finally, exploring intuitive logic beyond heuristics-and-biases tasks can offer insights into the design of interventions and policies that are aimed to mitigate biases in each domain (Milkman et al., 2009).

The present thesis addresses these issues and seeks to better understand the interplay of intuition and deliberation, across two key dimensions. The main axis, Axis 1, aims to test whether the evidence for correct intuitive responding (Chapters 1, 2, 3) and conflict sensitivity (Chapters 1 & 2) extends beyond classic heuristics-and-biases tasks. To achieve this, three broad domains are explored in which correct responding has been traditionally believed to result from effortful deliberation: decision making under risk, high-level semantic processing tasks, and low-level cognitive control tasks. A supplementary Axis 2 of this thesis focuses on the stability of responses to classic heuristics-and-biases tasks over time and the long-term impact of conflict sensitivity on answer change (Chapter 4). It thereby aims to generalize the earlier conflict detection findings in the reasoning field across a more extended temporal window.

Chapter summary

In **Chapter 1**, I focus on decision-making under risk. Although this field is central to prospect theory (Kahneman & Tversky, 1979), which lies at the heart of heuristics-and-biases research, there is no systematic empirical evidence to show whether deliberation is necessary to take expected-value maximizing risks into account (Mechera-Ostrovsky et al., 2022). When people take risks, they are often susceptible to the loss aversion bias, which makes them overestimate the negative impact of losses compared to the prospect of comparable potential gains (Kahneman & Tversky, 1979). According to dual process theories, deliberation is needed to overcome this bias and take the expected value of a gamble into account (Slovic et al., 2005). However, there is limited empirical evidence to support this idea. To directly test this, I presented participants with two-outcome positive expected-value gambles using the two response paradigm (Thompson et al., 2011), and they had to choose between a safe loss averse option and a risky expected-value-maximizing option. The findings show that in most of their choices people remained loss averse, both after mere intuitive processing and after deliberating. However, when they opted for the expected-value-maximizing choice

after deliberation, they had often arrived to this response already in the initial, intuitive stage. Additionally, even when people were loss averse, they often detected that their response conflicted with the problem's expected value principles, as shown by decreased confidence. These findings show that deliberation is not the primary route for expected-value-based responding in risky decision making. Most of the time when people manage to take expected-value maximizing risks, they do so using mere intuitive processing.

After showing evidence for sound intuitive responding and conflict sensitivity in decisions under risk, in **Chapter 2,** I investigate the nature of correct responding in semantic high-level language processing tasks. I specifically focus on semantic illusions, which are memory retrieval tasks that have a correct solution, but at the same time cue an incorrect, heuristic response (Erickson & Mattson, 1981). For instance, consider the following question: "What is the name of the kimono-clad courtesans who entertain Chinese men?". When faced with this question, most people would answer "Geisha", failing to notice that Geishas are part of the Japanese and not Chinese culture. Dual process theories attribute this bias to a failure to engage in deliberate processing (Koriat, 2017). According to this view, slow deliberate processing is required to detect anomalies in distorted sentences and correct superficial intuitive responses. Nevertheless, the available evidence does not tell us whether deliberation is always necessary to detect the anomalies in sentences. To test this hypothesis, I presented participants with semantic illusions using the two-response paradigm (Thompson et al., 2011). The results indicate that, more often than not, people need to engage in slow, deliberate processing to overcome the illusion. However, they still manage to provide correct intuitive responses in a non-negligible amount of cases. Additionally, even when people fall for the illusion, they are sensitive to the fact that their response is not fully warranted, as measured by decreased confidence. The findings of Chapter 2 reveal that correct intuitive responding is not limited to tasks that require a basic understanding of mathematical concepts, but that it is also present in high-level language comprehension tasks.

Building on the previous results that showed evidence for correct intuitive responding in both reasoning and non-reasoning high-level semantic tasks, in **Chapter 3**, I explore the nature of correct responding in low-level cognitive control tasks. These tasks have been used to directly tap into lower-level control processes, rather than higher order functioning such as reasoning (Botvinick et

al., 2001; Diamond, 2013). They typically require individuals to inhibit a task-irrelevant but potent response and select a less dominant one. It is believed that resisting the tempting, automatic responses demands controlled, effortful processing (Botvinick et al., 2001). In other words, cognitive control is assumed to have a corrective role; fast, incorrect responses are automatically generated and are then corrected by slower, controlled processes (Botvinick et al., 2001). This pattern is similar to the one proposed by dual process theories of reasoning. To empirically test the corrective pattern of deliberate control in low-level tasks, I presented participants with two of the most commonly used cognitive control tasks in a two-response format: the Stroop task (Stroop, 1935) and the Flanker task (Eriksen & Eriksen, 1974). The aim was to determine whether individuals could provide accurate responses when time and cognitive resources were limited. The results showed that good performance in these tasks is driven by accurate intuitive processing rather than by slow controlled correction of erroneous automatic responses. So, both in the Stroop task and the Flanker task correct responses are generated when deliberate control is constrained. As a second step, I explored the link between intuitive performance at the Stroop task and at heuristics-and-biases tasks (Abreu-Mendoza et al., 2020; De Neys et al., 2011; Handley et al., 2004). Results point to an—at best—weak correlation, for which I explore several theoretical explanations.

In sum, in the main Axis 1 of this thesis (Chapters 1-3) I found evidence that the alleged deliberate response can often result from mere intuitive processing across the fields of decision making under risk, high-level non-reasoning tasks, and low-level cognitive control tasks. Additionally, in Chapters 1 and 2, I showed that even when participants were providing a biased, heuristic response, they could detect that their answer conflicted with some underlying elements of the problem. Critically, this sensitivity to the correct response was not only found for deliberate responses, but also for intuitive ones, suggesting that the conflict detection mechanism can operate automatically in these domains.

In **Chapter 4**, I focus more closely on conflict sensitivity and examine its long-term impact on the stability of performance in heuristics-and-biases tasks. As previously mentioned, earlier two-response studies have shown that conflict detection predicts answer change on an intra trial level; the more conflicted a participant is about their response in the initial stage, the more likely they are to change it in the final stage when they are allowed to deliberate (Bago & De Neys,

2017; Thompson & Johnson, 2014). In Chapter 4, I test whether conflict sensitivity can also have a long-term impact on answer change. To do so, I asked participants to solve the same heuristics-and-biases tasks twice in two test sessions, two weeks apart. I used the two-response paradigm to test the stability of both initial intuitive and final deliberate responses. First, the results showed that participants' responses to heuristics-and-biases tasks are highly stable over time; participants rarely changed their intuitive and deliberate responses after they were first tested (see also Stango & Zinman, 2020, for similar findings). However, despite the high stability, there was still some variability in intuitive and deliberate responses after two weeks. Critically, this long-term variability was not entirely random, but could be predicted by conflict detection. The more conflicted people were about their intuitive response to a problem during the first test session, the more likely to change their (intuitive and deliberate) response to the same problem two weeks later.

In sum, supplementary Axis 2 demonstrates, in one of the rare direct tests of the stability of heuristics-and-biases tasks (see also Bialek & Pennycook, 2018; Stango & Zinman, 2020), that individual biases remain highly stable over time. Most importantly, it shows that intuitive conflict detection can predict variability in intuitive and deliberate responses over time. This way, it confirms that the conflict detection and answer change coupling can be generalized over a longer time window, which points to an interesting new application of the two-response paradigm.

After the four empirical chapters, I also present a Discussion chapter in which I present a summary of the findings, their implications, reflections for future research and concluding remarks.

References

Abreu-Mendoza, R. A., Coulanges, L., Ali, K., Powell, A. B., & Rosenberg-Lee, M. (2020). Children's discrete proportional reasoning is related to inhibitory control and enhanced by priming continuous representations. *Journal of Experimental Child Psychology*, *199*, 104931. https://doi.org/10.1016/j.jecp.2020.104931

Bago, B., Bonnefon, J.-F., & De Neys, W. (2021). Intuition rather than deliberation determines selfish and prosocial choices. *Journal of Experimental Psychology: General*, *150*(6), 1081–1094. https://doi.org/10.1037/xge0000968

Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition*, *158*, 90–109. https://doi.org/10.1016/j.cognition.2016.10.014

Bago, B., & De Neys, W. (2019a). The Smart System 1: Evidence for the intuitive nature of correct responding on the bat-and-ball problem. *Thinking & Reasoning*, *25*(3), 257–299. https://doi.org/10.1080/13546783.2018.1507949

Bago, B., & De Neys, W. (2019b). The intuitive greater good: Testing the corrective dual process model of moral cognition. *Journal of Experimental Psychology: General*, *148*(10), 1782–1801. https://doi.org/10.1037/xge0000533

Bago, B., & De Neys, W. (2020). Advancing the specification of dual process models of higher cognition: A critical test of the hybrid model view. *Thinking & Reasoning*, *26*(1), 1–30. https://doi.org/10.1080/13546783.2018.1552194

Beaulac, G., & Kenyon, T. (2018). The Scope of Debiasing in the Classroom. *Topoi*, *37*(1), 93–102. https://doi.org/10.1007/s11245-016-9398-8

Białek, M., & De Neys, W. (2016). Conflict detection during moral decision-making: Evidence for deontic reasoners' utilitarian sensitivity. *Journal of Cognitive Psychology*, *28*(5), 631–639. https://doi.org/10.1080/20445911.2016.1156118

Białek, M., & Neys, W. D. (2017). Dual processes and moral conflict: Evidence for deontological reasoners' intuitive utilitarian sensitivity. *Judgment and Decision Making*, *12*(2), 148–167. https://doi.org/10.1017/S1930297500005696

Bialek, M., & Pennycook, G. (2018). The cognitive reflection test is robust to multiple exposures. *Behavior Research Methods*, *50*(5), 1953–1959. https://doi.org/10.3758/s13428-017-0963-x

Bonnefon, J.-F., & Rahwan, I. (2020). Machine Thinking, Fast and Slow. *Trends in Cognitive Sciences*, S1364661320302229. https://doi.org/10.1016/j.tics.2020.09.007

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624–652. https://doi.org/10.1037/0033-295X.108.3.624

Burič, R., & Konrádová, ľubica. (2021). Mindware Instantiation as a Predictor of Logical Intuitions in Cognitive Reflection Test. *Studia Psychologica*, *63*(2), 114–128. https://doi.org/10.31577/sp.2021.02.822

Burič, R., & Šrol, J. (2020). Individual differences in logical intuitions on reasoning problems presented under two-response paradigm. *Journal of Cognitive Psychology*, *32*(4), 460–477. https://doi.org/10.1080/20445911.2020.1766472

De Neys, W. (2012). Bias and Conflict: A Case for Logical Intuitions. *Perspectives on Psychological Science*, *7*(1), 28–38. https://doi.org/10.1177/1745691611429354

De Neys, W. (2017). Bias, conflict, and fast Logic. In W. De Neys (Ed.), *Dual Process Theory 2.0* (pp. 47-65). Routledge.
https://doi.org/10.4324/9781315204550-4

De Neys, W. (Ed.). (2018). *Dual Process Theory 2.0*. Routledge. https://doi.org/10.4324/9781315204550

De Neys, W. (2022). Advancing theorizing about fast-and-slow thinking. *Behavioral and Brain Sciences*, 1–68. https://doi.org/10.1017/S0140525X2200142X

De Neys, W., Novitskiy, N., Geeraerts, L., Ramautar, J., & Wagemans, J. (2011). Cognitive Control and Individual Differences in Economic Ultimatum Decision-Making. *PLOS ONE*, *6*(11), e27107. https://doi.org/10.1371/journal.pone.0027107

De Neys, W., & Pennycook, G. (2019). Logic, Fast and Slow: Advances in Dual-Process Theorizing. *Current Directions in Psychological Science*, *28*(5), 503–509. https://doi.org/10.1177/0963721419855658

Diamond, A. (2013). Executive Functions. *Annual Review of Psychology*, *64*(1), 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Epstein, S. (1994). Integration of the Cognitive and the Psychodynamic Unconscious. *American Psychologist*, *49*(8), 709–724. https://doi.org/10.1037/0003-066X.49.8.709

Erickson, T. D., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior*, *20*(5), 540–551. https://doi.org/10.1016/S0022-5371(81)90165-1

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143–149. https://doi.org/10.3758/BF03203267

Evans, J. St. B. T. (2002). Logic and human reasoning: An assessment of the deduction paradigm. *Psychological Bulletin*, *128*(6), 978–996. https://doi.org/10.1037/0033-2909.128.6.978

Evans, J. St. B. T. (2008). Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition. *Annual Review of Psychology*, *59*(1), 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Evans, J. St. B. T. (2011). Dual-process theories of reasoning: Contemporary issues and developmental applications. *Developmental Review*, *31*(2), 86–102. https://doi.org/10.1016/j.dr.2011.07.007

Evans, J. St. B. T. (2019). Reflections on reflection: The nature and function of type 2 processes in dual-process theories of reasoning. *Thinking & Reasoning*, *25*(4), 383–415. https://doi.org/10.1080/13546783.2019.1623071

Evans, J. St. B. T., & Over, D. E. (1996). Rationality in the selection task: Epistemic utility versus uncertainty reduction. *Psychological Review*, *103*(2), 356–363. https://doi.org/10.1037/0033-295X.103.2.356

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Frankish, K., & Evans, J. St. B. T. (2009). The duality of mind: An historical perspective. In J. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 1–30). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199230167.003.0001

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Frey, D., Johnson, E. D., & De Neys, W. (2018). Individual differences in conflict detection during reasoning. *Quarterly Journal of Experimental Psychology*, *71*(5), 1188–1208. https://doi.org/10.1080/17470218.2017.1313283

Gangemi, A., Bourgeois-Gironde, S., & Mancini, F. (2015). Feelings of error in reasoning—In search of a phenomenon. *Thinking & Reasoning*, *21*(4), 383–396. https://doi.org/10.1080/13546783.2014.980755

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences*, *6*(12), 517–523. https://doi.org/10.1016/S1364-6613(02)02011-9

Handley, S. J., Capon, A., Beveridge, M., Dennis, I., & Evans, J. S. B. (2004). : Working memory, inhibitory control and the development of children's reasoning. *Thinking & Reasoning*, *10*(2), 175–195. https://doi.org/10.1080/13546780442000051

Johnson, E. D., Tubau, E., & De Neys, W. (2016). The Doubting System 1: Evidence for automatic substitution sensitivity. *Acta Psychologica*, *164*, 56–64. https://doi.org/10.1016/j.actpsy.2015.12.008

Kahneman, D. (2000). A psychological point of view: Violations of rational rules as a diagnostic of mental processes. *Behavioral and Brain Sciences*, *23*(5), 681–683. https://doi.org/10.1017/S0140525X00403432

Kahneman, D. (2011). *Thinking, fast and slow* (p. 499). Farrar, Straus and Giroux.

Kahneman, D., & Frederick, S. (2002). Representativeness Revisited: Attribute Substitution in Intuitive Judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and Biases* (1st ed., pp. 49–81). Cambridge University Press. https://doi.org/10.1017/CBO9780511808098.004

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454. https://doi.org/10.1016/0010-0285(72)90016-3

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237–251. https://doi.org/10.1037/h0034747

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263–291. https://doi.org/10.2307/1914185

Kessler, J. B., Kivimaki, H., & Niederle, M. (2017). Thinking fast and slow: generosity over time. Retrieved https://users.nber.org/~kesslerj/papers/KesslerKivimakiNiederle_GenerosityOverTime.pdf

Koriat, A. (2017). Can People Identify "Deceptive" or "Misleading" Items that Tend to Produce Mostly Wrong Answers? *Journal of Behavioral Decision Making*, *30*(5), 1066–1077. https://doi.org/10.1002/bdm.2024

March, D. S., Olson, M. A., & Gaertner, L. (2023). Automatic threat processing shows evidence of exclusivity. *Behavioral and Brain Sciences*, *46*, e131. https://doi.org/10.1017/S0140525X22002928

Mata, A. (2020). Conflict detection and social perception: Bringing meta-reasoning and social cognition together. *Thinking & Reasoning*, *26*(1), 140–149. https://doi.org/10.1080/13546783.2019.1611664

Mata, A., & Ferreira, M. B. (2018). Response: Commentary: Seeing the conflict: an attentional account of reasoning errors. *Frontiers in Psychology*, *9*. https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00024

Mata, A., Ferreira, M. B., Voss, A., & Kollei, T. (2017). Seeing the conflict: An attentional account of reasoning errors. *Psychonomic Bulletin & Review*, *24*(6), 1980–1986. https://doi.org/10.3758/s13423-017-1234-7

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. https://doi.org/10.1016/j.cognition.2014.07.011

Mechera-Ostrovsky, T., Heinke, S., Andraszewicz, S., & Rieskamp, J. (2022). Cognitive abilities affect decision errors but not risk preferences: A meta-analysis. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-021-02053-1

Melnikoff, D. E., & Bargh, J. A. (2018). The Mythical Number Two. *Trends in Cognitive Sciences*, *22*(4), 280–293. https://doi.org/10.1016/j.tics.2018.02.001

Milkman, K. L., Chugh, D., & Bazerman, M. H. (2009). How Can Decision Making Be Improved? *Perspectives on Psychological Science*, *4*(4), 379–383. https://doi.org/10.1111/j.1745-6924.2009.01142.x

Newman, I. R., Gibb, M., & Thompson, V. A. (2017). Rule-based reasoning is fast and belief-based reasoning can be slow: Challenging current explanations of belief-bias and base-rate neglect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(7), 1154–1170. https://doi.org/10.1037/xlm0000372

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2014). Cognitive style and religiosity: The role of conflict detection. *Memory & Cognition*, *42*(1), 1–10. https://doi.org/10.3758/s13421-013-0340-7

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. https://doi.org/10.1016/j.cogpsych.2015.05.001

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

Purcell, Z. A., Wastell, C. A., & Sweller, N. (2021). Domain-specific experience and dual-process thinking. *Thinking & Reasoning*, *27*(2), 239–267. https://doi.org/10.1080/13546783.2020.1793813

Purcell, Z. A., Wastell, C. A., & Sweller, N. (2023). Eye movements reveal that low confidence precedes deliberation. *Quarterly Journal of Experimental Psychology*, *76*(7), 1539–1546. https://doi.org/10.1177/17470218221126505

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, *489*(7416), Article 7416. https://doi.org/10.1038/nature11467

Raoelison, M., Boissin, E., Borst, G., & De Neys, W. (2021). From slow to fast logic: The development of logical intuitions. *Thinking & Reasoning*, *27*(4), 599–622. https://doi.org/10.1080/13546783.2021.1885488

Raoelison, M., & De Neys, W. (2019). Do we de-bias ourselves?: The impact of repeated presentation on the bat-and-ball problem. *Judgment and Decision Making*, *14*(2), 170–178. https://doi.org/10.1017/S1930297500003405

Raoelison, M., Thompson, V. A., & De Neys, W. (2020). The smart intuitor: Cognitive capacity predicts intuitive rather than deliberate thinking. *Cognition*, *204*, 104381. https://doi.org/10.1016/j.cognition.2020.104381

Reber, A. S., & Allen, R. (2022). *The Cognitive Unconscious: The First Half Century*. Oxford University Press.

Sirota, M., Dewberry, C., Juanchich, M., Valuš, L., & Marshall, A. C. (2021). Measuring cognitive reflection without maths: Development and validation of the verbal cognitive reflection test. *Journal of Behavioral Decision Making*, *34*(3), 322–343. https://doi.org/10.1002/bdm.2213

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3–22. https://doi.org/10.1037/0033-2909.119.1.3

Slovic, P., Peters, E., Finucane, M. L., & MacGregor, D. G. (2005). Affect, risk, and decision making. *Health Psychology*, *24*(4, Suppl), S35–S40. https://doi.org/10.1037/0278-6133.24.4.S35

Šrol, J., & De Neys, W. (2021). Predicting individual differences in conflict detection and bias susceptibility during reasoning. *Thinking & Reasoning*, *27*(1), 38–68. https://doi.org/10.1080/13546783.2019.1708793

Stango, V., & Zinman, J. (2020). *Behavioral Biases are Temporally Stable* (w27860; p. w27860). National Bureau of Economic Research. https://doi.org/10.3386/w27860

Stanovich, K. E. (2018). Miserliness in human cognition: The interaction of detection, override and mindware. *Thinking & Reasoning*, *24*(4), 423–444. https://doi.org/10.1080/13546783.2018.1459314

Stanovich, K. E., & West, R. F. (2000). Advancing the rationality debate. *Behavioral and Brain Sciences*, *23*(5), 701–717. https://doi.org/10.1017/S0140525X00623439

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662. https://doi.org/10.1037/h0054651

Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning*, *20*(2), 215–244. https://doi.org/10.1080/13546783.2013.869763

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*(3), 107–140. https://doi.org/10.1016/j.cogpsych.2011.06.001

Travers, E., Rolison, J. J., & Feeney, A. (2016). The time course of conflict on the Cognitive Reflection Test. *Cognition*, *150*, 109–118. https://doi.org/10.1016/j.cognition.2016.01.015

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315. https://doi.org/10.1037/0033-295X.90.4.293

Vartanian, O., Beatty, E. L., Smith, I., Blackler, K., Lam, Q., Forbes, S., & De Neys, W. (2018). The Reflective Mind: Examining Individual Differences in Susceptibility to Base Rate Neglect with fMRI. *Journal of Cognitive Neuroscience*, *30*(7), 1011–1022. https://doi.org/10.1162/jocn_a_01264

Vega, S., Mata, A., Ferreira, M. B., & Vaz, A. R. (2021). Metacognition in moral decisions: Judgment extremity and feeling of rightness in moral intuitions. *Thinking & Reasoning*, *27*(1), 124–141. https://doi.org/10.1080/13546783.2020.1741448

Wason, P. C., & Evans, J. ST. B. T. (1974). Dual processes in reasoning? *Cognition*, *3*(2), 141–154. https://doi.org/10.1016/0010-0277(74)90017-1

Chapter 1

Fast and slow decisions under risk

Voudouri, A., Bialek, M., & De Neys, W. (under review). Fast & slow decisions under risk: Intuition rather than deliberation drives advantageous choices. *Cognition*.

Supplementary material for this chapter can be found in Supplementary material for Chapter 1.

Abstract

Would you take a gamble with a 10% chance to gain $100 and a 90% chance to lose $10? Even though this gamble has a positive expected value, most people would avoid taking it given the high chance of losing money. Popular "fast-and-slow" dual process theories of risky decision making assume that to take expected value into account and avoid a loss aversion bias, people need to deliberate. In this paper we directly test whether reasoners can also consider expected value benefit intuitively, in the absence of deliberation. To do so, we presented participants with bets and lotteries in which they could choose between a risky expected-value-based choice and a safe loss averse option. We used a two-response paradigm where participants made two choices in every trial: an initial intuitive choice under time-pressure and cognitive load and a final choice without constraints where they could freely deliberate. Results showed that in most trials participants were loss averse, both in the intuitive and deliberate stages. However, when people opted for the expected-value-based choice after deliberating, they had predominantly already arrived at this choice intuitively. Additionally, loss averse participants often showed an intuitive sensitivity to expected value (as reflected in decreased confidence). Overall, these results suggest that deliberation is not the primary route for expected-value-based responding in risky decision making. Risky decisions may be better conceptualized as an interplay between different types of "fast" intuitions rather than between two different types of "fast" and "slow" thinking per se.

Introduction

Imagine that you are faced with a gamble that gives you a 10% chance to gain $100 and a 90% chance to lose $10. Would you take it or avoid it? The outcomes and their probabilities imply that if you were to play this gamble ten times you would win once ($100) and lose nine times (−$90). At the end, you would have gained $10. So, based on the expected value, it is in one's best financial interest to take the above gamble. However, when faced with such gambles most people would avoid taking them given the high chance of losing money. In fact, people often make biased decisions when it comes to evaluating risk, as they overestimate the impact of losses compared to the prospect of comparable potential gains (Kahneman & Tversky, 1979). This bias–often referred to as "loss aversion" (Kahneman & Tversky, 1979)–has been widely studied and applied in a range of real-world contexts (Camerer, 2005).

A popular explanation for the loss aversion bias has been put forward by dual process theories. These theories support that reasoning involves two types of processes; a fast, effortless, intuitive process ("System 1") and a slower, effortful, deliberate one ("System 2"; e.g. Kahneman, 2003; Kahneman & Frederick, 2002; Slovic et al., 2005). When it comes to evaluating risk, researchers have termed these two ways in which risks are assessed as "risk-as-feelings" and "risk-as-analysis" respectively (Loewenstein et al., 2001; Slovic et al., 2005). Dual process theories of risky choice support that to take the probabilities and outcomes of a gamble into account, people need to engage in effortful, deliberate processing (Slovic et al., 2005). On the contrary, when people process a gamble intuitively they will not be responsive to its probabilities. Instead, they are susceptible to affect and, consequently, loss aversion. Thus, when intuitive processing contradicts a gamble's probabilities and outcomes, it leads to biased decisions, which need to be overridden with effortful deliberation. However, people often act as cognitive misers and tend to minimize mental effort, so they are not likely to engage in deliberate processing once they have already made a choice intuitively (Evans & Stanovich, 2013; Kahneman, 2011; Sirota et al., 2023). This is why when faced with gambles like the above, the majority of people stick to their intuitively cued loss averse choices. Only the most highly skilled and motivated reasoners will manage to deliberate and override them.

Introspectively, the dual process account of risky decision making does not seem unreasonable. When faced with the above gamble, for example, many may instantly feel they want to avoid any possible loss. Making the choice that will maximize our payoffs may seem to require more time and effort. However, there is little direct empirical evidence showing that reasoners who take expected value maximizing risks manage to do so only after deliberating. Studies that tried to manipulate intuitive and deliberate risk taking point to inconclusive results (e.g., Drichoutis & Nayga, 2020). To clarify, the experimental rationale here is that deliberation is assumed to require time and cognitive resources (Kahneman, 2011; Sirota et al., 2021). When people are deprived of these resources by making choices under time-pressure and/or a cognitive load, one would expect to see an increase in loss aversion (i.e., the alleged intuitive response).

In line with the classic dual-process view, some indeed observed that cognitive constraints make people more loss averse (e.g., Deck & Jahedi, 2015; Gerhardt et al., 2016). However, others found that people continue taking advantageous risks (they attain "economic rationality", Drichoutis & Nayga, 2020) even when they are cognitively burdened with a load task (e.g., Drichoutis & Nayga, 2020; Freeman & Muraven, 2010). In addition, a recent meta-analysis found no credible association between cognitive abilities and loss aversion (Mechera-Ostrovsky et al., 2022). Studies using time constraints also present mixed findings. While some found that time pressure makes people more loss averse (Kocher et al., 2013; Zur & Breznitz, 1981), others showed that it increased risk seeking (Dror et al., 1999; Madan et al., 2015). As a response to these inconsistent findings, recent studies have suggested that cognitive load or time pressure do not lead to systematic changes in risk preferences, but rather to changes in choice consistency (e.g., choices might simply become more random, Andersson et al., 2016; Olschewski & Rieskamp, 2021).

In sum, based on the current literature, it is difficult to infer whether cognitive resources are necessary for people to take advantageous risks, or whether such risks can also be taken intuitively. In theory, it is possible that in addition to the deliberate route to profit-maximizing risky choices, there also exists an intuitive route via which the expected value maximizing choice is cued. Put differently, in some cases people's intuitive choices might not be loss averse, but instead they might be based on an intuitive understanding of expected value.

Interestingly, recent evidence from the reasoning field lends some indirect credence to this "expected-value intuitor" account. These studies have shown that in a range of classic "bias" task from the heuristics-and-biases literature, such as base-rate neglect, bat-and-ball (e.g., Burič & Konrádová, 2021; Raoelison & De Neys, 2019), conjunction fallacy (Boissin et al., 2022), or belief-bias syllogism problems (e.g., Bago & De Neys, 2017; Raoelison et al., 2020), logico-probabilistic principles which were traditionally thought to be processed only after deliberation, can also be processed intuitively (e.g., Bago & De Neys, 2017). To demonstrate this, the studies adopted a so-called two-response paradigm (Thompson et al., 2011) in which participants are instructed to provide the first response that comes to mind as quickly as possible, and to then take their time to reflect on the given problem before providing their final response. To be maximally sure that participants do not deliberate during the initial stage, they are forced to give their initial response under time-pressure while performing a concurrent load task which burdens their cognitive resources (Bago & De Neys, 2017). Since deliberation requires time and cognitive resources, by restricting both, possible deliberation is minimized during the initial stage and participants are maximally forced to rely on intuitive processing. Results from these two-response studies showed that when participants manage to give a correct response after deliberation, they have often already arrived at this response in the initial, intuitive stage (e.g., Bago & De Neys, 2019; Burič & Konrádová, 2021; Thompson & Johnson, 2014). Hence, sound reasoners are often good at accurate intuiting, and not necessarily at deliberately correcting their erroneous intuitions.

In the present study we introduce a two-response paradigm to directly investigate the nature of expected-value-based choices in risky decision making. In all experiments participants played risky choice games. On each trial, they had to first make an initial choice as fast as possible (under time pressure and concurrent cognitive load), and immediately after they could take time to deliberate before making their final choice. They also indicated their confidence in their (initial and final) choices. Our main question was whether in the cases where people make expected-value-based choices after deliberation, they can also make such choices intuitively in the initial response stage (i.e., when deliberation is prevented).

A second question that we explore in this paper is whether, in the cases where people make loss averse choices, they are sensitive to the fact that these

choices violate expected value. Dual process theories generally assume that the reason people provide responses that contradict logico-probabilistic principles is because they do not consider these principles (Evans & Stanovich, 2013). However, the reasoning field findings we alluded to above have also shown that even biased responders often show some intuitive sensitivity to the fact that that their response conflicts with competing logical considerations (e.g., Bago & De Neys, 2017; Burič & Šrol, 2020; Stupple & Ball, 2008, Stupple et al., 2011). For example, when reasoners give a biased response in the initial response stage of the two-response paradigm, they typically show increased response doubt (as indexed, for example, by lowered response confidence compared to control problems, for a review see De Neys, 2017). So, as a second step in the present study, we examined whether in those cases where participants provide loss averse initial intuitive choices, they show conflict sensitivity (as measured by decreased confidence). To this end, we also included control trials in our games, in which the conflict between the loss averse and the expected value maximizing choice was removed or reduced (e.g., a gamble that gives you a 90% chance to gain $100 and only a 10% chance to lose $10). If people refrained from random guessing, we expected a strong preference for the expected value maximizing choice on these "no conflict" control problems. On the standard "conflict" problems, the conflict between the expected value maximizing and loss averse options is–in theory–more pronounced. Following the reasoning literature (e.g., De Neys, 2017), we expected that if people who make intuitive loss averse choices do not completely disregard expected value considerations, the conflict should decrease their response confidence.

To test the generality of the findings, in each of our studies participants played two different types of classic risky choice games: the betting game (Keysar et al.,2012) and the lottery game (Holt & Laury, 2002). The betting game consisted of two-outcome positive-expected-value bets (in addition to no-conflict control bets) that could result in either a gain or a loss, and participants chose whether they wanted to accept them or not. The lottery game consisted of lottery pairs which could lead to gains of different magnitudes, and participants had to select their preferred lottery (they were also presented with no-conflict lottery pairs, see Method for details).

We present three studies. In Study 1 we introduce the paradigm and in Study 2 and 3 we test the robustness of the findings.

Study 1

Method

Preregistration and data availability

The study design and hypothesis were preregistered on the Open science Framework (https://osf.io/5gqst). No specific analyses were preregistered. All data and material are also available on the Open Science Framework (https://osf.io/hzjt8/files/osfstorage).

Participants

We recruited our participants online on Prolific Academic (www.prolific.ac). Only native English speakers from Canada, Australia, New Zealand, the United States of America, or the United Kingdom were allowed to take part in the study. Participants were paid £2.10 for their participation (£5 hourly rate). One hundred participants (80 female, mean age = 35.8 years, *SD* = 12.8 years) participated in the study. A total of 32% of participants reported high school as their highest completed educational level, while 66% reported having a postsecondary education degree.

Given that this was, to our knowledge, the first study to test risky decision making in a two-response format, we based our sample size on previous two-response studies in the logical and moral reasoning field (Bago & De Neys, 2017, 2019), in which also approximately 100 participants per condition were tested.

Materials

Betting Game. The betting game was based on Keysar et al.'s (2012) loss aversion task. Participants were presented with a total of 15 bets (5 conflict , 5 no-conflict and 5 filler) that could result in either a gain or a loss. Every bet stated the probability of winning a certain amount of money and the probability of losing a certain amount of money. Participants were asked whether they wanted to take the bet or not. They indicated their choices by clicking on one of two options, labelled as "yes" (take the bet) and "no" (do not take the bet).

Conflict bets. All standard, conflict bets had a positive expected value and a high probability of losing money. Therefore, a conflict was created between

avoiding a potential loss (i.e., by not taking the bet) and taking a risk in order to acquire a bigger potential gain (i.e., by taking the bet). An example of a conflict bet is presented below:

If you take this bet you have:
5%** probability to **WIN €110
95%** probability to **LOSE €5
Do you take the bet?

- *Yes*
- *No*

Note that in the instructions it was stressed that the goal was to make as much profit as possible (see Supplementary Material A for the literal instructions). So, according to an objective outcome calculation and in the absence of a loss aversion bias, participants should always take the (positive- expected-value) bet. This is why, in all conflict items, taking the bet was labelled as the Expected Value (EV) maximizing choice, and not taking the bet was labelled as the loss averse choice.

Each conflict item had a different probability pair. This variation made the task engaging and ensured that our loss aversion results were not dependent on specific probabilities. At the same time, we made sure to keep the probabilities relatively similar between items, so as to avoid differential risk preference (i.e., participants being loss averse with the probabilities of one conflict item, but not with the next one). We varied the probability of winning (P_{win}) from 5% to 25%, and the probability of losing (P_{lose}) from 95% to 75%, in 5% intervals. Thus, the following pairs were created: P_{win} = 5% and P_{lose} = 95%; P_{win} = 10% and P_{lose} = 90%; P_{win} = 15% and P_{lose} = 85%; P_{win} = 20% and P_{lose} = 80%; P_{win} = 25% and P_{lose} = 75%. The values were chosen so that the expected value difference (P_{win} * $\text{Value}_{\text{win}}$ – P_{lose} * $\text{Value}_{\text{lose}}$) was kept as similar as possible between all conflict bets (see Supplementary Material B for all items), to make sure that the items were of equal complexity.

No-conflict bets. The control, no-conflict bets had a positive expected value and a low probability of losing money. These were constructed by reversing the P_{win} and P_{lose} of the conflict items, while keeping the values identical. For example, the no-conflict version of the above conflict item would be:

If you take this bet you have:

95%** probability to **WIN €110

5%** probability to **LOSE €5

Do you take the bet?

- *Yes*
- *No*

Hence, in the no-conflict items participants always had a very high probability of winning a large amount and only a very small probability of losing a small amount. Consequently, the items should not (or only minimally) cue loss aversion and should not (or only minimally) create conflict with expected value considerations. Consequently, if people refrained from random guessing, we expected a strong preference for the expected value maximizing choice in these items: everyone should take the bet and show high confidence. These control trials served as a baseline for our conflict sensitivity analysis. Critically, if loss averse individuals on conflict trials consider the conflicting expected value option, they should experience some minimal doubt and show decreased response confidence compared to control trials. However, if those individuals do not consider expected value, the conflict trials should be a no-brainer for them (i.e., not involve any processing conflict) and they should remain highly confident in not taking the bet.

Filler bets. The filler bets had a negative expected value. Therefore, the most advantageous choice for participants, both in terms of loss aversion and EV calculation, was to not take the bet. An example of a filler bet is presented below:

If you take this bet you have:

50%** probability to **WIN €10

50%** probability to **LOSE €15

Do you take the bet?

- *Yes*
- *No*

These bets allowed us to verify whether participants were using a "take the bet" heuristic. More specifically, some participants may have applied a heuristic strategy where they were always taking the bet during the study. In this case, their responses would align with the EV maximizing choice in every conflict and no-conflict bet, which would distort our findings. In the filler items however, these

participants would have a very low accuracy, which allowed us to detect the strategy.

Each filler item had a different probability pair, but we kept the probabilities as comparable as possible between items. We varied the probability of winning (P_{win}) from 50% to 70%, and the probability of losing (P_{lose}) from 50% to 30%, in 5% intervals. Thus, the following pairs were created: P_{win} = 50% and P_{lose} = 50%; P_{win} = 55% and P_{lose} = 45%; P_{win} = 60% and P_{lose} = 40%; P_{win} = 65% and P_{lose} = 35%; P_{win} = 70% and P_{lose} = 30%. The exact values were chosen so that the expected value difference (P_{win} * $Value_{win}$ – P_{lose} * $Value_{lose}$) was kept as similar as possible between filler bets (see Supplementary Material B for all items).

Lottery Game. The lottery game was a variation of the Holt-Laury lottery choice task (Holt & Laury, 2002). We presented participants with a total of 15 lottery pairs (5 conflict , 5 no-conflict and 5 filler), and they had to choose one lottery from each pair (lottery A or lottery B). Both lotteries (A & B) consisted of a large probability to gain a large amount of money and a small(er) probability to gain a small(er) amount of money. In each lottery pair, lottery A and lottery B had the same large and small probabilities. Participants indicated their lottery choice by clicking on one of two options, labelled as "A" (for lottery A) and "B" (for lottery B).

It is important to note that the loss aversion tested in the lottery game does not involve, strictly speaking, losses since all the lottery pairs have positive expected values. However, even in the absence of losses, people usually experience a risk aversion bias, in that they tend to prefer outcomes with low uncertainty compared to outcomes with high uncertainty but higher potential gains. For consistency with the betting game, we will refer to this risk aversion bias as loss aversion throughout the paper. Given that the lottery game did not involve losses, it allowed us to test the generalisability of our findings beyond the strict loss domain per se.

Conflict lottery pairs. In the standard conflict lottery pairs, one of the lotteries always had the highest expected value in the set, while the other lottery had a lower expected value but the highest guaranteed minimal gain. Therefore, a conflict was created between choosing a lottery with a potentially big but

uncertain gain, and a lottery with a lower but more certain gain. An example of a conflict lottery pair is presented below:

Lottery A	***Lottery B***
70% *probability to* ***win €350***	***70%*** *probability to* ***win €230***
30% *probability to* ***win €10***	***30%*** *probability to* ***win €160***

Which lottery do you choose?

- *A*
- *B*

In the above example, Lottery A has a higher expected value, but only guarantees a gain of €10, while Lottery B has a lower expected value, but guarantees a gain of €160. As mentioned, participants were instructed to try to make as much profit as possible. So, according to an objective outcome calculation and in the absence of a loss aversion bias, participants should always choose the lottery with the highest expected value. That is why, choosing the lottery with the highest expected value was labelled as the EV maximizing choice, while choosing the lottery with the highest guaranteed minimal gain was labelled as the loss averse choice.

Each conflict item had a different probability pair, but we kept the probabilities as similar as possible between items (for an explanation see Conflict bets subsection). The probability of getting the large gain (P_{large}) varied from 60% to 80%, and the probability of getting the smaller gain (P_{small}) varied from 40% to 20%, in 5% intervals. Thus, the following pairs were created: P_{large} = 60% and P_{small} = 40%; P_{large} = 65% and P_{small} = 35%; P_{large} = 70% and P_{small} = 30%. P_{large} = 75% and P_{small} = 25%; P_{large} = 80% and P_{small} = 20%. The values were chosen so that the expected value difference [(P_{A_large} * V_{A_large}) + (P_{A_small} * V_{A_small}) – (P_{B_large} * V_{B_large}) + (P_{B_small} * V_{B_small}) was kept as similar as possible between the conflict lottery pairs (see Supplementary Material B for all items).

No-conflict lottery pairs. In the control, no-conflict lottery pairs one of the lotteries always had both the highest expected value in the set and the highest guaranteed gain. Therefore, no conflict was created; participants were always expected to prefer one of the two lotteries, both in terms of certainty and potential gain. The no-conflict pairs were constructed by reversing the P_{large} and P_{small} in each lottery, while keeping the values identical. For example, the no-conflict equivalent of the above conflict item would be:

Lottery A	***Lottery B***
70% *probability to* ***win €10***	***70%*** *probability to* ***win €160***
30% *probability to* ***win €350***	***30%*** *probability to* ***win €230***

Which lottery do you choose?

- *A*
- *B*

In the above example we would expect participants to choose Lottery B. The no-conflict items also allowed us to verify whether participants were using a "pick the highest value" heuristic. More specifically, some participants may have applied a heuristic strategy where they would always pick the lottery that includes the highest value in the whole set. In this case, their responses would always align with the EV maximizing choice in the conflict and filler (see below) lottery pairs, which would distort our findings. In the no-conflict items however, these participants would have a very low accuracy, which would allow us to detect the strategy.

Filler lottery pairs. The filler lottery pairs were designed so that they did not cue a loss averse response. An example of a filler lottery pair is presented below:

Lottery A	***Lottery B***
90% *probability to* ***win €350***	***90%*** *probability to* ***win €260***
10% *probability to* ***win €310***	***10%*** *probability to* ***win €220***

Which lottery do you choose?

- *A*
- *B*

In the above example it is obvious that Lottery A is the most advantageous lottery. Some of the filler items had the same probability pairs, but the values always varied between items so that participants never saw the exact same filler item twice. When creating the items, the aim was to make the correct choice obvious for any adult with a basic understanding of the problems' probability rules. With this in mind, the following probability pairs were created: P_{large} = 100% and P_{small} = 0%; P_{large} = 50% and P_{small} = 50%; P_{large} = 90% and P_{small} = 10%. The filler items allowed us to test for a guessing confound. If people refrained from random guessing, we expected them to constantly choose the lottery with the highest EV.

Load task. In order to make maximally sure that the initial responses in the two-response paradigm were indeed intuitive, we used a load task to burden participants' cognitive resources during the initial stage. The reasoning behind this manipulation is simple. Dual process theories assume that deliberation requires more cognitive resources than intuition (Evans & Stanovich, 2013). By engaging participants' resources with a secondary load task it will be more likely that their responding to the main task will be intuitive. We used the dot memorization task because it has been previously shown to successfully prevent deliberation in logical reasoning and economic decision making tasks (De Neys et al., 2011; De Neys & Schaeken, 2007; De Neys & Verschueren, 2006; Franssens & De Neys, 2009). Before each bet or lottery pair, participants were presented with a 3x3 grid, in which four grid squares were filled with crosses. They were instructed to memorize the location of the crosses. It was also emphasized that participants first had to try to memorize the crosses and then respond to the bet or lottery pair. After participants responded to the bet or lottery, they were shown four different matrices and they had to choose the correct, to-be-memorized pattern. They then received feedback as to whether their choice was correct or not. The load was applied only during the initial response stage and not during the subsequent final response stage in which participants were allowed to deliberate (see Two-response games).

Procedure

One-response (deliberative-only) pretest. To obtain a baseline performance in both games, we ran a traditional one-response version of our study (without load or deadline). We recruited an independent sample of 50 participants (72% female; mean age = 38.52 years, SD = 14.8) online on Prolific Academic (www.prolific.ac). Only native English speakers from Canada, Australia, New Zealand, the United States of America, or the United Kingdom were allowed to take part in the study. Participants were paid £0.85 for their participation (£5 hourly rate). A total of 38% of the participants reported high school as their highest completed educational level, while 60% reported having a postsecondary education degree.

Following previous studies, we wanted to base the deadline in the initial stage of our main, two-response study on the average response time in the pretest (e.g., Bago & De Neys, 2017, 2020). For this reason, the pretest included the

same number of trials and same stimuli as the main study, but here participants had to provide only one answer to each bet and lottery pair without time restrictions. In the betting game the EV maximizing responses to the conflict bets took more time (7.8 s, SD = 7.9 s) than the loss averse responses (5.1 s, SD = 1.9 s). Thus, we decided to base the initial response deadline on the reaction time of the EV maximizing trials only. The first quartile of these trials was 4.45 s, so we rounded to the nearest decimal and set the deadline for the betting game to 4.5 s. In the lottery game the reaction times were overall longer, which was to be expected as the items were lengthier. In this game, the EV maximizing responses to the conflict lottery pairs took less time (7.8 s, SD = 5.4 s) than the loss averse responses (8.2 s, SD = 4.2 s). Thus, we based the initial response deadline on the overall reaction time across conflict trials. The first quartile of these trials was 5.6 s, so we rounded to the nearest decimal and set the deadline for the lottery game to 5.5 s.

To make sure that participants were indeed under time pressure during the initial stage, we compared the response times on the conflict trials between the one-response pretest and the initial stage of the main two-response study. To do so, we first excluded from the two-response study all trials with incorrect load memorization or a missed deadline (see further). The results revealed that participants responded much faster in the initial response stage of the two-response study, compared to the one-response pretest, both in the betting game ($M_{two\text{-}response}$ = 2.7 s; $M_{one\text{-}response}$ = 5.3 s) and the lottery game ($M_{two\text{-}response}$ = 2.9 s; $M_{one\text{-}response}$ = 7.9 s). Welch Two Sample t-tests indicated that this difference was significant both for the betting game, $t(51.94) = 6.52$, $p < .001$, and for the lottery game, $t(51.80) = 8.40$, $p < .001$.

The one-response pre-test also allowed us to rule out a potential consistency confound in our main two-response study. More specifically, when participants are asked to give two consecutive responses, they might stick to their initial response in the final stage because they want to appear consistent (Thompson et al., 2011). Thereby, the paradigm may underestimate the rate of response change from the initial to the final stage. To check for this consistency confound in our study, we contrasted the proportion of EV maximizing responses in the conflict trials of the one-response pretest and that of the final stage of the main two-response study. If a consistency confound was present, we would find a significantly lower number of EV maximizing responses in the two response study.

However, we found that the percentage of EV maximizing responses was very similar in the one-response pretest and in the final responses of the two-response study, both in the betting game ($M_{two\text{-}response}$ = 13.8%; $M_{one\text{-}response}$ = 14.0%) and the lottery game ($M_{two\text{-}response}$ = 29.8%; $M_{one\text{-}response}$ = 28.8%). Welch Two Sample t-tests indicated that this difference was not significant, neither for the betting game, $t(123.05) = 0.04$, $p = .96$, nor for the lottery game, $t(96.32) = -0.17$, $p = .87$.

Two-response games. The experiment was run online on the Qualtrics (www.qualtrics.com) software server. Participants were instructed at the beginning that the study consisted of two games, a betting game and a lottery game. After the general instructions, participants directly started playing one of the games. First, they were presented with game-specific instructions (see Supplementary Material A for full instructions). They were also told to imagine that they were playing for real money and that the aim was to make as much profit as possible. Afterwards, they were presented with an example bet or lottery pair, depending on the game they were playing. Participants were told that we were first interested in the initial answer that came to their mind and that they would have additional time afterwards to reflect on the problem and provide a final answer.

After the game instructions, participants started a practice session to familiarize themselves with the experimental procedure. First, they were presented with two practice bet/lottery pair trials in which they simply had to respond before the deadline. Next, they solved two practice dot matrix load problems (without concurrent bet/lottery pair). Finally, at the end of the practice, they had to solve the two earlier practice examples under cognitive load and deadline, just as in the main study. Then, they began the experimental trials.

Each trial started with the presentation of a fixation cross for 2 s followed by the load matrix that stayed on the screen for 2 s. Next, the bet or lottery pair appeared. From this point onward, participants had 4.5 s to enter their answer in the betting game and 5.5 s in the lottery game; 1 s before the deadline, the background of the screen turned yellow to warn participants that the time limit was approaching. If they did not provide an answer before the deadline, they were asked to pay attention to provide an answer within the deadline on subsequent trials. If they responded within the deadline, they were asked to rate their

confidence in the correctness of their initial response on a scale from 0 (absolutely not confident) to 100 (absolutely confident). After the confidence question, participants were presented with four matrix patterns and were asked to recall the correct, to-be memorized pattern. They were then given feedback on whether their recall was correct or not. Finally, participants saw the full problem again and were asked to provide their final answer. Next, they were asked to report their confidence in the correctness of their final response.

The colour of the answer options was green during the initial response and blue during the final response phase to visually remind participants which question they were answering. Under the question we also presented a reminder sentence: "Please indicate your very first, intuitive answer!" and "Please give your final answer," respectively, which was coloured like the answer options. At the end of the study, participants completed standard demographic questions and were shown a debriefing message.

The presentation order of the games was randomized. At the end of the first game participants were presented with a short transition message which informed them that they had finished the first game and that they could take a short break before continuing to the second game.

Counterbalancing. Participants were presented with a betting game and a lottery game. Each game was composed of five conflict, five no-conflict and five filler items. For the lottery game two sets of items were created (set A and set B) in which the conflict status of each item was counterbalanced. More specifically, all the conflict items of set A appeared in their no-conflict version in set B, and all the no-conflict items in set A appeared in their conflict version in set B. Half of the participants were presented with set A of problems while the other half was presented with set B. All participants were presented with the same filler items. In the betting game only one set of conflict and no-conflict items was created, since it was not possible to create a second set of items with the same values and, simultaneously, keep the EV difference similar. However, the no-conflict items had slightly different values (−/+ €5-€10) from the conflict ones so although all participants saw the same items, none of them saw the same value pair twice. In sum, in both games, the same content was never presented more than once to a participant and everyone was exposed to items with the same probabilities and

EV differences. This minimized the possibility that mere item differences influence the results.

Exclusion criteria

The trials in which participants failed the load and/or the deadline were excluded from subsequent analyses, since in these trials we could not ensure that deliberation was minimized during the initial stage.

Betting game. Participants failed to answer before the deadline on 2.2% of conflict trials, 1.4% of no-conflict trials, and 3.4% of filler trials. In addition, they failed the load task on 14% of conflict trials, 9.8% of no-conflict trials, and 12.6% of filler trials. Overall, by rejecting the missed deadline and missed load trials we kept 83.8% of conflict trials, 88.8% of no-conflict trials, and 84% of filler trials. On average, each participant contributed 12.8 trials (out of 15 trials, *SD* = 2.0).

To ensure that participants were not using an "always take the bet" heuristic in the betting game (see Filler bets subsection above), following our pre-registration, we ran a control analysis where we excluded participants who had an accuracy lower than 50% both in their initial and final filler trials (n = 12). In this control analysis all of our conclusions remained the same, suggesting that the heuristic did not bias our results. Therefore, in the results section below, we present the intended complete analysis without exclusions. The partial results excluding these participants are reported in Supplementary Material C.

Lottery game. Participants failed to answer before the deadline on 3% of conflict trials, 1.4% of no-conflict trials, and 1.4% of filler trials. In addition, they failed the load task on 20.4% of conflict trials, 8.6% of no-conflict trials, and 16.6% of filler trials. Overall, by rejecting the missed deadline and missed load trials we kept 76.6% of conflict trials, 90% of no-conflict trials, and 82% of filler trials. On average, each participant contributed 12.4 trials (out of 15 trials, *SD* = 2.2).

To ensure that participants were not using a "pick the highest value" heuristic in the lottery game (see No-conflict bets subsection), following our pre-registration, we ran a control analysis where we excluded participants that had an accuracy lower than 50% both in their initial and final no-conflict trials (n = 4). In this control analysis all of our conclusions remained the same, suggesting that the heuristic did not bias our results. Therefore, in the results section below, we

present the intended complete analysis without exclusions. The partial results excluding these participants are reported in Supplementary Material C.

Results and Discussion

Proportion of EV maximizing choices

Our main question in this study was whether people who manage to make an EV maximizing choice after deliberating on a risky problem can also make this choice intuitively. So, we calculated, for each participant, the average proportion of EV maximizing initial and final responses for the conflict items. Figure 1 provides a summary of the percentage of EV maximizing choices for the critical conflict trials, both for the initial, intuitive and the final, deliberate responses, separately for both games. Note that for the statistical tests that are mentioned below loss averse trials were recoded as 0 and EV maximizing trials as 1.

Betting game. As Figure 1 shows, in the critical conflict trials of the betting game, the majority of responses were loss averse, but people still managed to provide EV maximizing responses. The proportion of EV maximizing responses reached 20.5% (*SD* = 30.9%) in the initial stage and 13.8% (*SD* = 27.9%) in the final stage. Interestingly, this percentage was higher in the initial, intuitive compared to the final stage where people were allowed to deliberate, and a paired-samples t-test showed that this difference was significant, $t(99) = 2.92$, $p = .004$. One possible explanation for this pattern could be that the loss aversion heuristic required some minimal deliberation to be activated (see Bago & De Neys, 2017). Critically, these results indicate that although the loss aversion bias is very prevalent people still manage to intuitively generate EV maximizing responses.

In the control, no-conflict trials, it was in participants' best interest to take the bet, both in terms of avoiding losses and acquiring gains. So, as expected, the proportion of trials in which participants took the bet reached 96.6% (*SD* = 9.4%) in the initial stage and 98.4% (*SD* = 7.6%) in the final stage. Given that the initial stage of the games was challenging–participants had to respond under a deadline and a cognitive load–one might argue that the intuitive responses in our study resulted from mere guessing. However, if the cognitive constraints had forced people to randomly click on one of the answer options, we would have found a much lower accuracy on the no-conflict problems. So, the ceiled initial performance argues against an overall guessing confound.

Finally, as expected, the filler items had a high accuracy both in the initial (M = 78.6%, SD = 33.8%) and final (M = 84.0%, SD = 30.2%) stage. This shows that, overall, participants did not rely on a "take the bet" heuristic when responding to the bets, which would have resulted in a floored performance on the filler trials (see also our control exclusion analysis in Supplementary Material C).

Lottery game. As Figure 1 shows, in the critical conflict trials of the lottery game, the majority of responses were loss averse[1], but participants still managed to provide EV maximizing responses. The proportion of EV maximizing responses reached 40.8% (SD = 37.0%) in the initial stage and 29.8% (SD = 32.9%) in the final stage. As in the betting game, this percentage was higher in the initial, intuitive stage compared to the final, deliberate stage, and a paired-samples t-test showed that this difference was significant, $t(99) = 4.14$, $p < .001$.

In the control, no-conflict trials participants were expected to always prefer one of the two lotteries, both in terms of certainty and potential gain. So, as expected, the proportion of trials in which participants chose the expected "correct" lottery pair reached 83.8% (SD = 23.6%) in the initial stage and 88.0% (SD = 23.8%) in the final stage. This shows that, overall, participants did not rely on a "pick the highest value" heuristic when responding to the lottery pairs, which would have resulted in a floored performance on the no-conflict trials (see also our control exclusion analysis in Supplementary Material C).

In the filler items the correct answer was made obvious for any adult with a basic understanding of the problems' probability rules. So, as expected, the proportion of trials in which participants gave the correct response was high both in the initial (M = 79.8%, SD = 25.0%) and the final (M = 90.3%, SD = 19.5%) response stage. If the cognitive constraints of the initial stage had forced people to provide random responses in the lottery game, we would have found a much lower accuracy at the filler problems. Thus, the good filler accuracy argues against an overall guessing confound.

[1] We use the loss averse label for consistency here. As noted in the Method, the lottery game does not imply losses per se, but rather measures risk aversiveness.

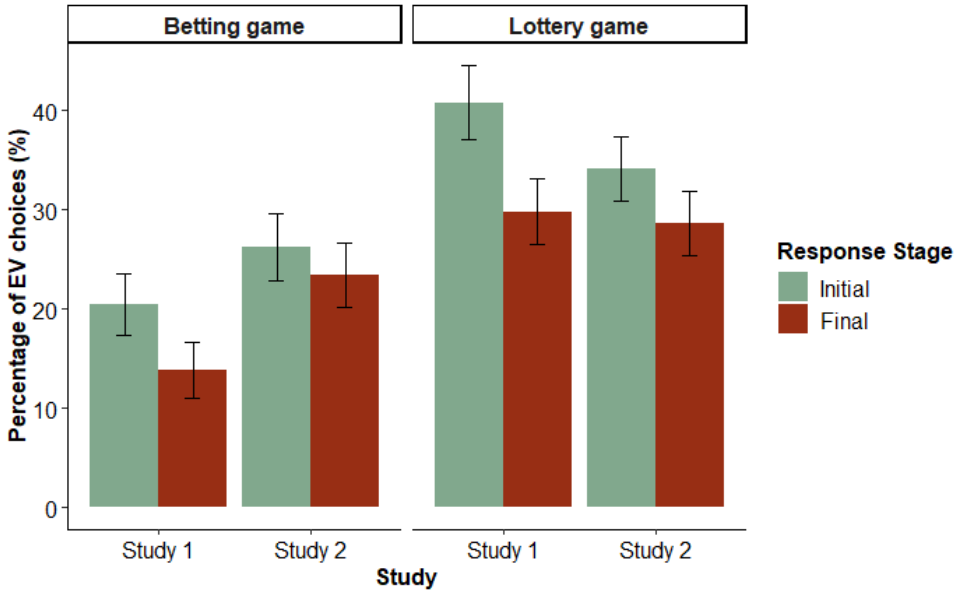

Figure 1. Percentage of Expected Value (EV) maximizing initial and final choices on conflict trials in the betting and lottery game, in Study 1 and Study 2. Error bars are standard errors of the mean.

Direction of change

To better understand how people changed (or did not change) their responses after deliberation we performed a direction of change analysis where we looked into how the accuracy changed from the initial to the final response stage (Bago & De Neys, 2017). In the conflict trials of both games participants could choose between two options: an EV maximizing choice and a loss averse choice. For simplicity, we coded the loss averse choice as "0", and the EV maximizing choice as "1". Consequently, four possible response patterns were possible on every trial: initial loss averse response and final loss averse response ("00"), initial loss averse response and final EV maximizing response ("01"), initial EV maximizing response and final loss averse response ("10"), and initial and final EV maximizing response ("11"). Figure 2 shows the mean proportions of each direction of change category in the conflict trials, separately for each game.

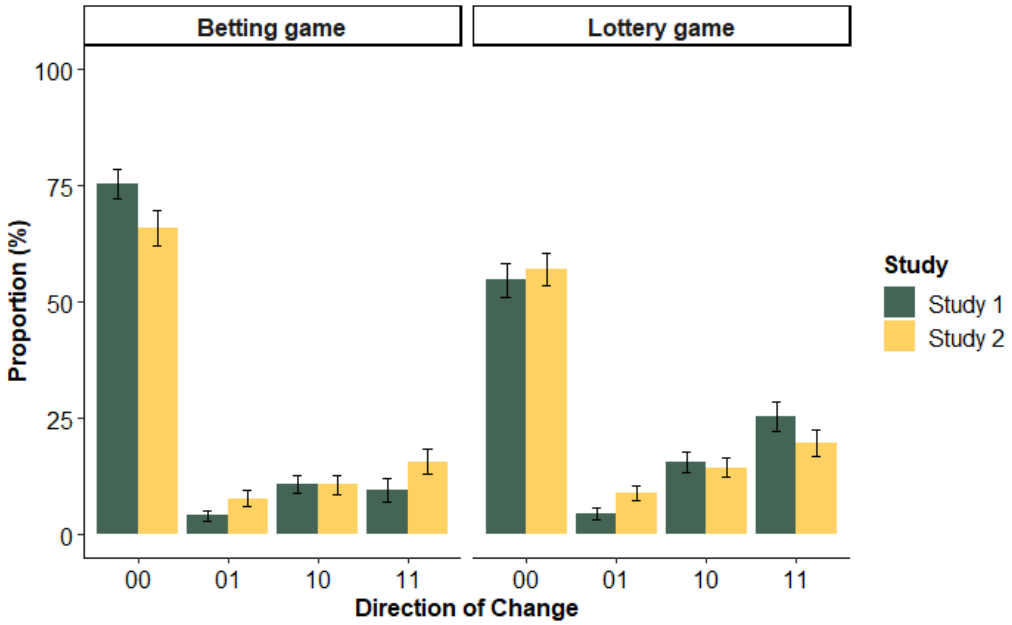

Figure 2. Proportion of each direction of change category in the conflict trials, separately for the betting game and the lottery game and for Study 1 and Study 2; "00" = initial and final loss averse response; "01" = initial loss averse response and final Expected Value (EV) maximizing response; "10" = initial EV maximizing response and final loss averse response; "11" = initial and final EV maximizing response. Error bars are standard errors of the mean.

The majority of conflict trials had a "00" pattern (75.4% in betting game; 54.7% in lottery game) which demonstrates that when taking risky decisions most participants remained loss averse even after they were given time to deliberate. Critically, "11" responses (9.7% in betting game; 25.3% in lottery game) were more frequent than "01" responses (4.1% in betting game; 4.5% in lottery game). This shows that in the cases where participants managed to generate an EV maximizing response after deliberation (i.e.,"01" and "11" cases), most of the time they had already arrived at this choice intuitively (i.e., "11" cases). In other words, deliberate correction exists, but it is relatively rare and not always necessary for EV-based responding. This is further highlighted by the percentage of "10" responses (10.8% in betting game; 15.5% in lottery game). These are the cases in which people intuitively provided EV maximizing answers and only after deliberating they changed them to loss averse ones. As in Bago and De Neys (2017) we also calculated the so-called non-correction rate (i.e., proportion 11/11+01). The non-correction rate indicates the proportion of final EV-maximizing choices which were already EV-maximizing in the initial response

stage. In other words, it shows the proportion of trials for which participants did not need to deliberate to make an EV-maximizing choice. The mean non-correction rate for the conflict items reached 70.3% in the betting game and 84.9% in the lottery game. So, when participants managed to make an EV-maximizing choice at the final stage of the conflict items, they had typically already made that choice at the initial stage most of the time.

Stability index

We also calculated a stability index on the standard, conflict trials. More specifically, for each participant we calculated on how many out of the five conflict trials they showed the same direction of change pattern (i.e., "00", "01", "10", or "11"). The average stability index in Study 1 was 85.6% (*SD* = 19.0%) in the betting game and 75.4% (*SD* = 21.0%) in the lottery game. If responding under load was prone to systematic guessing, we would expect more inconsistency in participants' responses across trials.

Confidence ratings

Following our preregistration and previous two-response studies on logical reasoning, we examined whether people who intuitively provide loss averse responses to conflict trials show some sensitivity to the fact that their answer goes against the items' EV. This would indicate that loss averse responders do not disregard EV altogether; instead they might be sensitive to EV principles, but they cannot overcome their loss aversion bias when making a choice. Note that in the no-conflict items loss aversion and EV calculations point to the same response. So, to see whether people are sensitive to the EV of the conflict problems, we can contrast their confidence at the correctly solved no-conflict items (i.e., their baseline confidence) with their confidence at the conflict items where they gave loss averse responses.

If loss averse people completely ignore the EV, they should process these conflict and no-conflict trials the same way. If, however, they detect that their loss averse responses are opposing the item's EV, they should show increased doubt in conflict trials. In other words, an increased doubt (or inversely a lowered confidence) in the conflict trials would be an indication that–despite their loss averse answer–people show some minimal sensitivity to EV.

Figure 3 shows the mean initial confidence ratings for conflict and no-conflict trials as a function of response type (EV maximizing; Loss averse; Other).[2] As it can be seen in Figure 3, the mean confidence ratings for conflict loss averse responses (72.4% in the betting game; 68.1% in the lottery game) were slightly lower than the mean confidence ratings for no-conflict correct responses (77.8% in the betting game; 69.8% in the lottery game). A Wilcoxon signed-ranked test showed that this difference was significant both in the betting game, $V = 1080$, $p = .002$, and in the lottery game, $V = 835$, $p = .04$. Thus, participants showed increased response doubt when making a loss averse choice on conflict trials, which suggests they were detecting to some extent that their answer conflicted with EV maximizing considerations. In other words, they considered the EV maximizing option, even though they eventually decided on the loss averse choice. Importantly, this doubt concerned people's initial responses for which deliberation was minimized[3]—suggesting that the conflict sensitivity was intuitive in nature.

For completeness, note that as Figure 3 indicates, we also observed a mean confidence decrease in the cases where people provided EV maximizing responses in the conflict items (44.2% in betting game; 60.1% in lottery game) compared to their mean confidence in the correct no-conflict items (77.8% in betting game; 69.8% in lottery game). A Wilcoxon signed-ranked test showed that this difference was significant both in the betting game, $V = 4$, $p < .001$, and the lottery game, $V = 407$, $p = .01$. Hence, when choosing the EV maximizing option, people also considered the alternative (loss averse) option, and detected that their answer was conflicting with the high chance of losing money.

For exploratory purposes we also looked into the mean confidence levels for each of the direction of change categories separately in Supplementary Material D. As in the logical reasoning field (e.g., Thompson et al., 2011), our results showed that initial confidence is lower on trials in which the initial response is changed after deliberation (i.e., "01" and "10" vs. "11" and "00" categories).

[2] In the no-conflict trials of the betting game participants could either choose to take the bet, which in Figure 3 is the "EV maximizing" choice, or to not take the bet, which is "Other" since it was neither loss averse nor EV maximizing. In the no-conflict items of the lottery game both expected value considerations and loss aversion led to the same choice. For consistency with the betting game, in Figure 3 this choice is named "EV maximizing". Note, however, that it could also be driven by loss aversion. When participants did not choose this option their choice was named "Other" since it was neither loss averse nor EV maximizing.

[3] As a reminder, note that the initial confidence rating was also given under load.

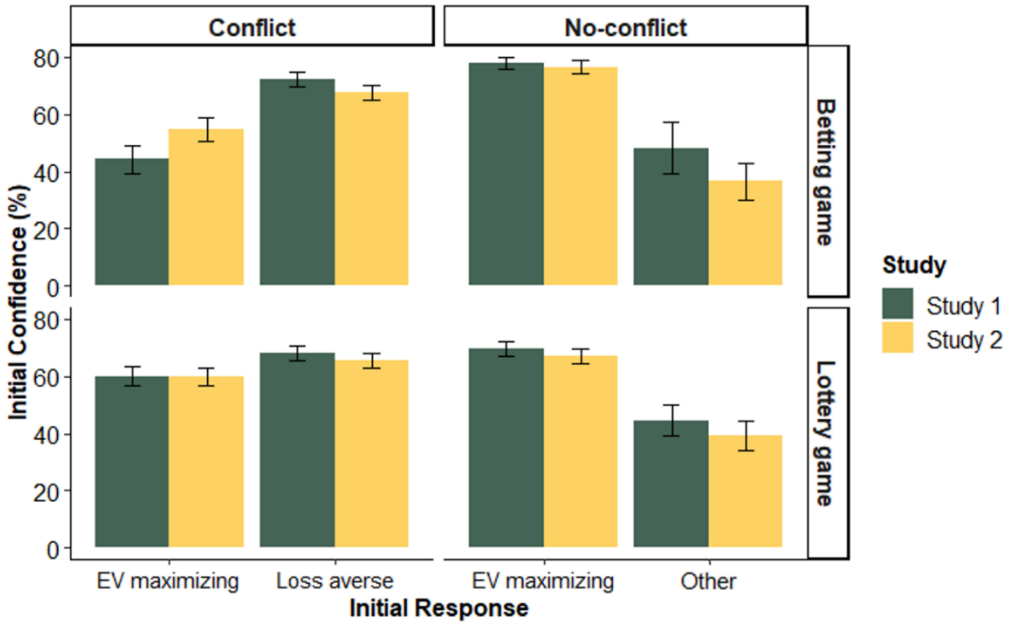

Figure 3. Initial response confidence ratings as a function of response type (Expected Value maximizing; Loss averse; Other) and conflict status (conflict; no-conflict) separately for the two games and for Study 1 and Study 2. Error bars are standard errors of the mean.

Study 2

The results of our first study demonstrate that people who manage to make an EV maximizing choice when deliberating on a risky decision, have typically already made this choice intuitively. Deliberate correction is, thus, not the prevalent route for EV-based responding. In addition, Study 1 also revealed that even when people make loss averse choices, they have an intuitive sensitivity to the fact that their decisions conflict with EV considerations (just like EV responders showed sensitivity to the presence of a conflicting loss averse option). This further indicates that taking EV considerations into account does not necessarily require deliberation.

In Study 2, we introduced methodological refinements to test the robustness of our findings. First, Study 1's sample consisted of 80% female participants, and previous studies have shown that loss aversion is susceptible to gender differences (see Croson & Gneezy 2009, for a review; but see also Filippin & Crosetto, 2016). So, in Study 2 we recruited a gender-balanced sample. Second, participants in Study 1 played the games for hypothetical pay-offs, but in Study 2

we incentivized our participants (see Hertwig & Ortmann, 2001 for a discussion on the importance of monetary incentives).

It should be noted that Study 2 was designed as the present Study 3 (which includes 5 "easy" conflict items on top of the items of Study 1, see further). However, due to a coding error, in Study 2 participants were presented with the items of Study 1 but saw each conflict item twice. To ensure that our results were not influenced by repeated exposure effects, in Study 2 we only kept the item from each conflict pair that was presented first. This way, the present study served as a refined robustness test of Study 1.

Method

Data availability

All data and material are available on the Open Science Framework (https://osf.io/hzjt8/files/osfstorage).

Participants

We recruited our participants online on Prolific Academic (www.prolific.ac). Only native English speakers from Canada, Australia, New Zealand, the United States of America, or the United Kingdom were allowed to take part in the study. Participants were paid £2.50 for their participation (£6 hourly rate[4]). They also received a possible monetary bonus payment of up to £1, depending on the game's outcome. One hundred participants (48 female, mean age = 35.7 years, *SD* = 12.0 years) participated in the study. A total of 40% of participants reported high school as the highest completed educational level, while 60% reported having a postsecondary education degree.

Procedure

One-response (deliberative-only) pretest. The initial response deadline of Study 2 was calculated based on the one-response pretest of Study 3 (see One-response pretest section in Study 3). For the betting game it was set to 4 s and for the lottery game it was set to 4.5 s. Note that this was a stricter deadline than in Study 1.

[4] The hourly rate in this study is £6 instead of the £5 hourly rate of Study 1, as in between these studies Prolific increased their minimum pay.

Two-response games. The experiment was run online on the Qualtrics (www.qualtrics.com) software server. The general instructions were the same as those in Study 1. However, here participants were told that, at the end of the study, one of their initial or final choices would be selected at random from each game, and would be played to determine the earnings for the option that they selected. It was specified that any money they made from the randomly selected bet and lottery pair would be added together and multiplied by a factor of 0.0013, meaning that they could earn from £0 to £1 extra in addition to their standard payment (i.e., a potential 40% increase in their total earnings).

With the exception of the initial response deadline, the practice and experimental trials were the same as in Study 1. As mentioned above, once the bet or lottery pair appeared in this study, participants had 4 s to enter their answer in the betting game and 4.5 s in the lottery game. As in Study 1, 1 s before the deadline the background of the screen turned yellow to warn participants that the time limit is approaching.

Finally, in both games the items were randomly presented, apart from the final item, which was always the same conflict item and was followed by a justification question. After they responded to this item, participants saw a screen that read "We are interested in the reasoning behind your response to the final bet/lottery". They were shown the item again and were asked to justify, in an open-response format, why they felt their previously entered response to the item was the most advantageous choice for them to make. We added this exploratory question to get an insight into the rationale participants used to arrive at their answers (see Supplementary Material E for an analysis of the justifications).

Counterbalancing. Each game (betting and lottery) was composed of ten[5] conflict, five no-conflict and five filler items. For each game separately, two sets of items (set A and set B) were created in which the conflict status of the conflict and no-conflict items was counterbalanced. More specifically, all the conflict items of set A appeared in their no-conflict version in set B, and all the no-conflict items in set A appeared in their conflict version in set B. It should be noted that in the betting game of this study, the values of the bets were slightly different compared to those of Study 1. This change allowed us to create a set B of items for

[5] As noted, each conflict item was presented twice due to a coding error. However, we only analyzed the first presentation of each conflict item and discarded the repeated items.

counterbalancing, which was not possible with the bets of Study 1 (all items can be found in Supplementary Material B). In addition, we created two variations of the filler items for set A and set B respectively. The items of set B were created by increasing the items' values by €5-10, so that the EV difference remained the same in the respective items of both sets. Half of the participants were presented with set A while the other half was presented with set B. So, in both games, the same content was never presented more than once to a participant and everyone was exposed to the same items.

Exclusion Criteria

As in Study 1, the trials in which participants failed the load and/or the deadline were excluded from subsequent analyses.

Betting game. Participants failed to answer before the deadline on 4.6% of conflict trials, 1.4% of no-conflict trials, and 4.6% of filler trials. In addition, they failed the load task on 7.4% of conflict trials, 7% of no-conflict trials, and 8% of filler trials. Overall, by rejecting the missed deadline and missed load trials we kept 88% of conflict trials, 91.6% of no-conflict trials, and 87.4% of filler trials. On average, each participant contributed 13.4 trials (out of 15 trials, *SD* = 1.8).

As in Study 1, to ensure that participants were not using an "always take the bet" heuristic in the betting game (see Filler bets subsection above), following our pre-registration, we ran a control analysis where we excluded participants that had an accuracy lower than 50% both in their initial and final filler trials (*n* = 8). In this control analysis all of our conclusions remained the same, suggesting that the heuristic did not bias our results. Therefore, in the results section below, we present the intended complete analysis without exclusions. The partial results excluding these participants are reported in Supplementary Material C.

Lottery game. Participants failed to answer before the deadline on 1.8% of conflict trials, 3.2% of no-conflict trials, and 3.4% of filler trials. In addition, participants failed the load task on 10.6% of conflict trials, 3.2% of no-conflict trials, and 11.6% of filler trials. Overall, by rejecting the missed deadline and missed load trials we kept 87.6% of conflict trials, 93.6% of no-conflict trials, and 85% of filler trials. On average, each participant contributed 13.3 trials (out of 15 trials, *SD* = 1.9).

As in Study 1, to ensure that participants were not using a "pick the highest value" heuristic in the lottery game, following our pre-registration, we ran a control analysis where we excluded participants that had an accuracy lower than 50% both in their initial and final no-conflict trials (n = 3). In this control analysis all of our conclusions remained the same, so in the results section below we present the intended complete analysis without exclusions. The partial results excluding these participants are reported in Supplementary Material C.

Results and Discussion

Proportion of EV maximizing choices

As Figure 1 shows the accuracy results were very similar across Study 1 and Study 2, for both games.

Betting game. In the critical conflict trials of the betting game, the majority of responses were loss averse, but people still managed to provide EV maximizing responses. The proportion of EV maximizing responses reached 26.2% (SD = 34.4%) in the initial stage and 23.4% (SD = 32.3%) in the final stage, but a paired-samples t-test showed that this difference was not significant, $t(99) = 1.02$, $p = .31$. Critically, these results again show that people managed to generate EV maximizing responses intuitively.

In the control, no-conflict trials, participants' accuracy remained at ceiling. The proportion of trials in which participants took the bet reached 95.8% (SD = 12.0%) in the initial stage and 97.7% (SD = 7.3%) in the final stage. Finally, as expected, the filler items had a high accuracy both in the initial (M = 77.1%, SD = 30.8%) and the final (M = 83.0%, SD = 26.8%) stage.

Lottery game. Similar to our previous results, most responses in the critical conflict trials of the lottery game were loss averse, but participants managed to provide EV maximizing responses. The proportion of EV maximizing responses reached 34.1% (SD = 32.2%) in the initial stage and 28.6% (SD = 32.4%) in the final stage, but a paired-samples t-test showed that this difference was not significant, $t(99) = 1.95$, $p = .05$.

Concerning the control, no-conflict lottery pairs the proportion of trials in which participants chose the expected, "correct" lottery pair reached 86.8% (SD = 26.1%) in the initial stage and 93.1% (SD = 16.7%) in the final stage. In the

filler items, as expected, the proportion of trials in which participants gave the correct response was high both in the initial (*M* = 82.5%, *SD* = 21.8%) and the final (*M* = 92.0%, *SD* = 15.2%) response stage.

Direction of change

In Study 2, the same direction of change analysis as in Study 1 was performed. As it can be seen in Figure 2, the direction of change analysis results were very similar across the two studies, for both games. The majority of conflict trials had a "00" pattern (66.0% in betting game; 56.9% in lottery game) which shows that, even after deliberation, most people remained loss averse. Critically, the "11" responses (15.6% in betting game; 19.6% in lottery game) were again more frequent than the "01" (7.8% in betting game; 9.0% in lottery game). The non-correction rate reached 66.7% for the betting game and 68.5% for the lottery game. It is worth noting that in Study 2 the proportion of "11" response slightly decreased and that of "01" slightly increased when compared to Study 1. However, the main response pattern remained the same in the sense that when people managed to provide an EV maximizing response after deliberation, they had often already arrived to this choice intuitively.

Stability index

The average stability index in Study 2 was 81.5% (*SD* = 21.1%) in the betting game and 73.1% (*SD* = 21.3%) in the lottery game. This response consistency further indicates that participants were not systematically responding randomly.

Confidence Ratings

As in Study 1, we looked at participants' initial confidence ratings to see whether people are intuitively sensitive to the EV of the conflict problems when making loss averse choices. Figure 3 shows the mean initial confidence ratings for conflict and no-conflict trials as a function of response type (EV maximizing; Loss averse; Other). The mean confidence ratings for conflict loss averse responses (67.4% in the betting game; 65.5% in the lottery game) were lower than the mean confidence ratings for no-conflict correct responses (76.6% in the betting game; 67.1% in the lottery game). A Wilcoxon signed-ranked test showed that this difference was significant in the betting game, $V = 791$, $p < .001$, but not in

the lottery game, $V = 1261$, $p = .12$. Thus, in the betting game, participants showed an increased response doubt when making a loss averse choice, which suggests they were detecting that their answer conflicted with EV maximizing considerations. Importantly, this happened in the initial response stage where deliberation was minimized.

As in Study 1, we also observed a confidence decrease for conflict EV maximizing responses. More specifically, the confidence ratings for conflict EV maximizing responses (54.9% in the betting game; 59.8% in the lottery game) were lower than the mean ratings at the correct, no-conflict items (76.6% in the betting game; 67.1% in the lottery game). A Wilcoxon signed-ranked test showed that this difference was significant both in the betting game, $V = 33.5$, $p < .001$, and the lottery game, $V = 577$, $p = .01$. Hence, EV responders also showed sensitivity to the alternative loss averse option.

Study 3

Results from Study 1 and Study 2 showed that in most cases that people arrive at an EV maximizing response after deliberation, they have already generated this response intuitively. So, although deliberate correction exists in risky decision making, it is not the primary route for EV-based responding. However, one cannot ignore that, across our two studies, the majority of responses were loss averse, even in the final, deliberate stage ($M = 81.4\%$ in the betting game; $M = 72.6\%$ in the lottery game). This implies that participants found it particularly difficult to opt for the expected value option in these items. In turn, this raises the concern that perhaps only the participants with very high cognitive capacities were able to provide EV maximizing responses. This could explain why, in our studies, EV maximizing responses are mostly generated intuitively (i.e., because highly gifted reasoners are particularly good at logical intuitive responding, e.g., Thompson et al., 2018). Hence, our results may only be representative for harder EV items. With easier items (where a wider range of participants arrives at the EV maximizing choice), the EV calculation might still require deliberation—as predicted by the standard dual process model. To test this, and to ensure that our findings are generalizable to different item types, in Study 3 we included "easy" conflict items (i.e., items that have a higher expected value difference than those of Studies 1 and 2). These "easy" items should show

a higher selection of EV choices and allow us to test whether deliberate correction is more common than among the "hard" conflict items from Study 1 and 2. Similarly to Study 2, Study 3 was also incentivized and had a gender balanced sample.

Method

Preregistration and data availability

The study design and hypothesis were preregistered on the Open science Framework (https://osf.io/rdj7h). No specific analyses were preregistered. All data and material are also available on the Open Science Framework (https://osf.io/hzjt8/files/osfstorage).

Participants

We recruited our participants online on Prolific Academic (www.prolific.ac). Only native English speakers from Canada, Australia, New Zealand, the United States of America, or the United Kingdom were allowed to take part in the study. Participants were paid £2.50 for their participation (£6 hourly rate). They also received a possible monetary bonus payment of up to £1, depending on the game's outcome. One hundred participants (49 female, mean age = 38.5 years, *SD* = 12.2 years) participated in the study. A total of 29% of participants reported high school as the highest completed educational level, while 71% reported having a postsecondary education degree.

Materials

Betting Game. The hard-conflict, no-conflict and filler bets were the same as in Study 2. All easy-conflict bets had the same probability of losing money as their respective hard-conflict bets, but a larger expected value. Therefore, the conflict between avoiding a potential loss (i.e., by not taking the bet) and taking a risk in order to acquire a(n) (even bigger) potential gain (i.e., by taking the bet) was reduced. Below we present an example of an original hard-conflict bet (left) and its equivalent easy-conflict bet (right):

Hard-conflict bet:	Easy-conflict bet:
If you take this bet you have:	*If you take this bet you have:*
5%** probability to **WIN €110	***5%** probability to **WIN €290***
95%** probability to **LOSE €5	***95%** probability to **LOSE €1***
Do you take the bet?	*Do you take the bet?*
○ *Yes*	○ *Yes*
○ *No*	○ *No*

The easy-conflict items were constructed based on the hard-conflict items by keeping all probabilities the same and decreasing the losing value (V_{lose}) by €4-€5 while increasing the winning value (V_{win}) by €120-€180. In all cases, an easy-conflict bet had more than double the winning value (V_{win}) of their equivalent hard-conflict bet. The exact values of the easy-conflict items were chosen so that the expected value difference (P_{win} * V_{win} – P_{lose} * V_{lose}) was kept as similar as possible between them (see Supplementary Material B for all items), to make sure that the items were of equal complexity.

Lottery Game. The hard-conflict , no-conflict and filler lottery pairs were the same as in Study 2. All easy-conflict lottery pairs had the same probabilities (P_{A_large}, P_{A_small}, P_{Blarge}, P_{B_small}) and the same smaller values (V_{A_small} & V_{B_small}) as their respective hard-conflict lottery pairs. However, they had a higher V_{large} in the most profitable but uncertain lottery and a lower V_{large} in the lottery with the highest guaranteed minimal gain and lower overall profit. Therefore, less conflict was created (when compared to the hard-conflict items). Below is an example of an original hard-conflict lottery pair (left) and its equivalent easy-conflict pair (right):

Hard-conflict lottery pair:		Easy-conflict lottery pair:	
Lottery A	***Lottery B***	***Lottery A***	***Lottery B***
70%** probability to **win €350	***70%** probability to **win €230***	***70%** probability to **win €440***	***70%** probability to **win €180***
30%** probability to **win €10	***30%** probability to **win €160***	***30%** probability to **win €10***	***30%** probability to **win €160***
Which lottery do you choose?		*Which lottery do you choose?*	
○ *A*		○ *A*	
○ *B*		○ *B*	

The easy-conflict lottery pairs were constructed on the basis of the hard-conflict items, by keeping all "smaller" values (V_{A_small} & V_{B_small}) the same and decreasing the "large" value (V_{B_large}) of the certain but less profitable lottery pair

by €20-€50, while increasing the "large" value (V_{B_large}) of the uncertain but more profitable lottery pair by €90-€150. The values were chosen so that the expected value difference [(P_{A_large} * V_{A_large}) + (P_{A_small} * V_{A_small}) – (P_{B_large} * V_{B_large}) + (P_{B_small} * V_{B_small})] was kept as similar as possible between all easy-conflict lottery pairs (see Supplementary Material B for all items).

Procedure

One-response (deliberative-only) pretest. Since participants were presented with 5 extra items in this study, we decided to recalibrate the deadline. As before, we ran a traditional one-response version of our study (without load or deadline) to obtain a baseline performance. We recruited an independent sample of 50 participants (48% female; mean age = 39.1 years, *SD* = 15.0) online on Prolific Academic (www.prolific.ac). Only native English speakers from Canada, Australia, New Zealand, the United States of America, or the United Kingdom were allowed to take part in the study. Participants were paid £1.2 for their participation (£6 hourly rate). A total of 44% of the participants reported high school as highest completed educational level, while 54% reported having a postsecondary education degree.

In the betting game, EV maximizing responses to the conflict bets took more time (5.3 s, *SD* = 2.7 s) than loss averse responses (4.8 s, *SD* = 2.5 s)[6]. So, following the same rationale as in Study 1, we based our deadline on the reaction time of the EV maximizing trials only. The first quartile of these trials was 3.3 s. However the mean reaction time for the correct no-conflict items was 3.9 s. To avoid excess missed trials and to be consistent with Study 1 (where the deadline was higher than the no-conflict baseline), we rounded the reaction time to the nearest integer and set the deadline for the betting game to 4 s (i.e., 0.5 s less than in Study 1). In the lottery game, the EV maximizing responses to the conflict lottery pairs (7.0 s, *SD* = 4.6 s) also took more time than the loss averse responses (6.1 s, *SD* = 3.4 s). So, following the same rationale as with the betting game we based our deadline on the reaction time of the EV maximizing trials only. The first quartile of these trials was 4.45 s, so we rounded to the nearest decimal and set the deadline for the lottery game to 4.5 s (i.e., 1 s less than in Study 1).

[6] The reaction times presented in this section are the average reaction time across the hard-conflict and easy-conflict items.

To ensure that participants were under time pressure during the initial stage, we compared the conflict trials' response times between the one-response pretest and the initial responses of the main two-response study. Participants responded much faster in the initial response stage of the main, two-response study, compared to the one-response pretest both in the betting game ($M_{two\text{-}response}$ = 2.6 s; $M_{one\text{-}response}$ = 4.8 s) and the lottery game ($M_{two\text{-}response}$ = 2.7 s; $M_{one\text{-}response}$ = 6.4 s). Welch Two Sample t-tests indicated that this difference was significant for the betting game, $t(86.68) = 9.42$, $p < .001$, and the lottery game, $t(103.11) = 10.73$, $p < .001$.

The one-response pre-test also allowed us to rule out a potential consistency confound in our main two-response study. To do so, we contrasted the proportion of EV maximizing responses in the conflict trials of the one-response pretest and those of the final stage of the main two-response study. Our results showed that the percentage of EV maximizing responses in the conflict trials of the one-response pretest, was very similar to this of the final conflict responses in the two-response study for the lottery game ($M_{two\text{-}response}$ = 41.6%; $M_{one\text{-}response}$ = 43.2%) and to a lesser extent for the betting game ($M_{two\text{-}response}$ = 40.2%; $M_{one\text{-}response}$ = 31.6%). Welch Two Sample t-tests indicated that the difference was not significant for the lottery game, $t(223.03) = 0.39$, $p = .70$, and that it was significant for the betting game, $t(220.67) = -2.00$, $p = .05$, but with the final trials of the two-response study having a higher accuracy than those of the one-response pretest. These results directly show that a consistency confound cannot account for the lack of deliberate EV correction.

Two-response games. The experiment was run online on the Qualtrics (www.qualtrics.com) software server. Apart from the differences in the items, the instructions and the procedure were the same as those in Study 2.

Counterbalancing. Each game (betting and lottery) was composed of five easy-conflict, five hard-conflict, five no-conflict and five filler items. For each game separately, two sets of items (set A and set B) were created in which the conflict status of the no-conflict and the hard-conflict items was counterbalanced. More specifically, all the hard-conflict items of set A appeared in their no-conflict version in set B, and all the no-conflict items in set A appeared in their hard-conflict version in set B. We also created two variations of the easy-conflict and filler items

for set A and set B respectively. The items of set B were created by increasing the items' values by €5 to €10. The exact values were determined so that the EV difference remained the same for the respective items of both sets. Half of the participants were presented with set A while the other half were presented with set B. So, in both games, the same content was never presented more than once to a participant and everyone was exposed to the same items, which minimized the possibility that mere item differences influence the results.

Exclusion Criteria

As before, the trials in which participants failed the load and/or the deadline were excluded from subsequent analyses.

Betting game. Participants failed to answer before the deadline on 5% (easy- and hard-) conflict trials, 2% of no-conflict trials, and 3.2% of filler trials. In addition, they failed the load task on 11.5% of conflict trials, 6.6% of no-conflict trials, and 9.6% of filler trials. Overall, by rejecting the missed deadline and missed load trials we kept 83.5% of conflict trials, 91.4% of no-conflict trials, and 87.2% of filler trials. On average, each participant contributed 17.3 trials (out of 20 trials, *SD* = 2.2).

We also ran a control analysis where we excluded participants that had an accuracy lower than 50% both in their initial and final filler trials (n = 10). In this control analysis all of our conclusions remained the same (see Supplementary Material C). Therefore, in the results section below, we present the intended complete analysis without exclusions.

Lottery game. Participants failed to answer before the deadline on 2.3% of conflict trials, 2.4% of no-conflict trials, and 4.4% of filler trials. In addition, participants failed the load task 15.2% of conflict trials, 5.8% of no-conflict trials, and 14.2% of filler trials. Overall, by rejecting the missed deadline and missed load trials we kept 82.5% of conflict trials, 91.8% of no-conflict trials, and 81.4% of filler trials. On average, each participant contributed 16.9 trials (out of 20 trials, *SD* = 2.5).

We also ran a control analysis where we excluded participants that had an accuracy lower than 50% both in their initial and final no-conflict trials (n = 2). In this control analysis all of our conclusions remained the same (see Supplementary Material C). Below we present the intended complete analysis without exclusions.

Results and Discussion

Proportion of EV maximizing choices

Betting game. As Figure 4 shows, regarding the hard-conflict items, the proportion of EV maximizing responses reached 30.6% (*SD* = 33.3%) in the initial stage and 27.1% (*SD* = 34.7%) in the final stage, but a paired-samples t-test showed that this difference was not significant, $t(99) = 1.48$, $p = .14$. Regarding the easy-conflict items, the proportion of EV maximizing responses reached 45.1% (*SD* = 34.0%) in the initial stage and 53.3% (*SD* = 36.4%) in the final stage, and a paired-samples t-test showed that this difference was significant, $t(99) = -3.02$, $p = .003$. Overall, these results show that for the hard-conflict items people manage to provide EV maximizing responses intuitively around 30% of the time, while for the easy-conflict items they manage to do so around 45% of the time. Thus, as expected, the proportion of initial (and final) EV maximizing responses was higher in the easy compared to the hard items. This implies that our manipulation of the items' difficulty was successful and that participants were indeed more likely to take bets with higher EV.

In the control, no-conflict items, participants' accuracy remained at ceiling. For these items, the proportion of trials where participants took the bet reached 92.3% (*SD* = 16.0%) in the initial stage and 92.9% (*SD* = 16.8%) in the final stage. Finally, the filler items had a high accuracy both in the initial (*M* = 77.0%, *SD* = 31.7%) and the final (*M* = 83.7%, *SD* = 28.6%) stage.

Lottery game. Regarding the hard-conflict items, the proportion of EV maximizing responses reached 40.7% (*SD* = 36.2%) in the initial response stage and 31.8% (*SD* = 35.7%) in the final stage, and a paired-samples t-test showed that this difference was significant, $t(99) = 2.86$, $p = .01$. For the easy-conflict items, the proportion of EV maximizing responses reached 53.6% (*SD* = 33.9%) in the initial stage and 51.3% (*SD* = 37.8%) in the final stage, and a paired-samples t-test showed that this difference was not significant, $t(99) = 0.82$, $p = .42$. So, overall, for the hard-conflict items people managed to provide EV maximizing responses intuitively around 40% of the time, and for the easy-conflict items around 54% of the time.

In the control, no-conflict problems the proportion of trials in which participants chose the expected, "correct" lottery pair reached 88.4% (*SD* =

19.2%) in the initial stage and 93.7% (*SD* = 14.8%) in the final stage. Finally, as expected, in the filler items the proportion of trials in which participants gave the correct response was high both in the initial (*M* = 79.8%, *SD* = 23.3%) and the final response stage (*M* = 92.7%, *SD* = 18.67%).

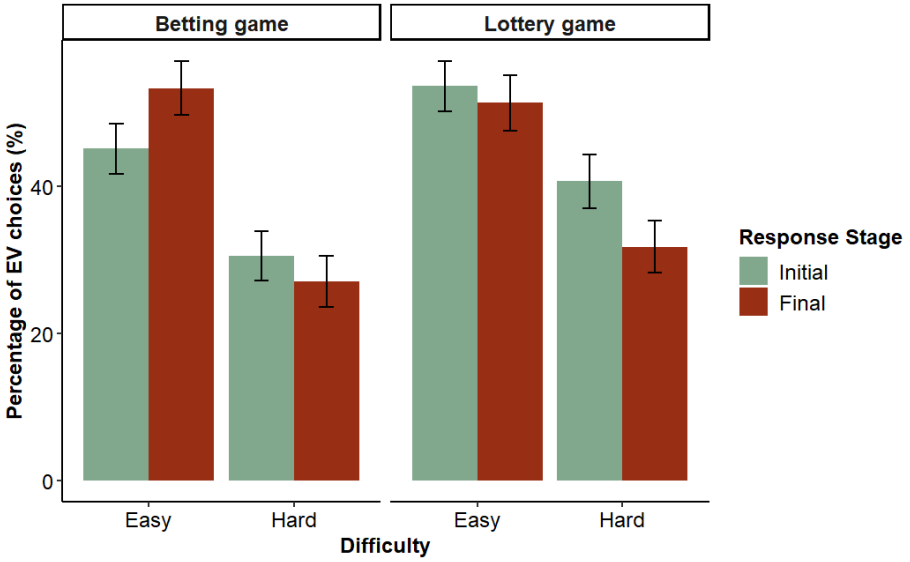

Figure 4. Percentage of Expected Value (EV) maximizing initial and final choices on the conflict trials of Study 3, separately for the betting and lottery game and for easy and hard conflict items. Error bars are standard errors of the mean.

Direction of change

As it can be seen in Figure 5, for the hard-conflict items, the majority of conflict trials had a "00" pattern (64.0% in betting game; 52.4% in lottery game) which indicates that, even after deliberating, most people remained loss averse. Critically, the "11" responses (21.6% in betting game; 24.8% in lottery game) were more frequent than the "01" responses (5.4% in betting game; 7.0% in lottery game), and the non-correction rate reached 80.0% in the betting game and 78.0% in the lottery game.

For the easy-conflict items, most conflict trials had either a "00" pattern (39.9% in betting game; 34.8% in lottery game) or a "11" pattern (38.3% in betting game; 39.7% in lottery game). There were fewer cases of "01" (15.0% in betting game; 11.6% in lottery game) patterns. Hence, although people were less loss averse when responding to the easy-conflict (compared to the hard-conflict) items, intuitive and deliberate loss averse responses were still prevalent. Critically, the "11" responses were again more frequent than the "01" responses, and the non-correction rate reached 71.9% in the betting game and 77.4% in the lottery

game. So, even with the easier items, people predominantly made EV choices without deliberating.

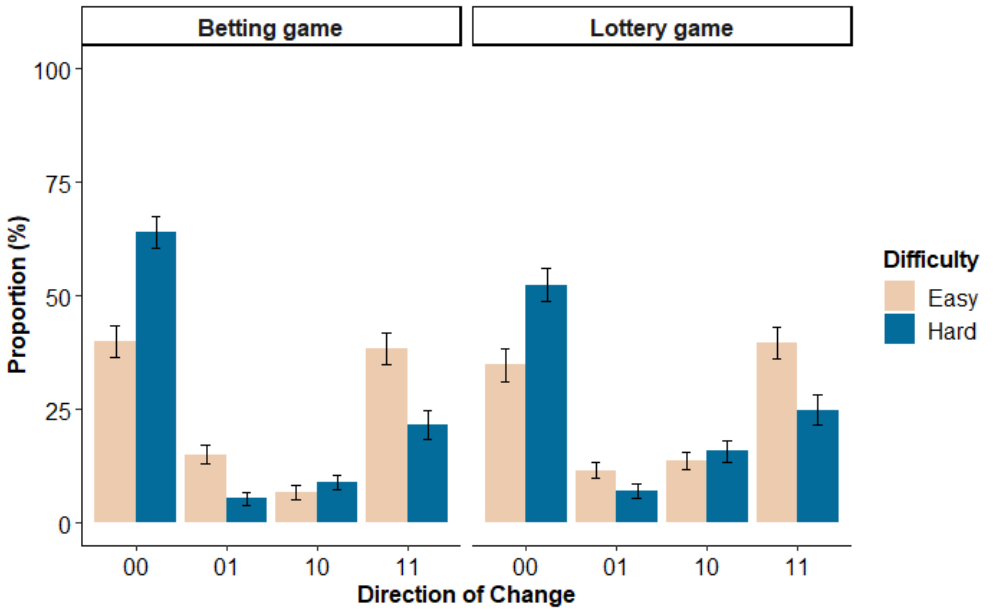

Figure 5. Proportion of each direction of change category in the conflict trials of Study 3, separately for the betting game and the lottery game and for easy-conflict and hard-conflict items; "00" = initial and final loss averse response; "01" = initial loss averse response and final Expected Value (EV) maximizing response; "10" = initial EV maximizing response and final loss averse response; "11" = initial and final EV maximizing response. Error bars are standard errors of the mean.

Stability index

Regarding the hard-conflict items, the average stability index was 80.0% (*SD* = 20.4%) in the betting game and 75.3% (*SD* = 21.5%) in the lottery game. For the easy-conflict items the index reached 71.2% (*SD* = 20.9%) in the betting game and 69.8% (*SD* = 22.4%) in the lottery game. This response consistency further indicates that participants were not systematically responding randomly.

Confidence Ratings

Figure 6 shows the mean initial confidence ratings for hard-conflict, easy-conflict and no-conflict trials as a function of response type (EV maximizing; Loss averse; Other). Starting with the hard-conflict items, the mean confidence ratings for hard-conflict loss averse responses (70.0% in the betting game; 68.1% in the lottery game) tended to be lower than the mean confidence ratings for no-conflict correct responses (79.6% in the betting game; 68.6% in the lottery game). A

Wilcoxon signed-ranked test showed that this difference was significant in the betting game, $V = 695.5$, $p < .001$, but not in the lottery game, $V = 1053.5$, $p = .20$. Thus, in the hard-conflict items of the betting game, participants showed an increased response doubt when they were choosing the loss averse option, suggesting that they were detecting that their answer conflicted with EV maximizing considerations.

Concerning the easy-conflict items, the mean confidence ratings for easy-conflict loss averse responses (65.7 % in the betting game; 64.7% in the lottery game) were lower than the mean confidence ratings for no-conflict correct responses (79.6% in the betting game; 68.6% in the lottery game). A Wilcoxon signed-ranked test showed that this difference was significant both in the betting game, $V = 396$, $p < .001$, and the lottery game, $V = 859$, $p = .007$.

As in our previous studies, we also observed a confidence decrease when people provided EV maximizing responses in the conflict problems. In the hard-conflict items the confidence ratings for conflict EV maximizing responses (52.4% in the betting game; 64.1% in the lottery game) were lower than those of the correct, no-conflict items (79.6% in the betting game; 68.6% in the lottery game). A Wilcoxon signed-ranked test showed that this difference was significant in the betting game, $V = 27.5$, $p < .001$, but not in the lottery game, $V = 645.5$, $p = .15$.

In the easy-conflict items the same pattern was observed. More specifically, the confidence for conflict EV maximizing responses (63.6% in the betting game; 64.1% in the lottery game) was lower than that at the correct, no-conflict items (79.6% in the betting game; 68.6% in the lottery game). A Wilcoxon signed-ranked test showed that this difference was significant in the betting game, $V = 460$, $p < .001$, but not in the lottery game, $V = 1072$, $p = .09$. Hence, in the betting game EV responders showed sensitivity to the alternative loss averse option.

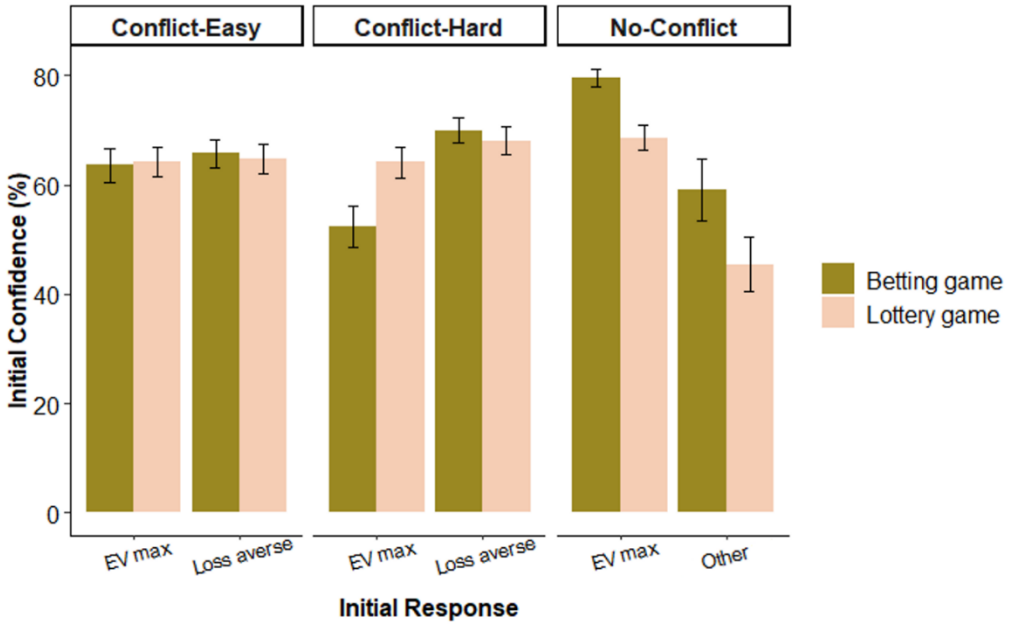

Figure 6. Initial response confidence ratings in Study 3 as a function of response type (Ex maximizing; Loss averse; Other) and conflict status (conflict-easy; conflict-hard; no-conflict) separately for the two games. Error bars are standard errors of the mean.

General Discussion

Our main goal in this paper was to test the corrective dual process assumption of risky decision making. According to this popular view, when people take risky decisions, deliberation is necessary for them to take the expected value of their decision into account and avoid the loss aversion bias (Slovic et al., 2005). In this paper we directly tested whether expected value maximizing decisions can also be intuitively generated, in the absence of deliberate processing. Across our three studies we found that, when playing two-outcome risky choice games, participants usually opted for the loss averse choice (instead of the expected value maximizing one), both after mere intuitive processing and after deliberating. However, in the cases where people chose the expected value maximizing choice after deliberation, they had predominantly already arrived at this choice intuitively. In Study 3 we replicated this finding with items that had a larger expected value. With these items more participants managed to provide expected value choices after deliberation and, here too, they had predominantly already made the expected-value maximizing choice in the intuitive stage. In sum, across our three

studies we found that deliberation is not the primary route for expected-value-based responding.

As a second step in our paper, we examined whether people were sensitive to expected value principles even when they intuitively chose the loss averse option. We found that in the conflict trials of the betting game, people consistently displayed decreased intuitive confidence levels (compared to baseline confidence) when they made a loss averse choice. Interestingly, they also showed decreased confidence in the conflict trials where they chose the expected value maximizing option. This suggests that, no matter what option people end up selecting, they intuitively process the alternative option as well. However, in the lottery game the results were less consistent, as participants reported a decreased confidence in some conflict trials (i.e., in Study 1 and in the easy-conflict items of Study 3), but not in others. In sum, these results indicate that when people opt for the loss averse choice, the expected-value maximizing choice is also often activated (and vice-versa).

Recent two-response studies in the reasoning field have reported similar results. More specifically, they have shown that logico-probabilistic principles which were traditionally thought to be processed only after deliberation, can also be processed intuitively (e.g., Bago & De Neys, 2017; but see also Ghasemi et al., 2023; Meyer-Grant et al., 2022). This finding has been replicated in a range of classic "bias" tasks, such as the bat-and-ball (e.g., Burič & Konrádová , 2021; Raoelison & De Neys, 2019), base-rate neglect, conjunction fallacy (Boissin et al., 2022), and syllogistic reasoning problems (e.g., Bago & De Neys, 2017; Raoelison et al., 2020). The present findings point to a similar upgraded role of intuition in risky decisions.

Decision making under risk is in itself a field of particular importance. To start with, risky decisions are the focus of prospect theory (Kahneman & Tversky, 1979) which lies at the heart of Kahneman and Tversky's work on heuristics and biases. So, it is important to understand their underlying mechanisms. Our results inform risky decision making research by showing that when people manage to take the expected value of their decisions into account, they often do so intuitively.

At the theoretical level, the current findings are also relevant for an ongoing debate in the dual process field (e.g., De Neys, 2022; Pennycook et al., 2015). Inspired by the two-response findings, recent advances in the dual process literature claim that intuitive responses to classic "bias" tasks are determined by

the interaction of different types of intuitions (e.g., "logical" intuitions, which are founded on an automated knowledge of simple logical rules, and "heuristic" intuitions, Bago & De Neys, 2017). According to this approach, the intuition that is strongest in activation will eventually prevail and the more similar the two competing intuitions are in strength, the less confident the responder will be about their answer (e.g., De Neys, 2022; De Neys & Pennycook, 2019; Pennycook et al., 2015). The findings in the present paper are in line with this view as they suggest that when faced with a risky decision, people generate both a loss averse intuition and an expected-value maximizing intuition.

Investigating the nature of expected-value-based choices is also essential for practical and methodological reasons. In his influential work on the Cognitive Reflection Test (CRT), Frederick (2005) suggests that higher cognitive reflection is associated with more expected-value-based choices. If, however, these choices are often generated intuitively, future studies should be wary of using expected value maximizing choices as an index of one's deliberate reflection capacity per se. Rather, it might more likely reflect the accuracy of one's intuitive processing.

Finally, risk is ubiquitous in investment and managerial business decisions and such decisions are frequently made under time pressure and cognitive load (e.g., traders having to make split-second decisions under stress, Lo & Repin, 2001). If intuition is often sufficient for the generation of profit-maximizing choices, then the idea that lack of deliberation and time-pressure is necessarily detrimental for financial and administrative business decisions might be reconsidered (Kahneman, 2011; McAfee, 2012; World Bank Group, 2015). This also may be linked to evidence suggesting that people higher in cognitive capacity have more accurate intuitions (e.g., Reyna & Brainerd, 2011; Thompson et al., 2018) and that top traders have superior intuitive processing skills (e.g., Kandasamy et al., 2016). Likewise, Reyna and colleagues (e.g., Reyna, 2012) have long stressed the importance of intuitive processing for optimal decision making and cognitive functioning. Their fuzzy-trace theory suggests that cognitive processing can switch from verbatim to more intuitive, gist-based representations (see Reyna et al., 2017 for a review). Although our findings do not inform us on the underlying representations, they agree with the central claim of the fuzzy-trace theory, which is the importance of sound intuitive reasoning in human cognition and more specifically risky decisions (Reyna, 2004).

It is worth noting that across our studies there was a non-negligible amount of conflict trials where participants selected the expected value maximizing choice in the initial, intuitive stage and after deliberating opted for the loss averse choice instead. These "10" cases ($M_{proportion}$ = 9.3% in the betting game; $M_{proportion}$ = 14.6% in the lottery game) are the main reason why the proportion of expected value maximizing choices is higher in the initial, intuitive than in the final, deliberate stage. This pattern seems contradictory since deliberation is thought, if anything, to make people more likely to consider logico-probabilistic principles like expected value (e.g., Slovic et al., 2005). Indeed, most two-response studies in the reasoning field that used classic "bias" tasks, found that "10" trials were negligible (see Boissin et al., 2021 for bat-and-ball problems; see Bago & De Neys, 2017 for base-rate neglect problems and syllogisms). However, in line with our findings, some studies researching the conjunction fallacy found that people also tended to provide more logical responses intuitively than after deliberation (Boissin et al., 2022; Dujmović et al., 2021; Voudouri et al., 2022). The authors suggested that the processing of the biased response might require some minimal deliberation (Dujmović et al., 2021). In other words, at the constrained intuitive stage, the biased response might not yet be fully activated and might need some time and resources to reach its peak strength (Bago & De Neys, 2017; Pennycook et al., 2015). In line with the theoretical model mentioned above, we can assume that in the initial stage both the loss averse and the expected value maximizing intuitions get activated at various speeds. If the latter is stronger it will be selected as the initial response. However, in the deliberate stage, the loss averse intuition might peak in strength and prevail as the final response (see De Neys, 2022; Pennycook et al., 2015).

To conclude, the present paper shows that expected value maximizing choices in risky decision making are most of the time the result of mere intuitive processing and not that of effortful deliberation. Consistent with recent advances in dual process theorizing this suggest that risky decision making may be better conceptualized as an interplay between different types of "fast" intuitions rather than two different types of "fast" and "slow" thinking per se.

References

Andersson, O., Holm, H. J., Tyran, J.-R., & Wengström, E. (2016). Risk Aversion Relates to Cognitive Ability: Preferences or Noise? *Journal of the European Economic Association*, *14*(5), 1129–1154. https://doi.org/10.1111/jeea.12179

Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition*, *158*, 90–109. https://doi.org/10.1016/j.cognition.2016.10.014

Bago, B., & De Neys, W. (2019). The Smart System 1: Evidence for the intuitive nature of correct responding on the bat-and-ball problem. *Thinking & Reasoning*, *25*(3), 257–299. https://doi.org/10.1080/13546783.2018.1507949

Bago, B., & De Neys, W. (2020). Advancing the specification of dual process models of higher cognition: A critical test of the hybrid model view. *Thinking & Reasoning*, *26*(1), 1–30. https://doi.org/10.1080/13546783.2018.1552194

Ben Zur, H., & Breznitz, S. J. (1981). The effect of time pressure on risky choice behavior. *Acta Psychologica*, *47*(2), 89–104. https://doi.org/10.1016/0001-6918(81)90001-9

Boissin, E., Caparos, S., Raoelison, M., & De Neys, W. (2021). From bias to sound intuiting: Boosting correct intuitive reasoning. *Cognition*, *211*, 104645. https://doi.org/10.1016/j.cognition.2021.104645

Boissin, E., Caparos, S., Voudouri, A., & De Neys, W. (2022). Debiasing System 1: Training favours logical over stereotypical intuiting. *Judgment and Decision Making*, *17*(4), 646–690. https://doi.org/10.1017/S1930297500008895

Burič, R., & Konrádová, ľubica. (2021). Mindware Instantiation as a Predictor of Logical Intuitions in Cognitive Reflection Test. *Studia Psychologica*, *63*(2), 114–128. https://doi.org/10.31577/sp.2021.02.822

Burič, R., & Šrol, J. (2020). Individual differences in logical intuitions on reasoning problems presented under two-response paradigm. *Journal of Cognitive Psychology, 32*(4), 460–477. https://doi.org/10.1080/20445911.2020.1766472

Camerer, C. (2005). Three Cheers—Psychological, Theoretical, Empirical—For Loss Aversion. *Journal of Marketing Research*, *42*(2), 129–133. https://doi.org/10.1509/jmkr.42.2.129.622

Croson, R., & Gneezy, U. (2009). Gender Differences in Preferences. *Journal of Economic Literature*, *47*(2), 448–474. https://doi.org/10.1257/jel.47.2.448

De Neys, W. (2017). Bias, conflict, and fast logic: Towards a hybrid dual process future? In W. De Neys (Ed.), *Dual Process Theory 2.0*. Oxon, UK: Routledge.

De Neys, W. (2022). Advancing theorizing about fast-and-slow thinking. *Behavioral and Brain Sciences*, 1–68. https://doi.org/10.1017/S0140525X2200142X

De Neys, W., Novitskiy, N., Geeraerts, L., Ramautar, J., & Wagemans, J. (2011). Cognitive Control and Individual Differences in Economic Ultimatum Decision-Making. *PLOS ONE*, *6*(11), e27107. https://doi.org/10.1371/journal.pone.0027107

De Neys, W., & Pennycook, G. (2019). Logic, Fast and Slow: Advances in Dual-Process Theorizing. *Current Directions in Psychological Science*, *28*(5), 503–509. https://doi.org/10.1177/0963721419855658

De Neys, W., & Schaeken, W. (2007). When people are more logical under cognitive load: Dual task impact on scalar implicature. *Experimental psychology*, *54*(2), 128–133. https://doi.org/10.1027/1618-3169.54.2.128

De Neys, W., & Verschueren, N. (2006). Working Memory Capacity and a Notorious Brain Teaser: The Case of the Monty Hall Dilemma. *Experimental Psychology*, *53*(2), 123–131. https://doi.org/10.1027/1618-3169.53.1.123

Deck, C., & Jahedi, S. (2015). The effect of cognitive load on economic decision making: A survey and new experiments. *European Economic Review*, *78*, 97–119. https://doi.org/10.1016/j.euroecorev.2015.05.004

Drichoutis, A. C., & Nayga, R. M., Jr. (2020). Economic Rationality under Cognitive Load. *The Economic Journal*, *130*(632), 2382–2409. https://doi.org/10.1093/ej/ueaa052

Dror, I. E., Basola, B., & Busemeyer, J. R. (1999). Decision making under time pressure: An independent test of sequential sampling models. *Memory & Cognition*, *27*(4), 713–725. https://doi.org/10.3758/BF03211564

Dujmović, M., Valerjev, P., & Bajšanski, I. (2021). The role of representativeness in reasoning and metacognitive processes: An in-depth analysis of the Linda problem. *Thinking & Reasoning*, *27*(2), 161–186. https://doi.org/10.1080/13546783.2020.1746692

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Filippin, A., & Crosetto, P. (2016). A Reconsideration of Gender Differences in Risk Attitudes. *Management Science*, *62*(11), 3138–3160. https://doi.org/10.1287/mnsc.2015.2294

Franssens, S., & De Neys, W. (2009). The effortless nature of conflict detection during thinking. *Thinking & Reasoning*, *15*(2), 105–128. https://doi.org/10.1080/13546780802711185

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Freeman, N., & Muraven, M. (2010). Self-Control Depletion Leads to Increased Risk Taking. *Social Psychological and Personality Science*, *1*(2), 175–181. https://doi.org/10.1177/1948550609360421

Gerhardt, H., Biele, G. P., Heekeren, H. R., & Uhlig, H. (2016). *Cognitive load increases risk aversion*. SFB 649 Discussion Paper.

Ghasemi, O., Handley, S. J., & Howarth, S. (2023). Illusory intuitive inferences: Matching heuristics explain logical intuitions. *Cognition, 235*, 105417. https://doi.org/10.1016/j.cognition.2023.105417

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences*, *24*(3), 383–403. https://doi.org/10.1017/S0140525X01004149

Holt, C. A., & Laury, S. K. (2002). Risk Aversion and Incentive Effects. *The American Economic Review*, *92*(5), 1644–1655. https://doi.org/10.1257/000282802762024700

Kahneman, D. (2003). Maps of Bounded Rationality: Psychology for Behavioral Economics. *The American Economic Review*, *93*(5), 1449–1475. https://doi.org/10.1257/000282803322655392

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

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgement. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), H*euristics and biases: The psychology of intuitive judgment* (pp. 49–81). Cambridge, MA: Cambridge University Press.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263–291. https://doi.org/10.2307/1914185

Kandasamy, N., Garfinkel, S. N., Page, L., Hardy, B., Critchley, H. D., Gurnell, M., & Coates, J. M. (2016). Interoceptive Ability Predicts Survival on a London Trading Floor. *Scientific Reports*, *6*(1). https://doi.org/10.1038/srep32986

Keysar, B., Hayakawa, S. L., & An, S. G. (2012). The Foreign-Language Effect: Thinking in a Foreign Tongue Reduces Decision Biases. *Psychological Science*, *23*(6), 661–668. https://doi.org/10.1177/0956797611432178

Kocher, M. G., Pahlke, J., & Trautmann, S. T. (2013). Tempus Fugit: Time Pressure in Risky Decisions. *Management Science*, *59*(10), 2380–2391. https://doi.org/10.1287/mnsc.2013.1711

Lo, A. W., & Repin, D. V. (2002). The psychophysiology of real-time financial risk processing. *Journal of cognitive neuroscience*, *14*(3), 323–339. https://doi.org/10.1162/089892902317361877

Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. *Psychological bulletin*, *127*(2), 267. https://doi.org/10.1037/0033-2909.127.2.267

Madan, C. R., Spetch, M. L., & Ludvig, E. A. (2015). Rapid makes risky: Time pressure increases risk seeking in decisions from experience. *Journal of Cognitive Psychology*, *27*(8), 921–928. https://doi.org/10.1080/20445911.2015.1055274

McAfee, A. (2012, February 8). Managerial Intuition Is a Harmful Myth. *Harvard Business Review*. https://hbr.org/2012/02/managerial-intuition-is-a-harm

Mechera-Ostrovsky, T., Heinke, S., Andraszewicz, S., & Rieskamp, J. (2022). Cognitive abilities affect decision errors but not risk preferences: A meta-analysis. *Psychonomic Bulletin & Review*, *29*, 1719–1750. https://doi.org/10.3758/s13423-021-02053-1

Meyer-Grant, C. G., Cruz, N., Singmann, H., Winiger, S., Goswami, S., Hayes, B. K., & Klauer, K. C. (2022). Are logical intuitions only make-believe? Reexamining the logic-liking effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. https://doi.org/10.1037/xlm0001152

Olschewski, S., & Rieskamp, J. (2021). Distinguishing three effects of time pressure on risk taking: Choice consistency, risk preference, and strategy selection. *Journal of Behavioral Decision Making*, *34*(4), 541–554. https://doi.org/10.1002/bdm.2228

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2014). Cognitive style and religiosity: The role of conflict detection. *Memory & Cognition*, *42*, 1-10. https://doi.org/10.3758/s13421-013-0340-7

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. https://doi.org/10.1016/j.cogpsych.2015.05.001

Raoelison, M., & De Neys, W. (2019). Do we de-bias ourselves?: The impact of repeated presentation on the bat-and-ball problem. *Judgment and Decision making*, *14*(2), 170–178. https://doi.org/10.1017/S1930297500003405

Raoelison, M., Thompson, V. A., & De Neys, W. (2020). The smart intuitor: Cognitive capacity predicts intuitive rather than deliberate thinking. *Cognition*, *204*, 104381. https://doi.org/10.1016/j.cognition.2020.104381

Reyna, V. F. (2004). How People Make Decisions That Involve Risk: A Dual-Processes Approach. *Current Directions in Psychological Science*, *13*(2), 60–66. https://doi.org/10.1111/j.0963-7214.2004.00275.x

Reyna, V. F. (2012). A new intuitionism: Meaning, memory, and development in Fuzzy-Trace Theory. *Judgment and Decision Making*, *7*(3), 332–359. https://doi.org/10.1017/S1930297500002291

Reyna, V. F., & Brainerd, C. J. (2011). Dual Processes in Decision Making and Developmental Neuroscience: A Fuzzy-Trace Model. *Developmental Review, 31*(2–3), 180–206. https://doi.org/10.1016/j.dr.2011.07.004

Reyna, V. F., Rahimi-Golkhandan, S., Garavito, D. M. N., & Helm, R. K. (2017). The fuzzy-trace dual-process model. In W. De Neys (Ed.), *Dual Process Theory 2.0* (pp. 90–107). Routledge.

Sirota, M., Dewberry, C., Juanchich, M., Valuš, L., & Marshall, A. C. (2021). Measuring cognitive reflection without maths: Development and validation of the verbal cognitive reflection test. *Journal of Behavioral Decision Making*, *34*(3), 322-343. https://doi.org/10.1002/bdm.2213

Sirota, M., Juanchich, M., & Holford, D. L. (2023). Rationally irrational: When people do not correct their reasoning errors even if they could. *Journal of Experimental Psychology: General*. Advance online publication. https://doi.org/10.1037/xge0001375

Slovic, P., Peters, E., Finucane, M. L., & MacGregor, D. G. (2005). Affect, risk, and decision making. *Health psychology*, *24*(4S), S35. https://doi.org/10.1037/0278-6133.24.4.S35

Stupple, E. J., & Ball, L. J. (2008). Belief–logic conflict resolution in syllogistic reasoning: Inspection-time evidence for a parallel-process model. *Thinking & Reasoning, 14*(2), 168-181. https://doi.org/10.1080/13546780701739782

Stupple, E. J., Ball, L. J., Evans, J. S. B., & 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

Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning*, *20*(2), 215–244. https://doi.org/10.1080/13546783.2013.869763

Thompson, V. A., Pennycook, G., Trippas, D., & Evans, J. St. B. T. (2018). Do smart people have better intuitions? *Journal of Experimental Psychology: General*, *147*(7), 945–961. https://doi.org/10.1037/xge0000457

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*(3), 107–140. https://doi.org/10.1016/j.cogpsych.2011.06.001

Voudouri, A., Białek, M., Domurat, A., Kowal, M., & De Neys, W. (2022). Conflict detection predicts the temporal stability of intuitive and deliberate reasoning. *Thinking & Reasoning*, 1–29. https://doi.org/10.1080/13546783.2022.2077439

World Development Report 2015: Mind, Society, and Behavior. (n.d.). Retrieved 15 March 2023, from https://www.worldbank.org/en/publication/wdr2015

Chapter 2

Semantic illusions, fast and slow

Beucler, J., **Voudouri, A.**, & De Neys, W. (under review). Semantic illusions, fast and slow. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Supplementary material for this chapter can be found in Supplementary material for Chapter 2.

Abstract

When asked "How many animals of each kind did Moses take on the Ark?", most people answer "Two", failing to notice that it was Noah, and not Moses, who took the animals in the Ark. Traditional "fast-and-slow" dual process accounts of such semantic illusions posit that incorrect reasoners are not sensitive to their error and that overcoming the illusion requires deliberate correction of an intuitive erroneous answer. We present three studies that force us to revise this longstanding dual process view. We used a two-response paradigm in which participants had to give their first, initial answer under cognitive load and time pressure. Next, participants could take all the time they wanted to deliberate and select a final answer. This enabled us to identify the intuitively generated response that preceded the final response given after deliberation. Results show that participants do not necessarily need to deliberate to avoid the illusion and that incorrect respondents consistently display error sensitivity (as reflected in decreased confidence), even when deliberation is minimized. Both reasoning performance and error sensitivity in the initial, intuitive stage were driven by the semantic relatedness between the anomalous word (e.g., "Moses") and the undistorted word (e.g., "Noah"). We show how this leads to a revised model where the response to semantic illusions depends on the interplay of both incorrect and correct intuitions.

Introduction

When asked "How many animals of each kind did Moses take on the Ark?", most people answer "Two", failing to notice that it was Noah, and not Moses, who took the animals in the Ark (Erickson & Mattson, 1981). This tendency to overlook a semantic anomaly in a sentence is known as a *semantic illusion* (or the *Moses illusion*—after its most famous example). It is a very robust effect that attracted a lot of attention in the memory field and beyond (e.g., Cantor & Marsh, 2017; Hannon & Daneman, 2001; Kamas et al., 1996; Park & Reder, 2004; Reder & Kusbit, 1991; Shafto & MacKay, 2000; Speckmann & Unkelbach, 2021).

A key driving factor in the emergence of the illusion is the semantic relatedness between the anomalous (or distorted) word (e.g., "Moses") and the correct or undistorted word (e.g., "Noah," Erickson & Mattson, 1981; Hannon & Daneman, 2001; Van Oostendorp & De Mul, 1990). In particular, Noah and Moses share many semantic attributes such as biblical character, male, leader, and so on. The anomalous word will thus serve as an "impostor" and go unnoticed because of how semantically similar it is to "Noah". Indeed, if the semantic similarity between the distorted and undistorted word is low (e.g., "How many animals of each kind did Nixon take on the Ark?"), participants are much less likely to fall prey to the illusion and will respond correctly that the question is anomalous and cannot be answered (e.g., Erickson & Mattson, 1981; Hannon & Daneman, 2001; Van Oostendorp & De Mul, 1990).

More generally, semantic illusions seem a key example of the human tendency for miserly processing or satisficing such as it has been put forward in the dual process framework (Kahneman, 2011; Koriat, 2017; Stanovich & West, 2000). This influential framework conceives human cognition as an interplay of fast and effortless, intuitive ("System 1") processing, and slower, more effortful deliberate ("System 2") processing (Evans & Stanovich, 2013; Kahneman, 2011). Although fast intuitive processing may often cue valid responses, it can sometimes also cue responses that conflict with the slower, deliberate processing and will need to be corrected. However, because people will typically try to minimize spending cognitive effort, they will often refrain from engaging the effortful processing. Consequently, they will fail to detect that their intuitively cued response is erroneous and end up with a biased judgment (Evans & Stanovich, 2013; Kahneman, 2011).

In the case of semantic illusions, Park and Reder (2004) argued, for example, that people rely on an automatic partial matching mechanism that focuses on the coarse fit between a memory trace and the presented sentence. As long as there is sufficient semantic overlap, people will not engage in a more effortful in-depth analysis. Although the fast matching mechanism might often be useful for quick sentence comprehension, it can also give rise to semantic illusions. Hence, semantic illusions seem like a paradigmatic case of a more general human failure to switch from fast intuitive to slow deliberate processing when it is needed (Koriat, 2017; two). It should be no surprise then that semantic illusions also feature in widely used Cognitive Reflection Tests that are intended to measure people's capacity and disposition to engage in effortful deliberation rather than to stick to a mere intuitive hunch (e.g., Sirota et al., 2021).

At first sight, the dual process account of semantic illusions does not seem unreasonable. From an introspective point of view, many of us will have fallen for the illusion and attest that spotting it requires taking more time for deeper reflection and having a closer, second look. There is also some empirical evidence that is consistent with the account. For example, several lines of research have demonstrated the importance of cognitive resources to avoid the illusion. Experiments using a concurrent cognitive load have shown that burdening participants' cognitive resources increases the likelihood of falling prey to the illusion (Büttner, 2012; Mata et al., 2013). In addition, working memory capacity is positively correlated with the ability to detect the illusion, suggesting that cognitive resources are needed to overcome it (Hannon & Daneman, 2001).

Nevertheless, the available evidence does not tell us whether deliberation is always necessary to correctly detect the anomaly and avoid the illusion. In theory, it is possible that in addition to the slow route there is also a fast route to anomaly detection in which the correct answer is cued intuitively. That is, rather than correcting an incorrect hunch (i.e., "Two") after having taken the time to deliberate, people might generate the correct answer from the outset. Clearly, if one intuitively detects the anomaly and avoids the illusion, there is no further need to deliberately correct it. Obviously, such a fast route would imply that it would be problematic to use people's performance on semantic illusion items as a measure of cognitive reflection or deliberation.

Recent dual process studies in the reasoning field lend some credence to this theoretical possibility (e.g., see De Neys & Pennycook, 2019). In these studies,

participants solve logico-mathematical "bias" problems in which intuitive processing can lead them astray (e.g., the notorious bat-and-ball problem, "A bat and a ball together cost $1.10. The bat costs $1 more than the ball. How much does the ball cost?"; correct answer: "5 cents", incorrect intuitive answer: "10 cents"). To isolate more intuitively and deliberately generated responses the studies use a two-response paradigm (Thompson et al., 2011), in which participants have to initially give the first response that comes to mind as quickly as possible, and are subsequently given all the time they want to deliberate and select a final response. To be maximally sure that participants do not engage in deliberation in the initial stage, they are forced to give their first response under time-pressure while performing a concurrent cognitive load task which burdens their cognitive resources (Bago & De Neys, 2017). Since deliberation takes time and cognitive resources, depriving people from these resources minimizes the possibility that reasoners will deliberate before giving their initial response. The traditional corrective dual process view predicts that correct responses will only emerge after deliberation in the final response stage (Kahneman, 2011). However, contrary to this view, results across a range of reasoning problems have now shown that on those trials where participants manage to give the correct response after deliberation, they often already generate the same correct response during the initial response stage (e.g., Bago & De Neys, 2017, 2019a; Burič & Konrádová, 2021; Burič & Šrol, 2020; Raoelison et al., 2020; Thompson & Johnson, 2014). Hence, sound reasoners are not necessarily good at deliberately correcting erroneous intuitions, but rather at accurate intuiting (Raoelison et al., 2020; Thompson et al., 2018).

Further evidence against the postulated role of deliberation during reasoning comes from intuitive error detection findings (e.g., De Neys & Pennycook, 2019). Just as with the dual process view on semantic illusions, it is traditionally assumed that detecting that an intuitively cued solution is incorrect, requires that people engage in effortful deliberation (Evans & Stanovich, 2013; Kahneman, 2011). Contrary to this assumption, however, it has been found that reasoners who fail to generate the correct response to logical bias problems often show some intuitive sensitivity to their error (e.g., De Neys et al., 2011, Stupple & Ball, 2008, Stupple et al., 2011, Voudouri et al., 2022; but see also Mata et al., 2014, Mata et al., 2017). For example, they show lower confidence in their erroneous responses than in their correct responses to control problems. Critically,

this error sensitivity is also observed in the initial stage of the two-response studies (i.e., when deliberation is minimized with time-pressure and load, e.g., Bago & De Neys, 2017, 2019b; Bialek & De Neys, 2017; Burič & Konrádová, 2021; Burič & Šrol, 2020; Johnson et al., 2016; Pennycook et al., 2014; Thompson & Johnson, 2014). Hence, even in the absence of proper deliberation, participants seem to be able to detect that their erroneous answer is not fully warranted.

Taken together, recent findings in the logical reasoning field suggest that the traditional role of deliberation in the dual process framework can be questioned (De Neys, 2022; Evans, 2019; Stanovich, 2018). If we were to generalize this pattern to semantic illusions, this suggests that detecting the anomaly and avoiding falling prey to semantic illusions may also be done intuitively and might not necessarily require slow and effortful deliberation. This would have critical implications for our conceptualization of semantic illusions and their use as an index of deliberate processing abilities.

In the present studies we test this hypothesis directly by introducing a two-response paradigm of semantic illusions. Participants were presented with a range of trivia questions that are known to elicit semantic illusions. Half of the problems were presented in an undistorted format (e.g., "In the tale, who found the glass slipper left at the ball by *Cinderella*?") and served as control problems on which intuitive processing is expected to cue the correct response. The other half of the problems were classic "anomaly" problems (e.g., "In the tale, who found the glass slipper left at the ball by *Snow White*?") in which the undistorted word was replaced by a semantically related distorted "impostor" word which may give rise to a semantic illusion. Participants had to give an initial response as fast as possible (under time pressure and concurrent load) and immediately after could take the time to deliberate and give a final response. Participants also indicated their response confidence. Our key research questions were: First, whether people who answer anomaly problems correctly and avoid the illusion after deliberation, can also provide a correct response to these questions intuitively. Second, whether people who give an incorrect response to anomaly problems in the intuitive response stage, show error sensitivity (i.e., by contrasting their response confidence in the anomaly and control no-anomaly problems).

We present a set of three studies: Study 1 introduces the paradigm, Study 2 tests the robustness of the results with methodological refinements, and Study 3

introduces a direct manipulation of the semantic similarity factor to validate the findings.

Study 1

Method

Open Science and Data

The research question and study design were preregistered on the OSF platform (https://osf.io/bpmc8). No specific analyses were preregistered. All data, material and analysis scripts can be retrieved from https://osf.io/bvy3u/.

Participants

In Study 1, we recruited 100 participants (78 females, *M* age = 32.6 years, *SD* = 12.2 years) on the Prolific platform (app.prolific.co). Only native English speaking American (USA) participants were allowed to participate in the study. They were paid at a rate of £5 per hour for their participation. Among them, 42 reported high school as their highest level of education, 1 less than high school and 57 a higher education degree.

We based our sample size decision for Study 1, 2 and 3 on previous two-response work in the logical, moral, and economic reasoning field (Bago et al., 2021; Bago & De Neys, 2017, 2019a), which also tested approximately 100 participants. Note that this sample size gives us more power than most previous studies on semantic illusions (e.g., Kamas et al., 1996).

Materials

Trivia Questions. We selected 40 multiple-choice trivia question problems from the second experiment of Speckmann and Unkelbach (2021). Using the results of their knowledge-check (i.e., open-ended questions such as: "Which biblical figure took two animals of each kind on the Ark?" on 200 participants), we selected the questions that were above or closest to the sample median accuracy (Mdn = 74.4%). In addition, we discarded two items that had a low control no-anomaly accuracy in the actual study of Speckmann and Unkelbach (2021, Experiment 2), and replaced them with the questions that were closest to the knowledge check median accuracy. This helped to guarantee that on average our participants would know the correct answer to the original questions. We also introduced some

superficial content modifications to minimize question length differences. Supplementary Material A provides the complete list of anomaly and no-anomaly questions.

For each of the selected questions we created an anomaly version (i.e., "In the biblical story, how many animals of each kind did Moses take on the Ark?") and a control, no anomaly version (i.e., "In the biblical story, how many animals of each kind did Noah take on the Ark?")[1]. The control version used the original, undistorted word (e.g., "Noah") whereas the anomaly version used the semantically related "impostor" word (e.g., "Moses") as in Speckmann and Unkelbach (2021). Half of the 20 problems that each participant saw were anomaly problems and the other half control problems. Two question sets were created for counterbalancing. For each question, the control version was used in one set and the anomaly version in the other set. Participants were randomly assigned to one of the sets. Hence, participants never saw the same question content more than once. The presentation order of the questions was randomized in both sets.

Following Speckmann and Unkelbach (2021), each question had four different response options. The first option was the "undistorted" answer (e.g., "two" for the Moses question) and could be correct or incorrect depending on the question version (no-anomaly vs. anomaly). The second option (e.g., "three") was always incorrect. The third response option was "This question can't be answered in this form", and could be correct or incorrect depending on the question version (anomaly vs. no-anomaly). The fourth option was "Don't know", which was always coded as incorrect. The order of options 1 and 2 was randomized, but we kept the order of response options 3 and 4 fixed so as not to confuse participants. Note that the use of a multiple-choice (vs. open-ended question) design with these specific response options was tested and validated across four experiments by Speckmann and Unkelbach (2021). Semantic illusions were as prevalent in the multiple choice design as in previous open-ended studies. In addition, to be sure that participants understood the difference between the "Don't know" and the "This question can't be answered in this form" response options, the following examples were presented in the instructions:

[1] Note that in our preregistration we referred to conflict and no-conflict problems (in analogy with the reasoning field). Here we have opted for the more descriptive anomaly and (control) no-anomaly labels.

What is the name of former president's Obama's oldest son?

Charles
Jonathan
This question can't be answered in this form.
Don't know.

The above question cannot be answered because Obama doesn't have a son; he only has two daughters. So, the correct answer option to this question is: 'This question can't be answered in this form.'

Here is a different example:

What is the name of former president's Obama's oldest daughter?

Sasha
Malia
This question can't be answered in this form.
Don't know.

In the above example, the question can be answered, since Obama does have an oldest daughter. The correct answer option is 'Malia'. However, if you do not know the answer to this question, you should select 'Don't know'.

Cognitive Load Task. The use of cognitive resources has been advanced as a key feature of System 2 deliberation (Evans & Stanovich, 2013). Thus, to help prevent deliberation in the initial stage of our two-response paradigm, we imposed a concurrent cognitive load to participants during the trivia question answering. We used the dot memorization task (Miyake et al., 2001), which has been shown to successfully burden executive resources during verbal reasoning (e.g., De Neys & Schaeken, 2007; De Neys & Verschueren, 2006; Verschueren et al., 2004).

Before each trivia question, participants were presented with a 3 x 3 grid, in which four crosses were placed (Figure 1b). Participants were told that it was essential to memorize the location of the crosses while answering the questions. After their initial response, participants were shown four different matrices and they had to select the correct, to-be-memorized pattern. They then received feedback as to whether they chose the correct pattern. The load was only present during the initial response stage and not during the subsequent final response stage in which participants were allowed to deliberate (see further).

Procedure

One-Response Pre-test. To determine an appropriate deadline for the two-response paradigm we ran a traditional one-response version of the experiment without deadline or load (e.g., see Bago et al., 2021). The same material as in the main two-response study was used but participants only had to give a single answer for which they had all the time they wanted to deliberate, without any concurrent load. We recruited an independent sample of 50 online native English speaking American (USA) participants (40 females, *M* age = 27 years, *SD* = 10 years) on the Prolific platform.

Results indicated that participants took on average 7.2 s (*SD* = 3.2 s) to provide correct responses to the anomaly problems (i.e., to respond "This question can't be answered in this form"). In Study 1, we set the initial deadline at 5 s (see further), which corresponded to the first quartile of the correct, anomaly response latency of the one-response pre-test. To test whether participants were under time pressure in the initial stage of the two-response paradigm, we contrasted latencies for anomaly correct responses in the one-response pre-test and in the initial stage of the main two-response study. Participants responded significantly faster in the initial two-response stage (*M* = 3.8 s) than in the one-response pre-test (*M* = 7.2 s), $W = 104$, $p < .001$, $r = -.79$.

The one-response pre-test also allowed us to check for a possible consistency confound in the two-response study. When people are asked to give two consecutive responses to the same question, they might want to appear consistent in their responses. This may prevent participants to correct their initial response, which would lead to an underestimation of the true correction rate and to a lower accuracy in the final stage of the two-response study. However, the results show that participants had virtually the same accuracy in the one-response pre-test (*M* = 63.9%, *SD* = 17.6%) and in the final stage of the two-response paradigm (*M* = 64.9%, *SD* = 16.4%), $t(92.45) = -0.32$, $p = .75$. This directly argues against a possible consistency confound in the two-response paradigm.

Two-Response Paradigm. We used a procedure similar to Bago and De Neys (2017). The experiment was run online on the Qualtrics platform. The task was introduced with the following instructions:

Please read these instructions carefully!

In this experiment you will have to respond to 20 multiple-choice trivia questions and a couple of practice questions.

For every multiple choice question you will be presented with four answer options but **you can only pick one answer. Please respond as accurately as you can.**

Some of the questions are impossible to answer. In that case, select the answer option: "This question can't be answered in this form."

If you don't know the answer to a question, select the response option "Don't know".

Then, two examples were given to clarify the difference between the "Don't know" and the "This question can't be answered in this form" response options (see above). After this general introduction, participants were presented with a more specific instruction page about the procedure itself:

Critically, in this study we want to know what your **initial, intuitive response** to the questions is and **how you respond after you have thought about these questions for some more time.**

First, we want you to respond with the **very first answer that comes to mind**. You don't need to think about it. Just give the first answer that intuitively comes to mind as quickly as possible.

To make sure that you answer as fast as possible, **a time limit was set for the first response**, which is going to be 5 seconds. When there is 1 second left, the background colour will turn to yellow to let you know that the deadline is approaching. Please make sure to **answer before the deadline passes.**

Next, **the question will be presented again** and you can take all the time you want to actively reflect on it. Once you have made up your mind you give your **final response.**

After you made your choice and clicked on it, you will be automatically taken to the next page.

After you have entered your first and final answer we will also ask you to indicate your confidence in the correctness of your response.

Participants then responded to two practice trivia questions to familiarize themselves with the deadline procedure. Next, they solved two practice load

matrix problems without concurrent questions. Finally, at the end of the practice, they had to respond to the two earlier trivia questions using the complete two-response procedure (i.e., with deadline and load).

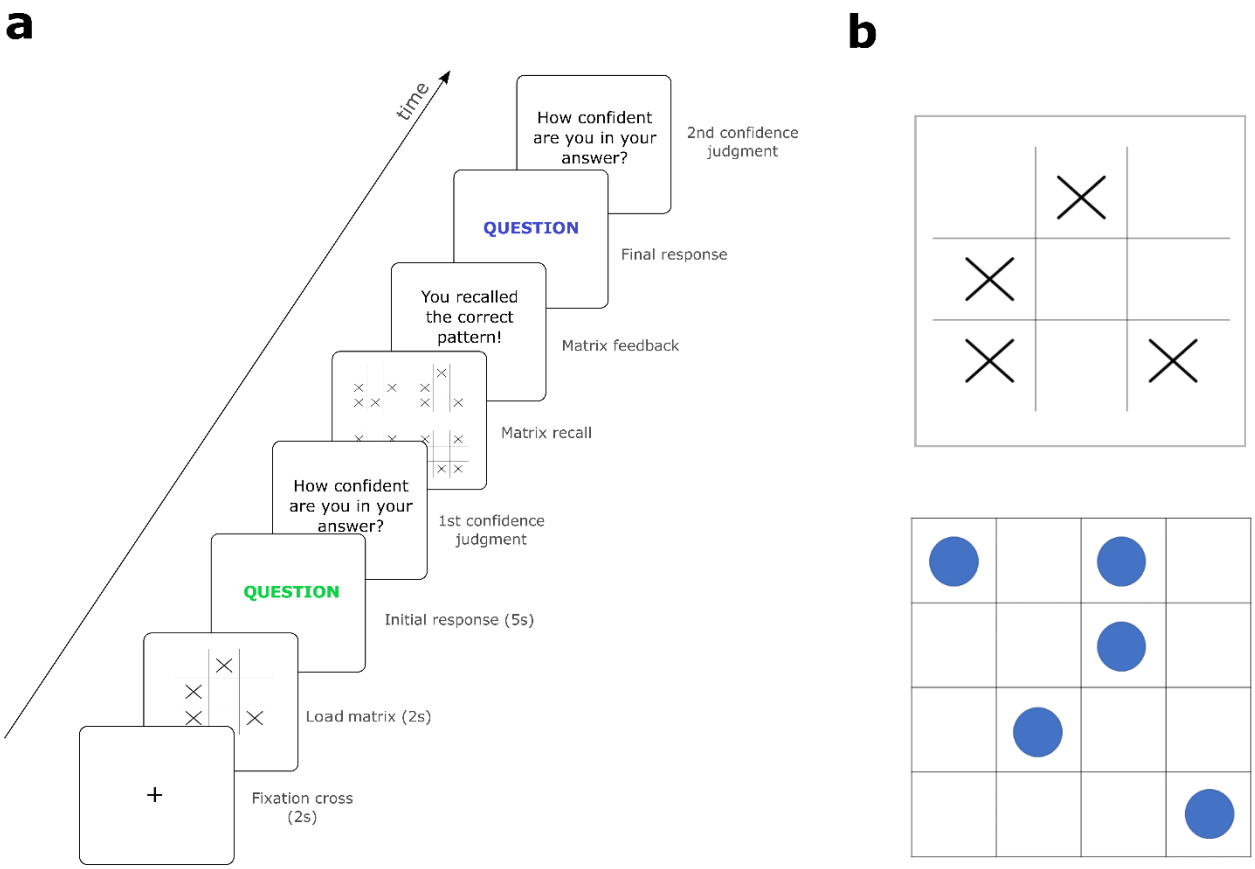

Figure 1. Study 1 trial sequence and examples of load patterns in Study 1-3. **a**) Example of one trial in Study 1. Participants had to respond to a trivia question twice, once with a deadline and a concurrent load and a second time without any constraint. **b**) Example of the to-be-memorized load patterns in Study 1 (upper panel) and Study 2-3 (lower panel).

Figure 1a shows a complete trial sequence for Study 1. At the beginning of each trial, a fixation cross was presented for 2 s. The load matrix was then presented for 2 s. After that, the question appeared. Participants had 5 s to respond. After 4 s, the background of the screen turned yellow to warn participants about the upcoming deadline. If they failed to provide an answer before the deadline, they were reminded to speed up and give an answer within the deadline on the next initial trials.

After the initial response, the question disappeared from the screen and participants had to enter their confidence in their response on a scale ranging from 0% to 100%, with the following instructions: "How confident are you in your

answer? Please type a number from 0 (absolutely not confident) to 100 (absolutely confident)". Then, participants had to select the to-be-memorized load pattern between four different load matrices. If they failed to select the correct pattern, they were reminded to make sure to remember the pattern correctly on the next trials.

The trivia question was then presented a second time, and participants could give their final response without any deadline, nor concurrent load. Once they had given their answer, they were automatically taken to the next page where they had to indicate their confidence level in their final answer.

To remind participants which question stage they were answering, the color of the answer options was green during the first response, and blue during the final response phase. In addition, we also added a reminder sentence under the question: "Please indicate your very first, intuitive answer." and "Please give your final answer.", respectively (see Supplementary Material B for the complete instructions).

After they had answered half of the questions, participants were allowed to take a short break. After they finished the experiment, they completed a page with standard demographic questions and were debriefed.

Exclusion Criteria. Participants failed to provide an initial response within the deadline on 6.8% of the trials, and did not recall the correct load pattern on 12.7% of the trials. We removed any of the trials in which the deadline was missed or recall was inaccurate (or both) from our analyses because we cannot be sure that participants did not already deliberate to produce their initial response in these cases. Indeed, if participants did not respond within the deadline, they might have engaged in slow deliberation. Similarly, if they failed the load memorization task, we cannot guarantee that their cognitive resources were successfully burdened by the cognitive load. Therefore, removing the trials that did not meet the inclusion criteria allowed us to be maximally sure that the initial responses were intuitive in nature.

Hence, a total of 18.2% of the trials were excluded, and we thus analyzed 1636 trials out of 2000. On average, each participant contributed a total of 16.4 valid trials (out of 20, *SD* = 2.4 trials).

Statistical Analysis. The data were analyzed using the following R packages: *afex* (Singmann et al., 2015), *emmeans* (Lenth et al., 2019), *ez* (Lawrence & Lawrence, 2016) and *tidyverse* (Wickham et al., 2019).

Results and Discussion

Accuracy

Our first question of interest was whether people who provide a correct response to anomaly trivia questions after deliberation, can also provide a correct response to these questions when reasoning more intuitively in the initial response stage. Figure 2a provides a summary of the initial and final accuracy for the critical anomaly and control no-anomaly problems. For the control questions, participants performed very well both at the initial (M = 89.1%, SD = 14%) and the final (M = 92%, SD = 9.9%) response stages. These results indicate that overall participants knew the correct responses to the control questions and typically managed to generate them intuitively. They also rule out a potential guessing confound, as they show that participants did not respond randomly in the initial response stage. The accuracy was lower for the anomaly questions, both in the initial response stage (M = 20.4%, SD = 22.6%) and in the final response stage (M = 35.9%, SD = 30.6%). In order to distinguish between errors due to lack of knowledge and errors due to a failure to spot the illusion, we looked at the types of errors that were made in the anomaly questions. Results showed that the vast majority of the incorrect responses were mainly due to participants giving the undistorted answer option (e.g. "Two" in the original Moses illusion) both at the initial (M = 86%) and final (M = 85.2%) response stages, and not because of choosing the "Don't know" or the incorrect filler answer options. The results for the anomaly questions thus indicate that our items succeeded in eliciting the Moses illusion. More importantly, they also show that participants managed to give a significant amount of correct answers at the initial response stage for anomaly questions, although they performed better at the final response stage.

To test these results statistically, we conducted a two-way within-subject ANOVA investigating the effect of anomaly presence (control no-anomaly; anomaly) and response stage (initial; final) on accuracy. There was a significant main effect of anomaly on accuracy, $F(1, 99) = 623.82$, $p < .001$, $\eta^2g = .69$., as well as response stage, $F(1, 99) = 93.66$, $p < .001$, $\eta^2g = .05$. Finally, the

difference between initial and final accuracy was higher for anomaly questions compared to control no-anomaly questions, as indicated by the response stage by anomaly interaction, $F(1, 99) = 37.61$, $p < .001$, $\eta^2g = .02$.

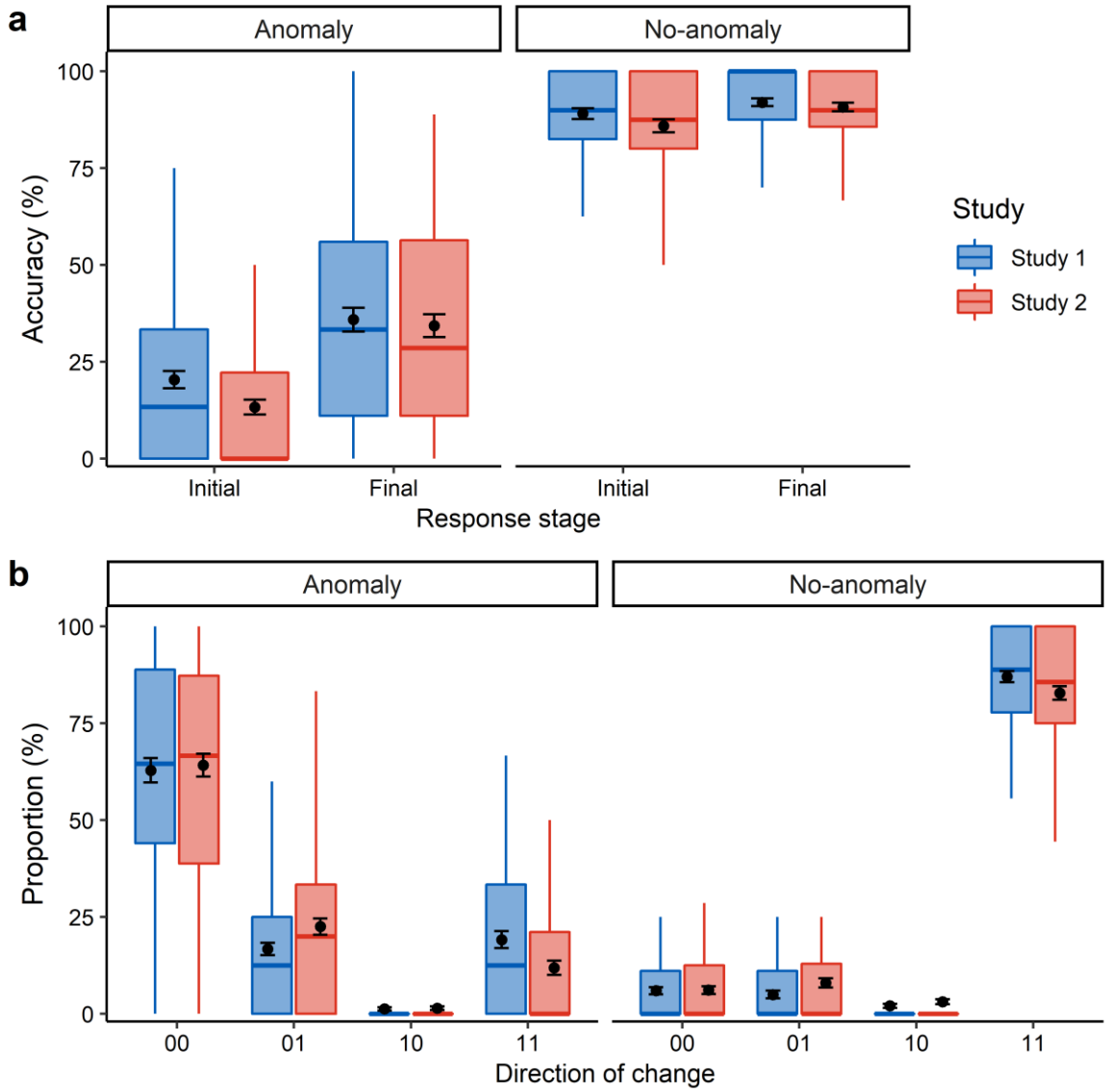

Figure 2. Accuracy and Direction of Change in Study 1 and Study 2. **a**) Response accuracy at anomaly and control no-anomaly trials as a function of response stage. **b**) Proportion of each direction of change category at anomaly and control no-anomaly trials; "00" = incorrect initial and incorrect final response; "01" = incorrect initial and correct final response; "10" = correct initial and incorrect final response; "11" = correct initial and correct final response. The lower and upper hinges of the boxplot correspond to the first and third quartiles, and the middle line shows the median. The lower (resp. upper) whiskers extend from the hinges to the smallest (resp. largest) value no further than 1.5 times the interquartile

range. Overlaid black dots represent the mean and black error bars are standard errors of the mean.

Direction of Change

To better understand how people changed (or did not change) their answers after deliberation, we performed a direction of change analysis by looking into how accuracy changed from the initial to the final response stage on every trial (Bago & De Neys, 2017). For each trial, participants have either an accuracy of "1" (i.e., correct response) or "0" (i.e., incorrect response) in each of the two response stages (i.e., initial and final). Hence, there are four possible directions of change: "00" (incorrect initial and incorrect final response), "01" (incorrect initial and correct final response), "10" (correct initial and incorrect final response) and "11" (correct initial and correct final response).

Figure 2b shows the mean direction of change frequencies as a function of anomaly presence. As expected, the vast majority of control no-anomaly questions yielded "11" responses (87%), followed by "00" (6%), "01" (5%) and "10" responses (2%). Regarding the critical anomaly questions, the majority of "00" responses (62.9%) is consistent with the literature, as it shows that participants tend to fall for the illusion even when they are allowed to deliberate. There were slightly more "11" responses (19.2%) than "01" responses (16.7%), whereas the "10" response pattern was rare (1.2%).

Testing the corrective deliberation assumption of classic dual process accounts of the Moses illusion requires to concentrate on trials where participants gave a correct answer in the final response stage. In order to get a more direct measure of how the proportion of "11" responses compared to that of "01" responses, we computed the mean "non-correction rate" across participants for the anomaly questions (i.e., proportion 11/11+01; Bago & De Neys, 2017). Note that here we computed the non-correction rate for each participant before computing the mean of these individual non-correction rates. We thus excluded the participants who never gave a final correct answer to anomaly questions (i.e., no "11" or "01" response whatsoever; n = 21). This measure indicates the proportion of correct final answers that were already correct in the initial response stage. Put differently, it shows the proportion of trials for which participants did not need to deliberate to find the correct answer. If deliberate correction is critical for correct responding, it should be at 0%. Instead, the mean non-correction rate

for anomaly questions was 46.1% (*SD* = 34.1%). It means that on average, when participants managed to give a correct answer to an anomaly question at the final stage, they already gave a correct answer at the initial stage about half the time. The full distribution of the individual non-correction rates is reported in Supplementary Material E. In summary, although deliberative correction did occur in Study 1, in many cases participants were able to generate the correct response when deliberation was minimized in the initial response stage.

Confidence Ratings

Error Sensitivity. Our second question of interest was to see whether people who give incorrect responses to anomalous trivia questions in the intuitive response stage show some sensitivity to the presence of the anomaly and detect that their answer is questionable. Participants who fall prey to the illusion typically answer with the undistorted response (i.e., "Two" in the Moses Illusion). Whereas this answer is correct for the undistorted control problems, it is obviously incorrect for the distorted anomaly problems. Hence, by contrasting participants' response confidence for correctly solved control no-anomaly problems and incorrectly solved anomaly problems we can test whether they display some basic error sensitivity. If incorrect responders do not register the anomaly, they should not process the two problem versions any differently and should be equally confident about their answers. If incorrect responders show increased doubt when they err, this indicates that they detect that their response is questionable and–despite their incorrect answer–show some minimal sensitivity to the presence of the anomaly.

Figure 3 displays the mean initial confidence ratings for anomaly and control no-anomaly trials as a function of accuracy (i.e., correct vs. incorrect responses). A Wilcoxon signed-ranked test showed that the confidence ratings for anomaly incorrect responses (M = 72.1%) were significantly lower than the confidence ratings for control correct responses (M = 90.2%), $p < .001$, $r = -.7$. Participants thus showed increased response doubt when they were making a mistake which suggests they were detecting that their answer was questionable, even in the initial response stage when deliberate reflection was minimized. The full distribution of the individual initial error sensitivity measures is reported in Supplementary Material E. One may argue that these findings may be better explained by the fact that we coded the (very rare) "Don't know" responses as incorrect (see Method section). As these responses may receive lower confidence

ratings on average, they could be responsible for our confidence findings. To rule out this alternative explanation, we also performed all our confidence analyses without the "Don't know" responses. This did not change the results.

For completeness, note that as Figure 3 indicates, for correct anomaly responses we did not observe a similar confidence decrease when contrasting the control and anomaly problems. In fact, when participants avoided the illusion and responded correctly, they actually gave slightly higher confidence ratings (M = 93.7%) than for the control correct responses (M = 89.8%), p = .005, r = −.26.

Direction of Change. For exploratory purposes we also looked at confidence findings for each of the direction of change categories separately in Supplementary Material C. In line with findings in the reasoning field (e.g., Thompson et al., 2011), results show that initial confidence is lower on trials in which the initial response is changed after deliberation (i.e., "01" and "10" vs. "11" and "00" categories).

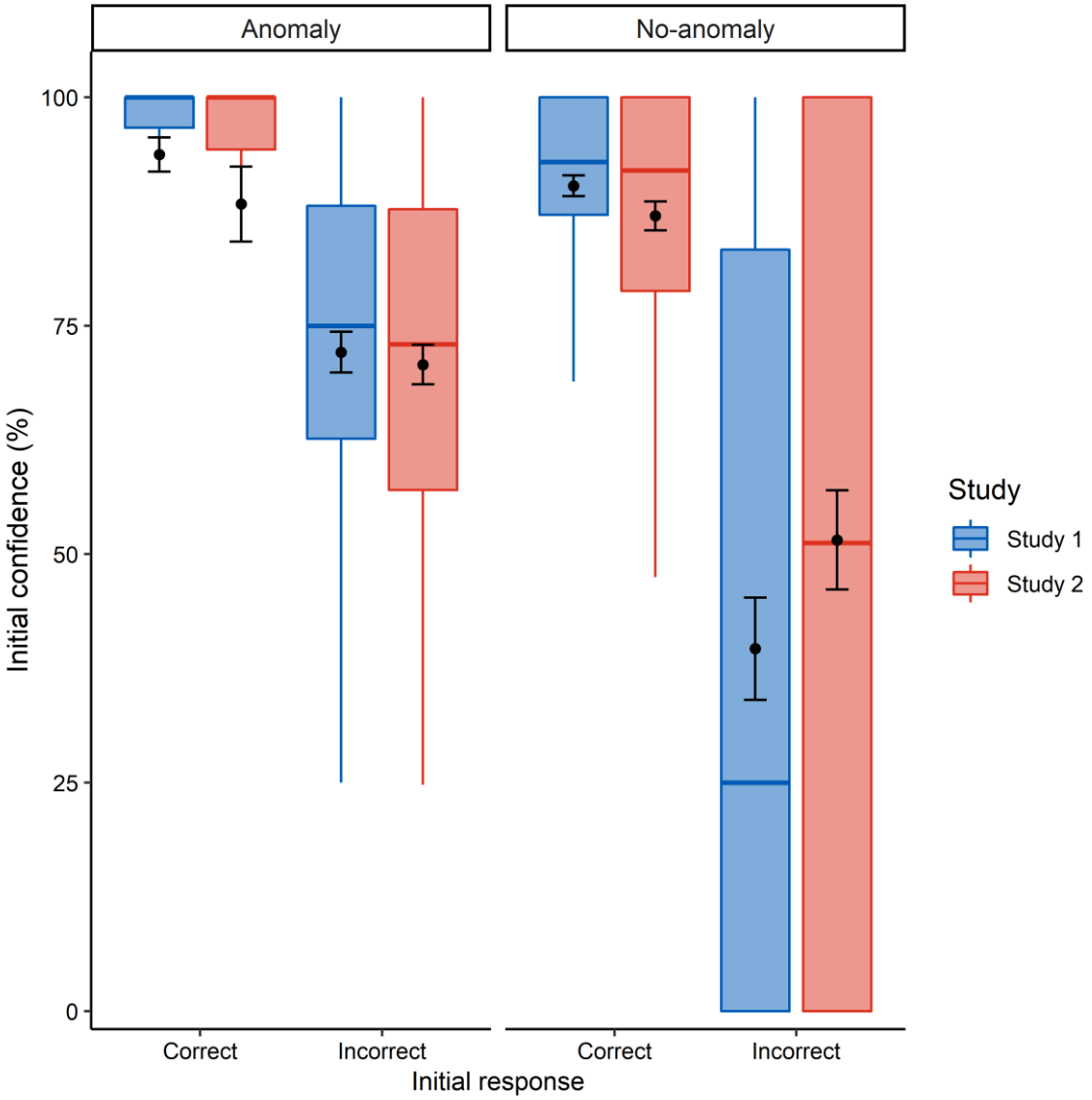

Figure 3. Initial confidence ratings at anomaly and control no-anomaly trials as a function of accuracy in Study 1 and Study 2. The lower and upper hinges of the boxplot correspond to the first and third quartiles, and the middle line shows the median. The lower (resp. upper) whiskers extend from the hinges to the smallest (resp. largest) value no further than 1.5 times the interquartile range. Overlaid black dots represent the mean and error bars are standard errors of the mean.

Illusion Strength Analysis. Overall, our anomaly items managed to trigger semantic illusions and we observed low average accuracy across our problems. However, not all individual items triggered the illusion to the same extent. For some problems the intuitive pull of the impostor word seemed stronger than others. This naturally occurring variance in "illusion strength" (i.e., how difficult it is to spot the illusion as operationalized by response accuracy) can be used to

further validate our confidence findings. Our overall evidence for intuitive error detection suggests that when responding incorrectly to the question "In the biblical story, how many animals of each kind did Moses take on the Ark?", the correct concept of "Noah" is also activated on some level. However, the strength of this correct intuition may differ across items. The stronger the illusion (i.e., the lower the correct intuition's strength / the lower the item's accuracy), the harder it will be to spot the anomaly, and the less likely that people will show error sensitivity and doubt their answer. In the reasoning field, such a link between illusion strength and error detection has already been established (e.g., Bago & De Neys, 2020; De Neys & Pennycook, 2019; Pennycook et al., 2015).

Interestingly, for correctly solved anomaly problems one can make the exact opposite prediction with respect to response confidence. That is, our overall analysis indicated that on average correct responders did not doubt their answer. However, if an illusion is particularly strong, even correct responders may feel more conflicted and less confident about their answer. This pattern has also been observed in the reasoning field (e.g., Bago & De Neys, 2020; De Neys & Pennycook, 2019; Pennycook et al., 2015).

Here, we ran a post-hoc analysis to explore these hypotheses. As an exploratory first test of these two predictions, we computed an "illusion strength" measure for each of our 20 items. For each question, we calculated the difference between the mean accuracy of the control no-anomaly version and the mean accuracy of the anomaly version in the initial, intuitive response stage. The mean accuracy of the control version thus served as a baseline: The lower the average anomaly version accuracy in comparison with the control version, the stronger the illusion. An items' illusion strength thus reflects its capacity to elicit the illusion in the initial stage at the group level. To recap, we expect that as illusion strength increases (i.e., the correct intuition decreases), participants will show more confidence in their errors (i.e., less error detection) and less confidence in correct responses.

Averaging the data over the 20 items would have raised statistical power issues, so we used linear mixed models to analyze the data on a trial-by-trial basis, while taking the by-participant variations into account (Baayen et al., 2008). We thus built a linear mixed model with random intercepts for participants and the initial confidence as the dependent variable. As fixed factors, we entered a variable which we will refer to as "response group", illusion strength, and their interaction.

The "response group" variable coded whether a given data point was a control trial on which the correct response was selected (intercept of the model), an anomaly trial on which the correct response was selected, or an anomaly trial on which the incorrect response was selected. Figure 4 plots the result of the regression (the full model is reported in Supplementary Material D). Note that a regression line parallel to the correct control baseline (dashed line) would indicate that confidence is not modulated by illusion strength.

Critically, the interaction term between illusion strength and the response groups was significant both for correct anomaly answers (b = −0.36, t(1487.61) = −2.63, p = .009, 95% CI [−0.63, −0.09]) and for incorrect anomaly answers (b = 0.19, t(1532.32) = 2.16, p = .03, 95% CI [0.02, 0.35]). As Figure 4 indicates, as illusion strength increased (i.e., the alleged correct intuition decreased), confidence decreased for correct anomaly responses and increased for incorrect anomaly responses. Hence, as could be expected, error sensitivity became less pronounced for "harder" problems whereas correct responses to these problems were doubted more.

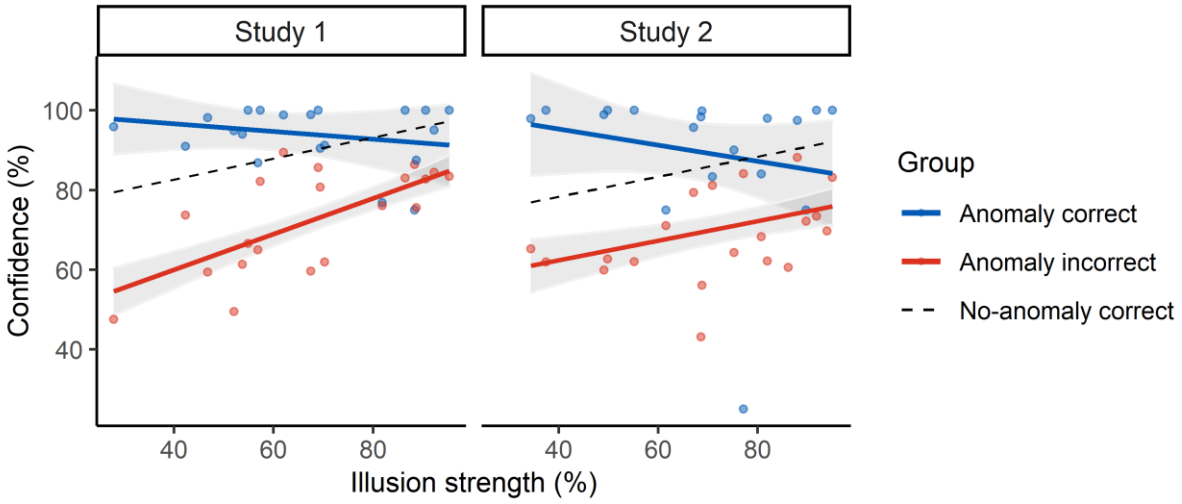

Figure 4. Regression results of initial confidence as a function of illusion strength for control no-anomaly correct (baseline), anomaly correct and anomaly incorrect responses. Illusion strength = mean initial no-anomaly accuracy − mean initial anomaly accuracy for each item. The shaded bands are 95% confidence bands. One dot represents the observed average initial confidence rating for one individual item for correct or incorrect anomaly trials.

Study 2

The results of our first study challenge the traditional dual process account of semantic illusions. When participants managed to give a correct answer to an anomaly problem at the final stage, they already gave a correct answer at the initial stage about half the time (mean non-correction rate = 46.1%). This result questions the corrective deliberation assumption in dual process accounts of semantic illusions. In addition, when participants failed to answer correctly in the initial response stage, they were still able to detect their error to some extent, as indicated by a decrease in their confidence.

In Study 2, we introduced methodological refinements to test the robustness of the findings. Although in Study 1 we combined three validated procedures (instructions, time pressure, and concurrent load) to minimize deliberation in the initial response stage, one concern is that the procedure was not stringent enough. If participants already deliberated in the initial response stage, this could explain the high non-correction rate and initial error sensitivity. To rule out this concern, we used more extreme load and time pressure manipulations (for a similar approach, see Bago & De Neys, 2019a).

Method

The research question and study design were preregistered on the OSF platform (https://osf.io/295bf). No specific analyses were preregistered.

Participants

We recruited 100 online participants (77 females, *M* age = 35.1 years, *SD* = 15.4 years) on the Prolific platform. Only native English speaking American (USA) participants were allowed to participate in the study. They were paid £5 per hour for their participation. Among them, 44 reported high school as their highest educational level, 1 less than high school, and 55 a higher education degree.

Materials and Procedure

Except for the initial response deadline and the load task, the materials and the procedure were the same as in Study 1.

Response Deadline. In Study 1, the initial response deadline was set to 5 seconds, based on our one-response pre-test. The results of Study 1 indicated

that on average participants respected the instructions and overall responded before the deadline. To further minimize the possibility that participants engage in deliberation during the initial response stage, we decided to use a more stringent time limit. The average correct, initial anomaly response latency in Study 1 was 3.8 s. On the basis of this result, we decided to round this value to the nearest integer (to give participants some minimal leeway) and decreased the deadline further to 4 s in Study 2. The screen turned yellow 1 s before the deadline to urge participants to enter their response.

To check whether the time pressure had increased between Study 1 and Study 2, we contrasted the response latencies in the initial response stage of the two studies. Participants responded significantly faster in Study 2 (M = 2.9 s) than in Study 1 (M = 3.2 s), $t(184.3) = 5.43$, $p < .001$, 95% CI [0.19, 0.42]. However, note that although we decreased the deadline by one entire second between the two studies, the mean initial latency difference was only 0.3 s. This indicates that participants were already responding near the minimal threshold in Study 1.

Load Task. In Study 2, we also increased the cognitive load during the initial response stage. In Study 1, participants had to memorize a complex four-cross pattern in a 3 x 3 grid. In Study 2, we presented a five-dot pattern in a 4 x 4 grid (Figure 1b; e.g., Bialek & De Neys, 2017; Trémolière & Bonnefon, 2014). This more extreme load has been shown to further burden participants' cognitive resources compared to the load task we used in Study 1 (Trémolière et al., 2012). Except for the more demanding five-dot patterns, the load task was the same as in Study 1.

Exclusion Criteria

Participants failed to provide an initial response within the deadline in 14.8% of the trials, and did not recall the correct dot pattern in 20.3% of the trials. A total of 31.6% of the trials were excluded, and we thus analyzed 1369 trials out of 2000. On average, each participant contributed a total of 13.7 valid trials (out of 20, SD = 3.3 trials).

Note that this higher proportion of excluded trials in Study 2 (i.e., 31.6% vs. 18.2% in Study 1) was to be expected, as we used a very stringent deadline and a more demanding load to be maximally sure that participants could not engage in slow deliberation during the initial response stage. However, since we

only discarded individual trials (rather than participants), this higher exclusion rate should not give rise to confounding selection effects (e.g., Bouwmeester et al., 2017).

Results and Discussion

Accuracy

Figure 2a summarizes the initial and final accuracies as a function of anomaly presence. These parallel the Study 1 findings. For the control no-anomaly questions, participants performed very well both at the initial (M = 85.9%, SD = 16.6%) and the final (M = 90.8%, SD = 11.2%) response stages. The accuracy was lower for the anomaly questions, again both in the initial (M = 13.3%, SD = 19%) and final response stage (M = 34.4%, SD = 29.3%). As in Study 1, the vast majority of the incorrect responses to anomaly questions were mainly due to participants giving the undistorted answer (e.g., "Moses") both at the initial (M = 82.2%) and final (M = 86.2%) response stages (rather than the "Don't know" answer).

A two-way within-subject ANOVA investigating the effect of anomaly presence (control no-anomaly; anomaly) and response stage (initial; final) on accuracy revealed a significant main effect of anomaly, $F(1, 98) = 697.6$, $p < .001$, $\eta^2g = .72$., as well as response stage, $F(1, 98) = 93.7$, $p < .001$, $\eta^2g = .09$. The difference between initial and final accuracy was higher for anomaly questions compared to control questions, as indicated by the response stage by anomaly interaction, $F(1, 98) = 39.68$, $p < .001$, $\eta^2g = .04$.

Overall, the pattern of results was similar in Study 1 and 2. Participants' performance on anomaly problems is better at the final response stage, but they still manage to give correct answers at the initial response stage. However, as Figure 2a indicates, the increased load and deadline in Study 2 did tend to have an impact on the initial accuracy. A Wilcoxon rank-sum test indicated that the initial accuracy for anomaly questions was indeed significantly higher in Study 1 (M = 20.4%) than in Study 2 (M = 13.3%), $W = 5909$, $p = .01$, $r = -.18$.

Direction of Change

To better understand how people changed (or did not change) their answers after deliberation, we once again performed a direction of change analysis by

looking into how accuracy changed from the initial to the final response stage on every trial. Figure 2b gives an overview of the results. As in Study 1, the majority of control no-anomaly questions yielded "11" responses (82.8%), followed by "01" (8%), "00" (6.1%) and "10" responses (3.1%). Similarly, there was a majority of "00" responses (64.2%) for anomaly questions. However, "01" responses were more frequent (22.5%) than "11" responses (11.9%) in Study 2. The "10" response pattern was again rare (1.4%).

In order to get a more direct measure of how the proportion of "11" responses compared to that of "01" responses, we again computed the mean "non-correction rate" across participants who had managed to give at least one correct answer to an anomaly question at the initial or final stage (n = 75). The mean non-correction rate for anomaly questions was 28.1% (*SD* = 34.1%). Put differently, when participants managed to give a correct answer to an anomaly question at the final stage, they already gave a correct answer at the initial stage 28.1% of the time. A Wilcoxon rank-sum test indicated that the non-correction rate for anomaly questions was significantly higher in Study 1 (*M* = 46.1%) than in Study 2 (*M* = 28.1%), *W* = 2109, *p* = .001, *r* = −.26. In summary, our data indicates that it was still possible to generate the correct answer intuitively in Study 2. However, the correction of an erroneous intuitive answer was more frequent than a correct intuitive answer.

Confidence Ratings

Error Sensitivity. To test whether incorrect anomaly problem responders detected their error, we again contrasted the confidence ratings of the correctly solved control no-anomaly problems and the incorrectly solved anomaly problems. As Figure 3 shows, we replicated the findings of Study 1. A Wilcoxon signed-ranked test showed that the initial confidence ratings for incorrectly solved anomaly problems (*M* = 70.7%) were significantly lower than the initial confidence ratings for correctly solved control problems (*M* = 87%), *p* < .001, *r* = −.42. Participants were thus able to detect that their answer was questionable, even in the initial response stage when deliberate reflection was minimized.

To test whether the magnitude of error detection differed between Study 1 and 2, we computed the individual error sensitivity effects by subtracting the mean initial confidence for incorrectly solved anomaly problems from the mean initial confidence for correctly solved control problems. A Wilcoxon rank-sum test

indicated that error detection did not differ significantly between Study 1 (M = 18.1%) and Study 2 (M = 15.9%), W = 4819, p = .94, r = −.01.

As in Study 1, for correct anomaly responses we did not observe a similar confidence decrease when contrasting the control no-anomaly and the anomaly problems (p = .13, r = −.16). When participants avoided the illusion and responded correctly, they were highly confident that their answer was indeed correct (M = 88.3% for correct anomaly problems vs. 89.1% for correct control problems).

Illusion Strength Analysis. As in Study 1, we also performed an illusion strength analysis. For each item, we computed the illusion strength (i.e., how difficult it is to spot the illusion) by subtracting the mean initial accuracy of the anomaly version from the mean initial accuracy of the control no-anomaly version. The higher the difference between the two, the stronger we assume the illusion to be. As in Study 1, we predict that as illusion strength increases, the experienced confidence will decrease for correct responses and increase for incorrect responses (i.e., less error detection).

To test these predictions statistically, we used a linear mixed model with random intercepts for participants and the initial confidence as the dependent variable. The fixed factors were the response group variable, the illusion strength and their interaction. The "response group" variable coded whether a given data point was a control no-anomaly trial on which the correct response was selected (intercept of the model), an anomaly trial on which the correct response was selected, or an anomaly trial on which the incorrect response was selected. Figure 4 plots the result of the regression.

Critically, the interaction term between illusion strength and the response group was significant for correct anomaly answers (b = −0.46, t(1221.44) = −2.26, p = .024, 95% CI [−0.85, −0.06]). However, the interaction was not significant for incorrect anomaly answers (b = −0.01, t(1203.1) = −0.07, p = .95, 95% CI [−0.22, 0.20]). Hence, when illusion strength increased (i.e., the alleged correct intuition decreased), confidence decreased for correct anomaly answers, as we expected. However, in this study, confidence was not significantly modulated by illusion strength for incorrect anomaly responses.

Study 3

Taken together, in Study 2, we found that correct intuitive responding was still possible for participants despite an even more stringent deadline and a higher cognitive load than in Study 1. These additional constraints decreased the initial accuracy at the anomaly problems, which led to a lower non-correction rate compared to Study 1. Nevertheless, even with extreme constraints in Study 2, correct intuiting was still present. These results suggest that when faced with semantic illusions, people may often need to engage in deliberation to respond correctly. However, there is still a substantial number of cases in which participants manage to generate correct responses to these problems intuitively. Sound intuiting is thus a (less prevalent but non-negligible) alternative route to correct responding. Concerning the error detection findings, we replicated the main results of Study 1. The harder deadline and load did not affect error sensitivity, suggesting the process mainly operates intuitively.

Study 1 and Study 2 also suggested that illusion strength modulated error detection. When illusion strength increased, response confidence tended to decrease for correct responses and to increase for incorrect anomaly responses. In other words, error detection became less likely for "harder" problems. However, these results were only correlational and did not always reach statistical significance. In Study 3, we sought to test the impact of illusion strength experimentally, by directly manipulating the semantic overlap between the impostor (e.g., "Moses") and the undistorted original term (e.g., "Noah", Hannon & Daneman, 2001; Van Oostendorp & De Mul, 1990). This allowed us to get a more controlled measure of illusion strength. We created "weak" (easy to spot) impostor questions (e.g., "How many animals of each kind did *Goliath* take on the ark?"), compared to the "strong" (hard to spot) impostor questions we used in Study 1-2 (e.g., "How many animals of each kind did *Moses* take on the ark?").

The first goal of the experiment was to test whether the weak-impostor versions of the questions would elicit more correct intuitive responses than the strong-impostor versions of the questions of Study 2. Indeed, if the correct intuition is made stronger (as we expect in the weak-impostor questions), we should observe more initial correct responses.

Second, we wanted to test whether response confidence would be modulated by the strength of the impostor in a similar fashion as in our illusion

strength analysis. Compared to the strong-impostor questions, the weak-impostor questions can be expected to increase the activation of the correct intuition (e.g., "It's definitely not Goliath"). We therefore expected that participants in the weak-impostor condition would be less confident than participants in the strong-impostor condition for incorrect initial anomaly responses (i.e., intuitive error sensitivity increases for easier problems), whereas for correct responses they would be more confident than participants in the strong-impostor conditions.

Note that we used the results of Study 2 as our strong-impostor baseline to be compared with our weak-impostor results of Study 3.

Method

The research question and study design were preregistered on the OSF platform (https://osf.io/64t9h). No specific analyses were preregistered.

Participants

We recruited 100 online participants (78 females, *M* age = 35.6 years, *SD* = 15.1 years) on the Prolific platform. Only native English speaking American (USA) participants were allowed to participate in the study. They were paid £5 per hour for their participation. Among them, 32 reported high school as their highest educational level, 1 less than high school, and 67 a higher education degree.

Materials and Procedure

Except for the semantic similarity manipulation, we used the exact same materials and procedure as in Study 2.

Semantic Similarity Manipulation. We constructed weaker versions of our anomaly problems (e.g., Hannon & Daneman, 2001). For instance, in the following anomaly question: "In the biblical story, how many animals of each kind did Moses take on the Ark?", we replaced the strong-impostor word "*Moses*" by "*Goliath*". Note that both the weak-impostor and the strong-impostor words were semantically related to the control no-anomaly target ("*Noah*"). Hence, completely unrelated words (e.g., "*Kennedy*") were avoided. However, the strong impostor was more strongly related to the control no-anomaly target than the weak impostor. For each problem, a set of possible candidate weak-impostor words were generated by the three co-authors (partly based on Hannon & Daneman, 2001).

After discussion, we decided on the best alternative for each problem. In case no agreement could be reached (5 problems) the top two alternatives were both included in the pretest rating study (see below). For these problems, we selected the alternative that received the most distinctive (i.e., lowest) similarity rating in the pretest.

Pretest. We recruited 25 native English speaking American (USA) participants (11 females, *M* age = 36.3 years, *SD* = 10.5 years) on the Prolific platform. For each question, participants had to first read the control, no-anomaly version of the question along with the correct answer. The strong- and the weak-impostor versions of the questions were displayed below in a random order, with the impostor words in upper case. To illustrate, here is an example of one complete trial for a given question:

> The undistorted question is: "In the biblical story how many animals of each kind did NOAH take on the Ark? (answer: Two)"
>
> How similar is each distorted sentence to the original undistorted question?
>
> Please type a number from 0 (Not at all similar) to 100 (Extremely similar) for each sentence.
>
> In the biblical story how many animals of each kind did GOLIATH take on the Ark?
>
> In the biblical story how many animals of each kind did MOSES take on the Ark?

On average, the weak-impostor versions received a significantly lower similarity rating (*M* = 23.2%, *SD* = 19.6%) than the strong-impostor versions, *M* = 38.7%, *SD* = 19.3%, $t(24) = 6.59$, $p < .001$. Furthermore, the mean rating of every individual item was higher for the weak-impostor version than for the strong-impostor version. See Supplementary Material A for the complete list of weak-impostor questions.

Exclusion Criteria. Participants failed to provide an initial response within the deadline in 14.8% of the trials, and did not recall the correct dot pattern in 18% of the trials. In addition, due to a coding error, the trials for one control no-anomaly version of an item ("In the biblical story, how many animals of each kind did Noah take on the Ark?") could not be analyzed. This accounted for 2.4% of the subset of the to-be-analyzed trials. A total of 30.5% of the trials were

excluded, and we thus analyzed 1390 trials out of 2000. On average, each participant contributed a total of 13.9 valid trials (out of 20, *SD* = 3.5 trials).

Results and Discussion

Accuracy

Figure 5a gives an overview of the initial and final accuracies as a function of the impostor strength condition (strong vs. weak). As expected, there seemed to be no difference in accuracy between Study 2 and Study 3 for the (identical) control no-anomaly questions. For the anomaly questions however, participants had a higher accuracy for weak-impostor questions both at the initial (Strong = 13.3%; Weak = 26.4%) and the final (Strong = 34.4%; Weak = 56.2%) response stages. Thus, our weak-impostor manipulation worked as intended in that it boosted performance. However, note that in a fair amount of cases participants still failed to solve the weak-impostor versions correctly, even in the final response stage.

To test these results statistically, we used two separate two-way mixed ANOVAs for the initial and final response stage separately, with impostor strength as our between-subject variable and anomaly presence as our within-subject variable. Results confirmed the visual inspection of Figure 5a. For the initial response stage, the main effects of anomaly version, $F(1, 197) = 1383.7$, $p < .001$, $\eta^2g = .72$, and impostor strength, $F(1, 197) = 7.6$, $p = 0.006$, $\eta^2g = .02$ were both significant, as well as their interaction, $F(1, 197) = 14.6$, $p < .001$, $\eta^2g = .03$. Post hoc t tests using the Holm correction revealed that the difference between the weak and the strong-impostor questions was significant for the anomaly versions of the questions ($p < .001$), but not for the (identical) control versions ($p = .88$).

For the final response stage, the main effect of anomaly presence, $F(1, 197) = 487$, $p < .001$, $\eta^2g = .50$, and impostor strength, $F(1, 197) = 21.4$, $p < .001$, $\eta^2g = .06$, were both significant, as well as their interaction, $F(1, 197) = 23.4$, $p < .001$, $\eta^2g = .05$. Post hoc t tests using the Holm correction revealed that the difference between the weak and the strong-impostor questions was significant for the anomaly versions of the questions ($p < .001$), but not for the (identical) control versions ($p = .63$). Participants were thus better at avoiding the illusion

when the impostor was weak compared to when it was strong, both at the initial and the final response stage.

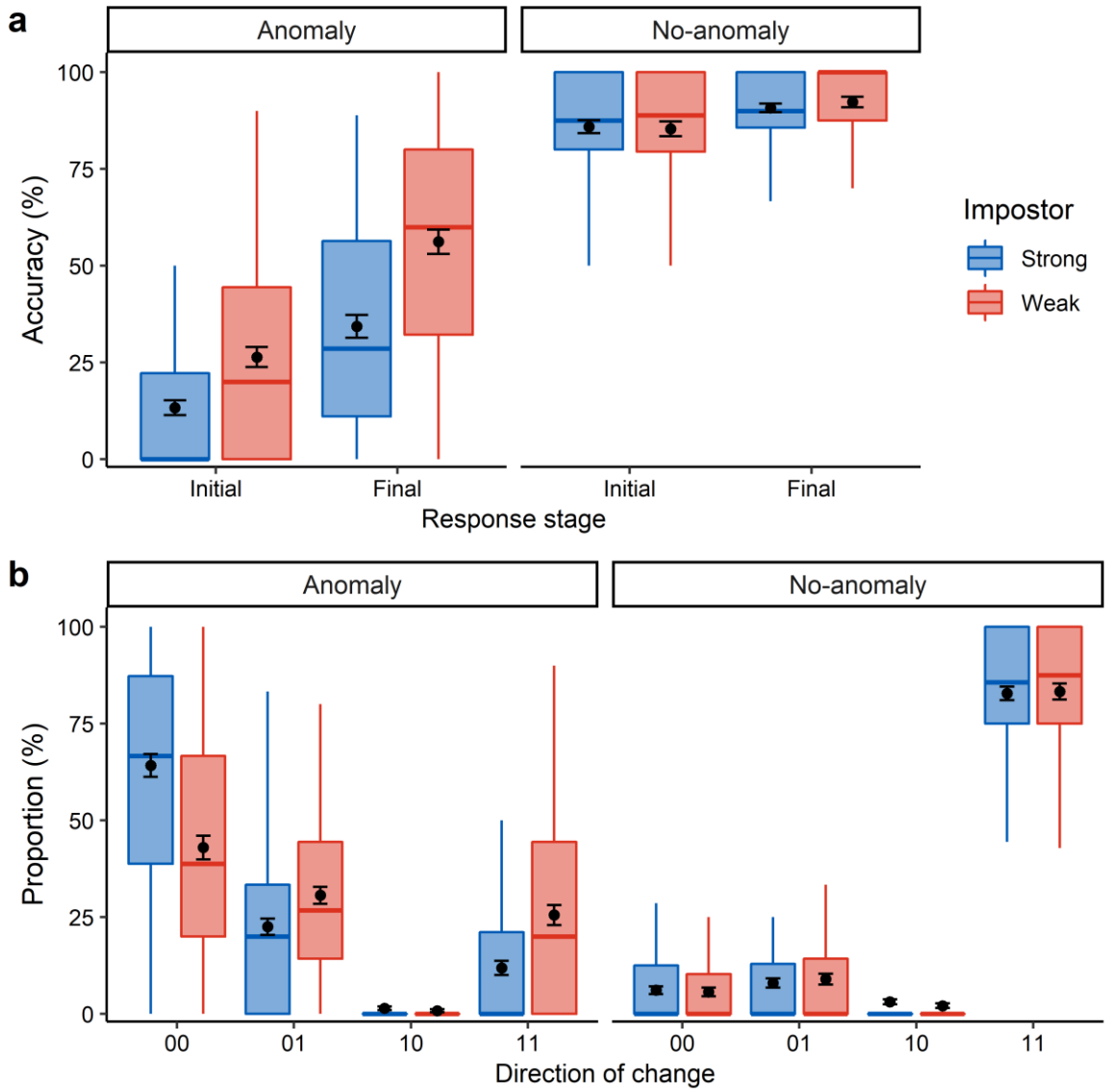

Figure 5. Accuracy and Direction of change as a function of impostor strength ("Strong" = Study 2; "Weak" = Study 3). **a**) Response accuracy at anomaly and control no-anomaly trials as a function of response stage and impostor strength. **b**) Proportion of each direction of change category at anomaly and control no-anomaly trials as a function of impostor strength; "00" = incorrect initial and incorrect final response; "01" = incorrect initial and correct final response; "10" = correct initial and incorrect final response; "11" = correct initial and correct final response. The lower and upper hinges of the boxplot correspond to the first and third quartiles, and the middle line shows the median. The lower (resp. upper) whiskers extend from the hinges to the smallest (resp. largest) value no further

than 1.5 times the interquartile range. Overlaid black dots represent the mean and error bars are standard errors of the mean.

Direction of Change

To better understand how people changed (or did not change) their answers after deliberation as a function of impostor strength, we once again performed a direction of change analysis by looking into how accuracy changed from the initial to the final response stage on every trial. Results are summarized in Figure 5b.

For the control no-anomaly problems, the results were very similar to Study 2, with a majority of "11" responses (83.3%), followed by "01" (9%), "00" (5.6%) and "10" responses (2.1%). For the anomaly problems, the proportion of "00" responses was lower for the weak-impostor questions (43%) than for the strong-impostor questions (64.2%). Crucially, the weak-impostor questions had a higher proportion of "11" (25.5%) responses compared to the strong-impostor questions (11.9%). The proportion of "01" responses was also higher for the weak-impostor questions (30.7%) than for the strong-impostor ones (22.5%).

To get a more direct measure of how the proportion of "11" responses compared to that of the "01" responses, we computed the mean non-correction rate across participants who had managed to give at least one correct answer to an anomaly question at the initial or final stage (n = 90). The mean non-correction rate for the weak impostor anomaly questions in Study 3 was 39.8% (*SD* = 32.5%). Put differently, when participants managed to give a correct answer to an anomaly problem at the final stage, they already gave a correct answer at the initial stage 39.8% of the time. To quantify the impact of our impostor strength manipulation over correct intuitive responding (vs. correction of an initial erroneous answer), we directly contrasted the non-correction rates as a function of impostor strength. A Wilcoxon rank-sum test indicated that the non-correction rate for anomaly problems was significantly higher in Study 3 (M = 39.8%) than in Study 2 (M = 28.1%), $W = 2664$, $p = .016$, $r = -.19$. The impact of the semantic similarity manipulation on accuracy was thus more linked to increased correct intuitive responding than to an increased deliberate correction of an initial erroneous answer.

Confidence Ratings

We expected that participants in the weak-impostor condition would be less confident than participants in the strong-impostor condition for incorrect initial anomaly responses (i.e., intuitive error sensitivity increases for easier problems), whereas for correct responses they would be more confident than participants in the strong-impostor conditions.

Figure 6 shows the confidence findings at the initial response stage. For correct control no-anomaly problems, the initial confidence is nearly identical in the weak- and strong-impostor conditions. For incorrect anomaly problems, a visual inspection of the figure shows a lower confidence for weak-impostor questions (i.e., a higher error sensitivity). For correct responses (where we expected higher confidence for weak impostor questions), there seems to be no clear difference between the weak- and the strong-impostor conditions.

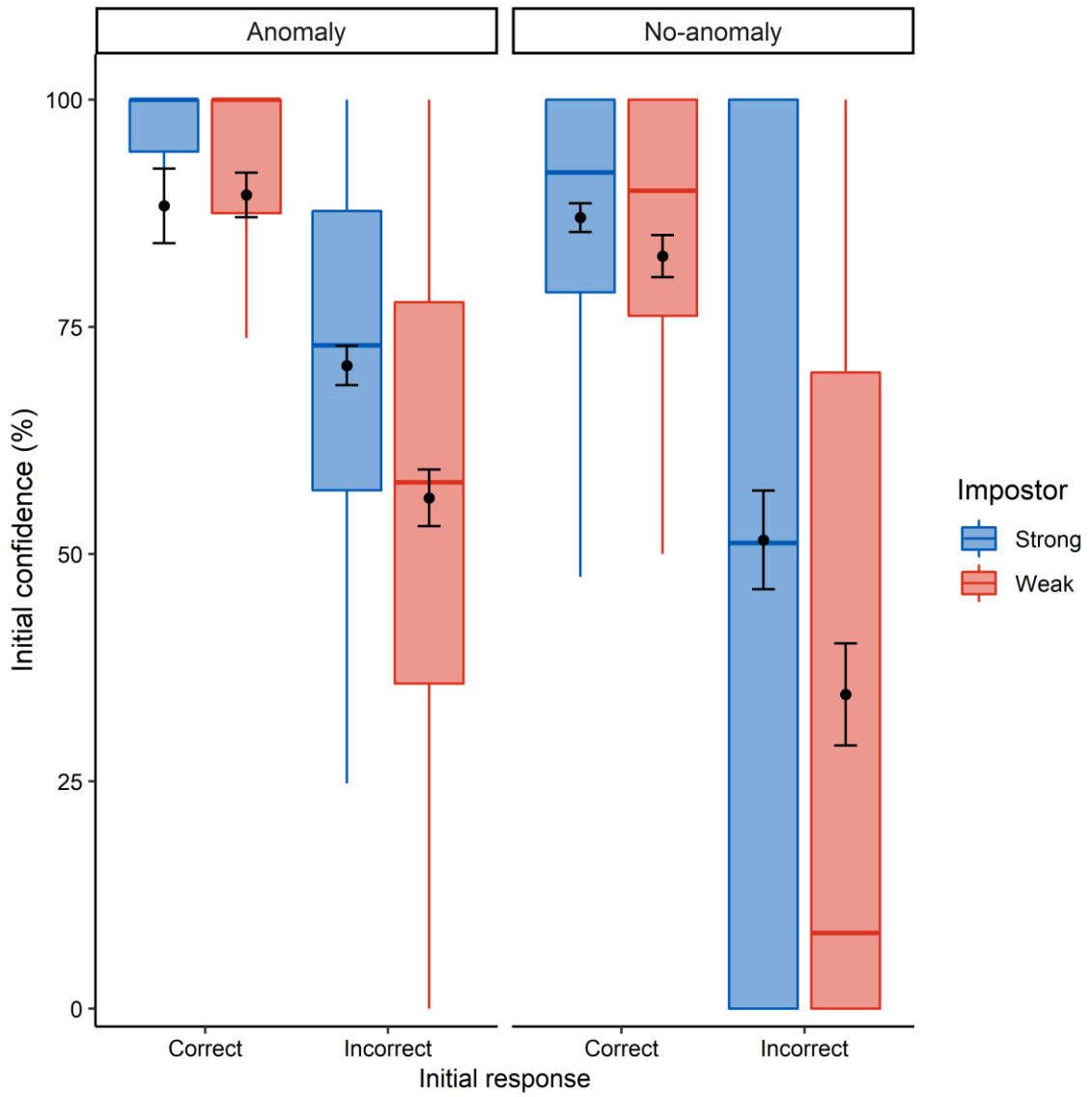

Figure 6. Initial confidence at anomaly and control no-anomaly trials as a function of impostor strength ("Strong" = Study 2; "Weak" = Study 3). The lower and upper hinges of the boxplot correspond to the first and third quartiles, and the middle line shows the median. The lower (resp. upper) whiskers extend from the hinges to the smallest (resp. largest) value no further than 1.5 times the interquartile range. Overlaid black dots represent the mean and error bars are standard errors of the mean.

To test these results statistically, we directly contrasted the initial confidence for the correct control no-anomaly trials (which served as baseline) with the confidence for correct and incorrect anomaly problems as a function of impostor strength. A Wilcoxon rank-sum test indicated that the decrease for incorrect responses was significantly higher for weak-impostor questions (*M* =

26.6%) than for strong-impostor questions (M = 15.9%), W = 3809.5, p = .007, r = −.19. Hence, as expected, when responding incorrectly, participants' error sensitivity increased for the easier weak-impostor questions. But there was no significant difference between strong (M = −1.4%) and weak-impostor questions (M = 0.8%) for correct responses, W = 1312, p = .59, r = −.05. Note, however, that correct absolute trial confidence was already approaching the maximum for strong-impostor anomaly questions. Thus, the lack of a significant further increase for correct weak anomaly problems may be due to the presence of a ceiling effect.

General Discussion

The present studies tested the popular dual process view of semantic illusions. According to this influential account, giving the correct answer to an anomaly question requires deliberate processing to overcome an intuitively cued incorrect answer. By adopting a two-response paradigm in which participants had to give an initial response under time pressure and a concurrent cognitive load, we first tested whether the correct response could also be generated intuitively. Across our three studies, we found that deliberate correction was more frequent than correct intuiting. Our results thus suggest that avoiding semantic illusions typically requires deliberation. However, participants still gave a correct intuitive answer in a non-negligible proportion of trials. Sound intuiting is thus a less prevalent but non-negligible alternative route to correct responding in the case of semantic illusions.

Second, the dual process account of semantic illusions further assumes that falling for the illusion results from a failure to engage in deliberate processing and to detect the anomaly in the sentence. In each of our three studies, we consistently observed that participants who gave incorrect responses to an anomaly question were sensitive to the erroneous nature of their response (as reflected in a decreased confidence). Critically, this error detection effect was present at the initial, intuitive stage, suggesting that this error sensitivity is an automatic and effortless process.

Third, the illusion strength analyses of Study 1-2 and the impostor strength manipulation of Study 3 suggest that both confidence and performance in the initial, intuitive stage depend on the strength of the illusion. When the illusion was

weaker, sound intuiting became more frequent and confidence tended to decrease in the case of an incorrect response (i.e., more error sensitivity).

These results have both theoretical and practical implications. At the theoretical level, our results allow us to better understand the nature of erroneous and correct responding in semantic illusions. Our error sensitivity findings indicate that errors do not result from a failure to detect the anomaly in the distorted sentence. Hence, when responding incorrectly to the question "In the biblical story, how many animals of each kind did Moses take on the Ark?", the correct concept of "Noah" must also be activated on some level. In addition, our results show that correct responding does not necessarily require deliberation but that it can frequently be achieved intuitively. Overall, these findings are consistent with recent advances in dual process theorizing in which the intuitive performance to classic "bias" tasks is determined by the interplay of both incorrect and correct intuitions (De Neys, 2022). According to this dual process model 2.0. (De Neys, 2017), the logical, correct response that has traditionally been considered to be cued by deliberation, can also be cued intuitively through System 1. Hence, the model assumes that System 1 will cue both an incorrect "heuristic" and correct "logical" intuition in a typical heuristics-and-biases task. Whichever intuition is strongest will be selected as initial response. In addition, the more similar the strength of the competing intuitions, the more conflict will be experienced, and the more one will doubt their decision. The current findings are in line with this view, as they also suggest that reasoners faced with semantic illusions automatically generate both a correct ("Noah") and an incorrect ("Two") intuition. The illusion strength analyses and the impostor strength manipulation further support the idea that the strength interplay of these intuitions determines the intuitive performance in semantic illusions (i.e., how likely it is one responds correctly and how likely it is one shows error detection).

These results also bear practical implications, as semantic illusions have been used as a measure of people's capacity and disposition to engage in effortful deliberation (e.g., Sirota et al., 2021). However, our results indicate that correct answers in the case of semantic illusions are often generated intuitively. Therefore, the mere use of correct answers on semantic illusion problems as an index of deliberate processing abilities can be problematic and distort conclusions. To clearly measure one process or the other (i.e., intuition or deliberation capacities),

we recommend to test how the answer has been generated (e.g., by using a two-response paradigm).

Critics of our work may argue that we observed correct intuiting because the items we selected might have been relatively easy compared to those typically used in the literature. To test this hypothesis, we can directly compare our illusion rate to the results in the literature. Following Speckmann and Unkelbach (2021), we computed the rate of incorrect undistorted responses (e.g., "Two") at anomaly questions in the final response stage across Study 1 and Study 2 (i.e., the % of trials in which participants fell prey to the illusion). The result (56%) was higher than what Reder and Kusbit (1991) reported (33% in Experiment 1, 35% in Experiments 2 & 3, 32% in Experiment 4; we only used the results from the comparable literal task condition). Similarly, Speckmann and Unkelbach (2021) also found lower rates than the ones reported here (49% in Experiment 1, 52.6% in Experiment 2). This slightly higher illusion rate in our study may be explained by the fact that we only selected items from Speckmann and Unkelbach (2021) which had a high knowledge check as well as a high control no-anomaly accuracy. Hence, if anything, our items were overall harder than those adopted in the literature which implies that sound intuiting will be even more prevalent in other studies.

Another critique might be that correct intuiting was possible in our studies because our design was not challenging enough and still allowed deliberation. To minimize the possibility that reasoners engage in deliberate processing in the initial stage, we combined 3 validated procedures (instructions, time pressure, and concurrent load). All these manipulations have been previously shown to minimize deliberation. In Studies 2-3, we used an even more challenging load task and deadline to further minimize the possibility that reasoners would deliberate in the initial response stage. Nevertheless, one may still argue that we could have used an even more demanding deadline and load task. However, the high number of missed trials in Study 2 (31.6%) and Study 3 (30.5%) shows that adding load or time pressure would have raised practical and statistical issues (i.e., selection effects due to a large portion of discarded trials, e.g., Bouwmeester et al., 2017). From a more theoretical standpoint, the problem is that dual process theories are underspecified (Kruglanski, 2013). The framework often entails that System 2 is slower and more demanding than System 1 but gives us no unequivocal a priori criterion that allows us to classify a process as intuitive or deliberate (e.g., takes

at least x time, or x amount of load). Consequently, as long as we keep on observing correct initial responses, one can always argue that these will disappear "with just a little bit more load/time pressure". However, note that the corrective assumption becomes unfalsifiable at this point. Any evidence for correct intuiting can always be explained away by arguing that the methodological design let room for deliberation.

Although the conflict detection effect we observed is in line with findings in the logical reasoning field (e.g., Bago & De Neys, 2017, 2020), sound intuiting was less prevalent in our semantic illusion task than what is typically found in the reasoning field—where correct intuitive responding tends to be the modal pattern over deliberative correction. For instance, across four experiments, Bago and De Neys (2017) reported higher "non-correction rates" for syllogistic reasoning (*M* = 87.6%) and base-rates (*M* = 74.8%) tasks than what we found (46.1% in Study 1; 28.1% in Study 2)[2]. We speculate that this may be explained by the different nature of our task. Classic reasoning tasks can be solved using a universal algorithm. Once you know the correct rule, you can apply it whatever the specific values in the problem are. For instance, in a base-rate task, you simply have to give weight to the priors/base-rates that are given in the problem. Likewise, in the bat-and-ball problem one might use the equation "x + y = $a. x = y + b. Solve for x", for example. The solution strategy can thus be automatized, which is assumed to be the nature of correct intuitions in these tasks (De Neys, 2012; Raoelison et al., 2020). However, in the case of semantic illusions, there is no general algorithm one could apply. Instead, one can only carefully search their semantic memory, but this search will be "unique" for each problem. Hence, this semantic search strategy might be less automatized than applying the correct rule in a reasoning problem. Therefore, "correct" responses might be less instantiated than in classic reasoning tasks which would explain the lower prevalence of correct intuiting in the case of semantic illusions.

To conclude, we believe that it is hard for the popular traditional dual process account of semantic illusions to account for our findings, and that they rather support recent models in which the absolute and relative strength of competing intuitions determines performance.

[2] Experiments using moral (Bago & De Neys, 2019a) and prosocial (Bago et al., 2021) reasoning tasks also yielded very high overall "non-correction rates" (83.8% and 83.1% respectively).

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Bago, B., Bonnefon, J.-F., & De Neys, W. (2021). Intuition rather than deliberation determines selfish and prosocial choices. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0000968

Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition*, *158*, 90–109. https://doi.org/10.1016/j.cognition.2016.10.014

Bago, B., & De Neys, W. (2019a). The intuitive greater good: Testing the corrective dual process model of moral cognition. *Journal of Experimental Psychology: General*, *148*(10), 1782–1801. https://doi.org/10.1037/xge0000533

Bago, B., & De Neys, W. (2019b). The smart System 1: Evidence for the intuitive nature of correct responding on the bat-and-ball problem. *Thinking & Reasoning*, *25*(3), 257–299. https://doi.org/10.1080/13546783.2018.1507949

Bago, B., & De Neys, W. (2020). Advancing the specification of dual process models of higher cognition: A critical test of the hybrid model view. *Thinking & Reasoning*, *26*(1), 1–30. https://doi.org/10.1080/13546783.2018.1552194

Bialek, M., & De Neys, W. (2017). Dual processes and moral conflict: Evidence for deontological reasoners' intuitive utilitarian sensitivity. *Judgment and Decision Making*, *12*(2), 148.

Bouwmeester, S., Verkoeijen, P. P., Aczel, B., Barbosa, F., Bègue, L., Brañas-Garza, P., Chmura, T. G., Cornelissen, G., Døssing, F. S., & Espín, A. M. (2017). Registered replication report: Rand, greene, and nowak (2012). *Perspectives on Psychological Science*, *12*(3), 527–542. https://doi.org/10.1177/1745691617693624

Burič, R., & Konrádová, L. (2021). Mindware instantiation as a predictor of logical intuitions in the Cognitive Reflection Test. *Studia Psychologica*, *63*(2), 114–128. https://doi.org/10.31577/sp.2021.02.822

Burič, R., & Šrol, J. (2020). Individual differences in logical intuitions on reasoning problems presented under two-response paradigm. *Journal of Cognitive Psychology*, *32*(4), 460–477. https://doi.org/10.1080/20445911.2020.1766472

Büttner, A. C. (2012). The effect of working memory load on semantic illusions: What the phonological loop and central executive have to contribute. *Memory*, *20*(8), 882–890. https://doi.org/10.1080/09658211.2012.706308

Cantor, A. D., & Marsh, E. J. (2017). Expertise effects in the Moses illusion: Detecting contradictions with stored knowledge. *Memory*, *25*(2), 220–230. ☐ https://doi.org/10.1080/09658211.2016.1152377

De Neys, W. (2012). Bias and Conflict: A Case for Logical Intuitions. *Perspectives on Psychological Science*, *7*(1), 28–38. https://doi.org/10.1177/1745691611429354

De Neys, W. (2017). Bias, conflict, and fast logic: Towards a hybrid dual process future? In *Dual process theory 2.0* (pp. 47–65). Routledge. https://doi.org/10.4324/9781315204550

De Neys, W. (2022). Advancing theorizing about fast-and-slow thinking. *Behavioral and Brain Sciences*, 1–68. https://doi.org/10.1017/S0140525X2200142X

De Neys, W., Cromheeke, S., & Osman, M. (2011). Biased but in doubt: Conflict and decision confidence. *PloS one, 6*(1), e15954. https://doi.org/10.1371/journal.pone.0015954

De Neys, W., & Pennycook, G. (2019). Logic, fast and slow: Advances in dual-process theorizing. *Current Directions in Psychological Science*, *28*(5), 503–509. https://doi.org/10.1177/0963721419855658

De Neys, W., & Schaeken, W. (2007). When people are more logical under cognitive load: Dual task impact on scalar implicature. *Experimental Psychology*, *54*(2), 128–133. https://doi.org/10.1027/1618-3169.54.2.128

De Neys, W., & Verschueren, N. (2006). Working memory capacity and a notorious brain teaser: The case of the Monty Hall Dilemma. *Experimental Psychology*, *53*(2), 123–131. https://doi.org/10.1027/1618-3169.53.1.123

Erickson, T. D., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior*, *20*(5), 540–551. https://doi.org/10.1016/S0022-5371(81)90165-1

Evans, J. S. B. (2019). Reflections on reflection: The nature and function of type 2 processes in dual-process theories of reasoning. *Thinking & Reasoning*, *25*(4), 383–415. https://doi.org/10.1080/13546783.2019.1623071

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Hannon, B., & Daneman, M. (2001). Susceptibility to semantic illusions: An individual-differences perspective. *Memory & Cognition*, *29*(3), 449–461. https://doi.org/10.3758/BF03196396

Johnson, E. D., Tubau, E., & De Neys, W. (2016). The Doubting System 1: Evidence for automatic substitution sensitivity. *Acta Psychologica*, *164*, 56–64. https://doi.org/10.1016/j.actpsy.2015.12.008

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.

Kamas, E. N., Reder, I. M., & Ayers, M. S. (1996). Partial matching in the Moses illusion: Response bias not sensitivity. *Memory & Cognition*, *24*(6), 687–699. https://doi.org/10.3758/BF03201094

Koriat, A. (2017). Can people identify "deceptive" or "misleading" items that tend to produce mostly wrong answers? *Journal of Behavioral Decision Making*, *30*(5), 1066–1077. https://doi.org/10.1002/bdm.2024

Kruglanski, A. W. (2013). Only one? The default interventionist perspective as a unimodel—Commentary on Evans & Stanovich (2013). *Perspectives on Psychological Science*, *8*(3), 242–247. https://doi.org/10.1177/1745691613483477

Lawrence, M. A., & Lawrence, M. M. A. (2016). Package "ez." *R Package Version*, *4*(0).

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2019). *Package "emmeans"*.

Mata, A., Ferreira, M. B., & Reis, J. (2013). A process-dissociation analysis of semantic illusions. *Acta Psychologica*, *144*(2), 433–443. https://doi.org/10.1016/j.actpsy.2013.08.001

Mata, A., Schubert, A. L., & Ferreira, M. B. (2014). The role of language comprehension in reasoning: how "good-enough" representations induce biases. *Cognition, 133*(2), 457–463. https://doi.org/10.1016/j.cognition.2014.07.011

Mata, A., Ferreira, M. B., Voss, A., & Kollei, T. (2017). Seeing the conflict: an attentional account of reasoning errors. *Psychonomic bulletin & review, 24*(6), 1980–1986. https://doi.org/10.3758/s13423-017-1234-7

Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. *Journal of Experimental Psychology: General*, *130*(4), 621. https://doi.org/10.1037/0096-3445.130.4.621

Park, H., & Reder, L. (2004). *Moses Illusion: Implication for Human Cognition*. https://doi.org/10.1184/R1/6617207.V1

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. https://doi.org/10.1016/j.cogpsych.2015.05.001

Pennycook, G., Trippas, D., Handley, S. J., & Thompson, V. A. (2014). Base rates: Both neglected and intuitive. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(2), 544. https://doi.org/10.1037/a0034887

Raoelison, M., Thompson, V. A., & De Neys, W. (2020). The smart intuitor: Cognitive capacity predicts intuitive rather than deliberate thinking. *Cognition*, *204*, 104381. https://doi.org/10.1016/j.cognition.2020.104381

Reder, L. M., & Kusbit, G. W. (1991). Locus of the Moses illusion: Imperfect encoding, retrieval, or match? *Journal of Memory and Language*, *30*(4), 385–406. https://doi.org/10.1016/0749-596X(91)90013-A

Shafto, M., & MacKay, D. G. (2000). The Moses, mega-Moses, and Armstrong illusions: Integrating language comprehension and semantic memory. *Psychological Science*, *11*(5), 372–378. https://doi.org/10.1111/1467-9280.00273

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2015). Afex: Analysis of factorial experiments. *R Package Version 0.13–145*.

Sirota, M., Dewberry, C., Juanchich, M., Valuš, L., & Marshall, A. C. (2021). Measuring cognitive reflection without maths: Development and validation of the verbal cognitive reflection test. *Journal of Behavioral Decision Making*, *34*(3), 322–343. https://doi.org/10.1002/bdm.2213

Speckmann, F., & Unkelbach, C. (2021). Moses, money, and multiple-choice: The Moses illusion in a multiple-choice format with high incentives. *Memory & Cognition*, *49*(4), 843–862. https://doi.org/10.3758/s13421-020-01128-z

Stanovich, K. E. (2018). Miserliness in human cognition: The interaction of detection, override and mindware. *Thinking & Reasoning*, *24*(4), 423–444. https://doi.org/10.1080/13546783.2018.1459314

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*(5), 645–665. https://doi.org/10.1017/S0140525X00003435

Stupple, E. J., & Ball, L. J. (2008). Belief–logic conflict resolution in syllogistic reasoning: Inspection-time evidence for a parallel-process model. *Thinking & Reasoning, 14*(2), 168-181. https://doi.org/10.1080/13546780701739782

Stupple, E. J., Ball, L. J., Evans, J. S. B., & 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

Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning*, *20*(2), 215–244. https://doi.org/10.1080/13546783.2013.869763

Thompson, V. A., Pennycook, G., Trippas, D., & Evans, J. S. B. (2018). Do smart people have better intuitions? *Journal of Experimental Psychology: General*, *147*(7), 945. https://doi.org/10.1037/xge0000457

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*(3), 107–140. https://doi.org/10.1016/j.cogpsych.2011.06.001

Trémolière, B., & Bonnefon, J.-F. (2014). Efficient kill–save ratios ease up the cognitive demands on counterintuitive moral utilitarianism. *Personality and Social Psychology Bulletin*, *40*(7), 923–930. https://doi.org/10.1177/0146167214530436

Trémolière, B., De Neys, W., & Bonnefon, J.-F. (2012). Mortality salience and morality: Thinking about death makes people less utilitarian. *Cognition*, *124*(3), 379–384. https://doi.org/10.1016/j.cognition.2012.05.011

Van Oostendorp, H., & De Mul, S. (1990). Moses beats Adam: A semantic relatedness effect on a semantic illusion. *Acta Psychologica*, *74*(1), 35–46. https://doi.org/10.1016/0001-6918(90)90033-C

Verschueren, N., Schaeken, W., & d'Ydewall, G. (2004). Everyday conditional reasoning with working memory preload. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *26*.

Voudouri, A., Białek, M., Domurat, A., Kowal, M., & De Neys, W. (2022). Conflict detection predicts the temporal stability of intuitive and deliberate reasoning. *Thinking & Reasoning*, 1-29. https://doi.org/10.1080/13546783.2022.2077439

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., & Hester, J. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Chapter 3

Reasoning and cognitive control, fast and slow

Voudouri, A., Bago, B., Borst, G., & De Neys, W. (2023). Reasoning and cognitive control, fast and slow. *Judgment and Decision Making, 18*, E33. https://doi.org/10.1017/jdm.2023.32.

Supplementary material for this chapter can be found in Supplementary material for Chapter 3.

Abstract

Influential "fast-and-slow" dual process models suggest that sound reasoning requires correction of fast, intuitive thought processes by slower, controlled deliberation. Recent findings with high-level reasoning tasks started to question this characterization. Here we tested the generalizability of these findings to low-level cognitive control tasks. More specifically, we examined whether people who responded accurately to the classic Stroop and Flanker tasks, could also do so when their deliberate control was minimized. A two-response paradigm, in which people were required to give an initial "fast" response under time-pressure and cognitive load, allowed us to identify the presumed intuitive answer that preceded the final "slow" response given after deliberation. Across our studies we consistently find that correct final Stroop and Flanker responses are often non-corrective in nature. Good performance in cognitive control tasks seems to be driven by accurate "fast" intuitive processing, rather than by "slow" controlled correction of these intuitions. We also explore the association between Stroop and reasoning performance and discuss implications for the dual process view of human cognition.

Introduction

Sometimes a solution to a problem pops in to mind instantly and effortlessly whereas at other times arriving at a decision can take time and effort. This distinction between what is often referred to as a more intuitive and deliberate mode of cognitive processing—or the nowadays more popular "System 1" and "System 2" labels—lies at the heart of the influential "fast-and-slow" dual process view that has been prominent in research on human reasoning in the last decades (Evans, 2008; Kahneman, 2011).

Although intuitive thinking is useful when it comes to fast decision-making, it often also relies on mental shortcuts, or heuristics, which can lead to cognitive biases (Kahneman, 2011). This bias susceptibility of System 1 is often demonstrated in the literature with the use of heuristics-and-biases tasks, like the following example:

> A psychologist wrote thumbnail descriptions of a sample of 1000 participants consisting of 5 women and 995 men. The description below was drawn randomly from the 1000 available descriptions.
> Sam is a 25 years old writer who lives in Toronto. Sam likes to shop and spends a lot of money on clothes.
>
> What is most likely?
> a. Sam is a woman.
> b. Sam is a man.

Intuitively, many people will be tempted to conclude that Sam is a woman based on stereotypical beliefs cued by the description. However, given that there are far more males than females in the sample (i.e., 995 out of 1000), the statistical base-rates favor the conclusion that a randomly drawn individual will most likely be a man. Hence, logically speaking, taking the base-rates into account should push the scale to the "man" side. Unfortunately, educated reasoners are typically tricked by their intuition and often fail to solve the problem correctly (e.g., De Neys & Glumicic, 2008).

The dual process framework presents a simple and elegant explanation for this bias phenomenon (Evans, 2008; Kahneman, 2011). Dual process theorists have traditionally highlighted that taking logical principles into account typically requires demanding System 2 deliberation (e.g., Evans, 2002, 2008; Evans &

Over, 1996; Kahneman, 2011; Stanovich & West, 2000). Because human reasoners have a strong tendency to minimize difficult computations, they will often refrain from engaging or completing the slow deliberate processing when mere intuitive processing has already cued a response (Evans & Stanovich, 2013; Kahneman, 2011). Consequently, most reasoners will simply stick to the intuitive response that quickly came to mind and fail to consider the logical implications. It will only be the few reasoners who have sufficient resources and motivation to complete the deliberate computations and override the initially generated intuitive response, who will manage to reason correctly and give the logical answer (Stanovich & West, 2000). Hence, sound reasoning is, in essence, believed to be corrective in nature.

However, studies in the last decade suggest we may need to reconsider this traditional view of the two systems (De Neys & Pennycook, 2019). These studies typically present heuristics-and-biases tasks using a two-response paradigm (Thompson et al., 2011). More specifically, participants are asked to provide two consecutive responses on each task trial. The first response is given under time-pressure and a cognitive load (e.g., a parallel task taxing cognitive resources), while in the final response stage participants have no restrictions and are allowed to deliberate (Bago & De Neys, 2017). Since System 2 is believed to be slow and burden our cognitive resources, the constraints that are imposed during the initial response minimize its involvement. This way, the paradigm allows for a direct comparison of more intuitive and deliberate responses. The key finding of these studies is that in many of the (infrequent) trials where participants provide a correct, final response, they had already provided a correct response during the initial stage (e.g., Bago & De Neys, 2017, 2019a; Newman et al., 2017; Raoelison & De Neys, 2019). Hence, System 2 does not always need to revise the intuitively generated responses, as the latter might already be correct.

Relatedly, a similar line of research using the two-response paradigm has shown that when people provide biased intuitive responses, they are often sensitive to the fact that they are erring (De Neys, 2017; Pennycook et al., 2015). In other words, participants seem not completely oblivious to the fact that their answers conflict with some (logical) elements of the problem. This has been found by comparing conflict/incongruent and no-conflict/congruent versions of the same heuristics-and-biases tasks. In congruent versions, both the heuristic and logical information in the problem cue the same answer. For instance, the congruent

version of the example given above would simply switch the base-rates around (e.g., "A psychologist wrote thumbnail descriptions of a sample of 1000 participants consisting of 995 women and 5 men"). Everything else stays the same. Hence, in the congruent case both the description and the base-rates cue the same response (i.e., "Sam is a woman"). If processing logical principles such as base-rate information requires deliberation, then reasoners' initial, intuitive responses to the incongruent and the congruent versions should not differ. However, when solving incongruent trials, participants typically report lower confidence in their initial responses in comparison to congruent trials. This response doubt has been referred to as conflict detection in the reasoning field and suggests that participants are intuitively processing the conflicting information in the incongruent problem (e.g., Bago & De Neys, 2017; Burič & Srol, 2020; Mata, 2020; Pennycook et al., 2014; Thompson & Johnson, 2014; but see also Mata et al., 2014, and Mata & Ferreira, 2018).

The above findings have led researchers to propose a revised dual process model–sometimes referred to as a "Dual Process model 2.0"–which posits that System 1 can generate two types of intuitions, a classic "heuristic" intuition, and an alleged "logical" intuition (e.g., Bago & De Neys, 2017, 2019a; De Neys & Pennycook, 2019; Handley et al., 2011; Newman et al., 2017; Pennycook et al., 2015; see De Neys, 2017, for review). The latter is believed to be based on an automated knowledge of mathematical and probabilistic rules (De Neys, 2012; Evans, 2019; Stanovich, 2018).

Interestingly, similar patterns have also been observed in other higher-order reasoning tasks on moral (Bago & De Neys, 2019b; Vega et al., 2021) and prosocial (Bago et al., 2021; Kessler et al., 2017) reasoning. The main result across these studies is that responses that are assumed to require deliberation by the traditional dual-process model (e.g., taking the consequences of a moral action into account or maximizing pay-offs for oneself or others), are often generated intuitively. It then seems that there is a need to upgrade our view of the fast and intuitive System 1. Responses that are traditionally believed to necessitate controlled deliberation, often seem to fall within the realm of more intuitive processing (De Neys, 2022; De Neys & Pennycook, 2019).

The key aim of the present paper is to explore the generalizability of these findings to classic cognitive control tasks, like the Stroop task (Stroop, 1935) and the Flanker task (Eriksen & Eriksen, 1974). These are tasks that have been used

to directly tap into lower-level cognitive control processes, rather than higher order functioning, such as reasoning. Cognitive control, according to a common definition, is a group of top-down processes that help us carry out cognitive tasks when automatic responding is not sufficient (Botvinick et al., 2001; Diamond, 2013). Similar to heuristics-and-biases tasks, classic cognitive control tasks usually contain two competing pieces of information: task-relevant and task-irrelevant information. In the incongruent versions, the task-irrelevant information cues an automatic, incorrect response, which conflicts with the response cued by the task-relevant information. Conversely, in the congruent version, both the task-relevant and task irrelevant information cue the same response.

For example, one of the most popular and frequently used tasks is the Stroop (Stroop, 1935). In the Stroop task participants are presented with words that denote a colour and are written in a coloured ink (e.g., the word "red" written in blue ink). Sometimes the ink colour and the word are congruent (e.g., the word "red" written in red ink), but other times, as in the first example, they are incongruent. Participants are asked to respond to the ink colour of each word. On average, participants have longer reaction times and higher error rates when solving the incongruent compared to the congruent stimuli. This is also known as the Stroop interference effect. The most common explanation for this effect is that, since reading is an automatic process for educated adults, reading the word will always come before identifying its ink colour (Stirling, 1979; Keele, 1972; but also see Kahneman & Chajczyk, 1983). Therefore, in the incongruent trials, participants need to take the time to inhibit their automatically generated (incorrect) answer (i.e., the read word), in order to arrive at the correct answer (i.e., the ink colour in which the word is written). In other words, not giving in to the luring, automatic response is thought to require controlled, effortful processing (e.g., Botvinick et al., 2001). Put differently, cognitive control is assumed to have a corrective role: fast (incorrect) responses are generated automatically, and are then corrected by slower controlled processes. This pattern is similar to the one that has been put forward by traditional dual process theories in the reasoning field: heuristic responses are generated automatically, and are later corrected by slow, deliberate processes (e.g., Evans & Stanovich, 2013). However, as it was mentioned before, the corrective role of deliberation in the reasoning field has been questioned, and evidence shows that correct responses are often generated automatically. Given the reasoning findings, our goal in the present paper is to

examine whether correct responding to cognitive control tasks is also possible when control is minimized.

It is worth mentioning that, in line with this research question, recent cognitive control findings have shown evidence for an automatically operating (cognitive) control (Desender et al., 2013; Jiang et al., 2015, 2018; Linzarini et al., 2017). These studies focus on a phenomenon observed in cognitive control tasks, where participants tend to more often respond correctly to an incongruent trial if it is preceded by an incongruent trial (instead of a congruent one, e.g., Braem & Egner, 2018). The explanation for this phenomenon is that the cognitive control that is recruited during the first trial facilitates correct responding in the upcoming trial. Critically, studies have found that this effect persists even when the first trial is presented unconsciously (e.g., Desender et al., 2013; Jiang et al., 2015, 2018; Linzarini et al., 2017). This suggests that cognitive control on the subliminal trial can, in theory, be exerted automatically (without the participants' intention, e.g., Abrahamse et al., 2016; Algom & Chajut, 2019). These findings lend some credence to the idea that correct responding in cognitive control tasks might be observed in the absence of deliberate correction.

In Studies 1, 2 and 3 of the present paper, we directly tested this hypothesis and examined whether correct responding in cognitive control tasks is also possible when participants' deliberate control is constrained. For this purpose, we focused on the Stroop task (Studies 1 & 3) and the Flanker task (Study 2). In the Flanker task participants were presented with a central arrow surrounded by two arrows on each side. The surrounding arrows either pointed in the opposite direction (incongruent trials) or in the same direction (congruent trials) as the central arrow (Stoffels & van der Molen, 1988) and participants' task was to indicate the direction of the central arrow. We designed a two-response version of both the Stroop task and the Flanker task. We were specifically interested in testing whether, in the incongruent trials where participants managed to provide a correct final response, they had already arrived at a correct response in the initial stage or not.

A second objective of the present paper (Study 3), was to explore in what way cognitive control and reasoning performance are related. There is existing evidence in the literature showing that classic cognitive control tasks can predict reasoning accuracy (Abreu-Mendoza at al., 2020; De Neys et al., 2011; Handley

et al., 2004). Participants who score better on cognitive control tasks, such as the Stroop, tend to show less biased responding on reasoning tasks.

Despite the links between these two measures, the way in which they are related is unclear. If we assume that correct responding in both cognitive control and reasoning tasks results from the same generic mechanism, we can imagine (at least) two possible alternative routes. On the one hand, it could be that both reasoning and cognitive control tasks tap onto the same deliberate control processes. In other words, people who successfully control (and later correct) their automatically generated Stroop responses, would also be good at controlling (and correcting) their intuitive responses in reasoning tasks. Under this "smart deliberator" view (see Raoelison et al., 2020), people's performance in the Stroop task would predict their ability to deliberately correct responses in reasoning tasks. On the other hand, it might also be the case that both reasoning and cognitive control tasks tap into intuitive or automatic control processes. In other words, people who provide correct Stroop responses when their cognitive resources are restricted, would also be able to intuitively provide correct responses to reasoning problems. When these people are allowed to deliberate, they will not need to correct their intuitive answers, as these will be already correct. Under this "smart intuitor" view, people's "intuitive" performance at the Stroop task would predict their ability to generate correct intuitive responses (rather than to deliberately correct their intuitions) in the Reasoning task. In Study 3 we presented participants with both a two-response Stroop task and a set of two-response reasoning tasks to explore this issue.

Study 1

In Study 1 we designed a two-response version of the Stroop task. On each trial participants were asked to give a first answer as fast as possible under cognitive constraints (time-pressure and secondary memorization task load), and to then take the time to reflect and provide a final constraint-free response. The key question is whether correct responding to the critical incongruent Stroop trials is also possible when participants' deliberate control is constrained. In those cases that participants managed to provide a correct final response, do they initially typically err or is the initial response already correct?

Method

Preregistration and data availability

The study design and hypothesis were preregistered on the Open science Framework (https://osf.io/9pz5j). No specific analyses were preregistered. All data and material are also available on the Open Science Framework (https://osf.io/gkhbm/).

Participants

We recruited our participants online on Prolific Academic (www.prolific.ac). Only native English speakers from Canada, Australia, New Zealand, the United States of America, or the United Kingdom were allowed to take part in the study. Participants were paid £2.60 for their participation (£5 hourly rate). Based on Aïte et al.'s (2016) Stroop study, we recruited 50 adult participants. The mean age of participants was 37.2 years (*SD* = 14.3) and 60% were female. Thirty-four percent of participants had a high-school degree as their highest education level, 50% had a bachelor's degree, 12% a Master's degree, and 4% had not completed high school.

Materials

Stroop stimuli. Based on Aïte et al. (2016), sixteen colour-word stimuli were created by combining four different colour names ('red', 'green', 'blue' and 'yellow') with four corresponding ink colours (RGB colour codes 255;0;0, 0;255;0, 0;0;255 and 255;255;0). We used these stimuli to create 64 congruent and 64 incongruent Stroop experimental trials. Before the main experiment, participants were presented with a set of practice trials (see "Two-response Stroop task" section below). For the colour practice, four circle stimuli were created each filled with either red, green, blue, or yellow ink (RGB colour codes 255;0;0, 0;255;0, 0;0;255 and 255;255;0).

All stimuli were presented in the center of the screen on a grey background (RGB code 135; 135; 135) in randomized order. Participants were instructed to press the key "d" if the word was presented in the colour red, the key "f" if it was presented in blue, they key "j" if it was presented in green and the key "k" if it was presented in yellow (we chose these four response keys as they have the same position in the three most common keyboard layouts: QWERTY, QWERTZ

and AZERTY). The response times were measured from the stimulus onset until the button press.

Congruent trials allowed us to test for a guessing confound and are reported in this context. Our main results concern the critical incongruent trials, unless otherwise stated.

Load task. In the two-response version of the Stroop task (see two-response section below), we used a secondary digit memorization task (Lavie 2005; Lavie et al., 2004), as this type of task has been shown to burden cognitive control in Stroop-like tasks (i.e., it has been found that this task increases the Stroop interference effect, e.g., de Fockert et al., 2001; Lavie & de Fockert, 2005; Lavie et al., 2004; but see also Gao et al., 2007). On each trial participants saw a sequence of six (black) digits (i.e., the memory set). All digits were randomly selected from 1 to 9 without replacement on a given trial. The memory probe consisted of a single black digit, a question mark and a message reminding participants of the keyboard response buttons. Participants were asked to indicate whether the probe had appeared in the memory set on that trial. They were instructed to press "d" for probe-present and "k" for probe-absent responses. For half of the trials the correct answer was "probe present".

Procedure

One-response (deliberative-only) pre-test. In order to obtain a baseline Stroop performance, we conducted a pre-test where participants performed a traditional one-response colour-word Stroop task, without a digit memorization load or a deadline. We recruited 25 participants (52% female; mean age = 35.4 years, *SD* = 17.3) online on Prolific Academic (www.prolific.ac). Only native English speakers from Canada, Australia, New Zealand, the United States of America, or the United Kingdom were allowed to take part in the study. Participants were paid £1.00 for their participation. A total of 40% of the participants reported a high-school degree as their highest education level, while 56% reported a bachelor's degree and 4% a Master's degree.

The idea was to base the response deadline of the initial response stage in our two-response design on the average response time in the one-response pretest (e.g., Bago & De Neys, 2017, 2020). Thus, in the one-response pretest participants were presented with the same amount of trials and the same stimuli

as in the main two-response study. The only difference from the main study was that participants were asked to provide a single response and they only received standard Stroop task instructions to respond "as fast and as accurate as possible". The average response time for the congruent trials was 755 ms (*SD* = 134 ms) and for the incongruent trials it was 893 ms (*SD* = 197 ms).[1] Based on these values we decided to set the maximum response deadline for the initial response to 750 ms (i.e., approximately the mean of congruent trials which do not require controlled processing to answer correctly).

To verify that participants were indeed under time pressure during the initial stage, we compared the response times for the critical incongruent trials between the one-response pre-test and the initial responses in the main two-response study. For this comparison, we excluded all trials with missed load memorization or missed deadlines in the initial stage of the two-response study. The results showed that participants responded much faster in the initial response stage of the main study (incongruent trials: 580.8 ms, *SD* = 55.4 ms), compared to that of the one-response study (incongruent trials: 893.3 ms, *SD* = 196.5 ms; i.e., responses were on average more than 1.5 SDs faster than in the one-response study). A Welch Two Sample t-test indicated that this difference was significant, $t(26.47) = 7.76$, $p < .001$.

In addition, the one-response pre-test allowed us to check for a potential consistency confound in our main two-response study. More specifically, since the study requires two consecutive responses, participants might provide the same response in the initial and the final stage, merely driven by the desire to appear consistent (Thompson et al., 2011). In this case, the correction rate from the initial to the final response would be underestimated. Previous two-response work in other fields has argued against the presence of this confound (Bago & De Neys, 2017, 2019a, 2020; Thompson et al., 2011). Here we tested for it by contrasting the proportion of correct responses in the incongruent Stroop trials of the one-response pretest and those of the final stage of the main two-response study. A consistency confound would result in a clear discrepancy between these accuracies. However, our results showed that the percentage of correct responses in the critical incongruent trials of the one-response pretest (*M* = 93.2%, *SD* =

[1] Before computing the average reaction times all trials with reaction times higher than 2 SDs above the general mean were removed from the analysis.

18.2%), was very similar to this of the incongruent final responses of the two-response study (*M* = 92.5%, *SD* = 18.6%). A Welch Two Sample t-test indicated that this difference was not significant, $t(52.32) = 0.14$, $p = 0.890$.

Two-response Stroop task. The experiment was run online on Gorilla Experiment Builder (gorilla.sc). Participants were informed that the study would take 30 minutes to complete and that it demanded their full attention. They were told that they would be presented with words and that they needed to respond to the colour that each word was presented in using their keyboard (for literal instructions see Supplementary Material section A). Then they were given instructions about the correct response key mapping.

To familiarise themselves with the colour-key pairs, participants first practiced only with the colours (without the words). They were presented with 32 colour stimuli (red, blue, green or yellow) and they were instructed to respond as fast and as accurately as possible. They were given feedback after each response and, in case of an incorrect response, they were shown a picture of a keyboard with the correct colour-key pairs. Figure 1A illustrates the time course of this practice round.

Then, participants were presented with a second practice round, which was identical to the first one, with the difference that now the stimuli were 12 congruent colour-word pairs. Participants were told that they needed to respond to the colour that each word was presented in.

After the second practice round, participants were introduced to the incongruent trials. They were informed that sometimes the ink colour in which the word appears would not match with the word, and they were asked to always respond to the colour of the word. This practice round was identical to the above two, with the difference that now the stimuli were eight incongruent colour-word pairs.

At the end of this practice round, participants were introduced to the two-response paradigm. They were told that we were interested in their initial, intuitive response to the colour of each word and wanted them to answer as fast as possible with the first response that popped up in mind. They were also informed that after the first response, they would have more time to reflect on the colour of the word and provide their final answer. Participants were introduced to the deadline of the initial response, and were shown an example of an initial trial. Then, they were

presented with 12 two-response colour-word trials. The time-course of this practice round can be seen in Figure 1B.

Following the two-response practice round, participants were presented with the load task. They were told that they also had to memorise a set of six numbers while responding to the colour-word pairs. Participants were informed that after the memory probe was shown, they would have to press "d" if the probe was part of the memory set, or "k" if the probe was not part of the memory set. At this point, they were presented with five load memorisation practice trials. Figure 1C illustrates the time course of this practice round.

After the load practice round participants were reminded that they had to memorise the set of numbers while responding to the colour-word pairs. They were instructed to first focus on the memorisation task, and then on the colour-word task. They were then presented with 24 two-response practice trials (with load and deadline). Critically, the first 12 practice trials had a looser initial response deadline (1 second instead of 750 ms). This was done to familiarise participants with the two-response format. For the last 12 practice trials the actual 750 ms deadline was applied. The time-course of this practice round was identical to that of the experimental trials and is illustrated in detail in Figure 1D.

After this practice session, participants started the experimental trials. The main task was composed of 128 trials which were grouped in 3 blocks. Participants were told that after each block they could take a short break. Before each new block started, they were shown a picture of a keyboard with the correct colour-key pairs to remind them of the response key mapping. At the end of the experiment, participants completed standard demographic questions and were presented with a debriefing message.

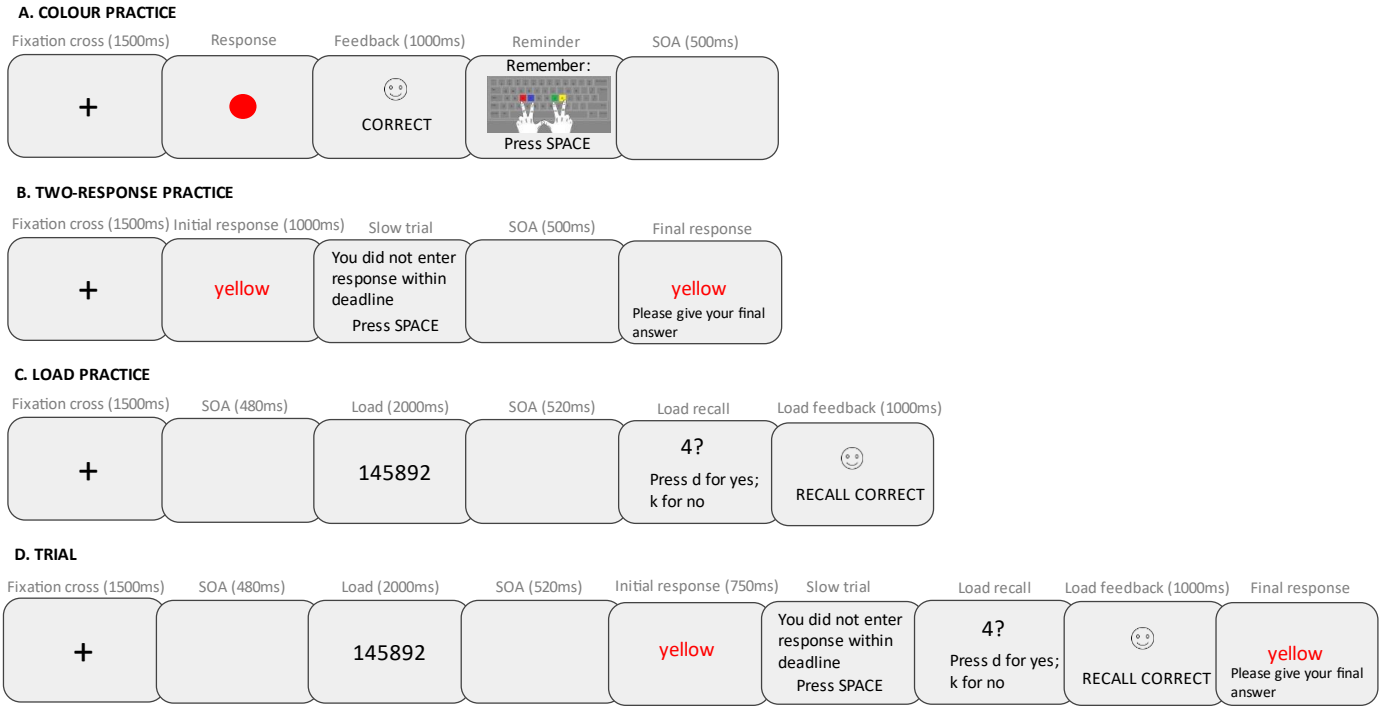

Figure 1. Time course of the practice trials and experimental trial. Panel A shows the time course of a colour-only practice trial. Panel B shows the time course of a deadline-only two-response practice trial. Panel C shows the time course of a load-only practice trial. Finally, Panel D shows the time-course of an experimental trial.

Exclusion criteria

Following our preregistration, we discarded from all analyses participants who scored lower than 50% on both their initial congruent and initial incongruent trials. This was done to sidestep the possibility that results would be distorted because some participants could not meet the initial trial constraints without guessing. Based on this criterion, 6 out of the 50 participants were excluded. We were thus left with a sample of 44 participants (59% female) with a mean age of 36.6 years (*SD* = 14.1).

In addition, we excluded the trials in which participants failed the load and/or the deadline, since in these trials we could not ensure that deliberation was minimized during the initial stage. Participants failed to answer before the deadline on 36.2% of incongruent initial trials (1019 out of 2816) and 25.4% of congruent initial trials (716 out of 2816). In addition, participants failed the load task on 9.6% of incongruent initial trials (269 out of 2816) and 12.6% of congruent initial

trials (355 out of 2816). Overall, we kept 58.1% of all trials (3273 out of 5632), by rejecting trials in which participants missed the deadline and failed the load task. On average, each participant contributed 74.4 trials (out of 128 trials, *SD* = 39.2). Clearly, the high amount of missed trials demonstrates that meeting the initial deadline and load constraints was challenging for participants. Note however that since we only discarded individual trials (rather than participants), this higher exclusion rate should not give rise to confounding individual selection effects (e.g., Bouwmeester et al., 2017).

Results and Discussion

Accuracy

Figure 2A gives an overview of the initial and final accuracies. As the figure indicates, overall, findings are in line with classic results. Participants typically managed to solve incongruent trials correctly when they were allowed to deliberate, although they performed better on congruent than incongruent trials. Regarding initial responses, we overall observed fairly high accuracy rates. For the congruent trials, the mean accuracy for initial responses was 82.6% (*SD* = 16.7%) and differed from 25% chance, $t(41) = 31.94$, $p < .001$, while for the critical, incongruent trials it was 67.3% (*SD* = 23.3%) and differed from 25% chance, $t(38) = 18.00$, $p < .001$. This suggests that participants were often able to produce correct responses when deliberation was minimized and they were forced to rely on intuitive, automatic processing. Although this is expected for congruent trials in which the intuitively cued response is correct, it suggests that correct responding on incongruent trials does not necessarily require deliberate controlled processing. To see if there was an effect of the response stage (initial; final) and the congruency status (congruent; incongruent) on the accuracy of the Stroop responses, a two-way within-subjects ANOVA was conducted. As Figure 2A shows, the accuracy for congruent trials was higher than for incongruent trials, $F(1, 44) = 15.06$, $p < .001$, $\eta^2g = 0.048$, and the accuracy at the final stage was higher than at the initial stage, $F(1, 44) = 65.83$, $p < .001$, $\eta^2g = 0.194$, indicating that accuracy improved after deliberation. Finally, the difference between initial and final accuracy was higher for incongruent compared to congruent trials, as indicated by the response stage by congruency interaction, $F(1, 44) = 11.08$, $p < .01$, $\eta^2g = 0.015$.

Note that, in theory, correct responding could result from random guessing. Since our test procedure is highly challenging, participants might not manage to process the stimuli, and might respond randomly instead. However, if that were true, accuracy rates should not differ between congruent and incongruent trials and should remain at chance levels throughout the study. It is clear from our findings that this is not the case.

In sum, the final accuracy findings are consistent with those of previous Stroop studies (e.g., AÏte et al, 2016). The key finding is the high initial accuracy rate on the incongruent trials. Although accuracy increased in the final stage, we frequently observed correct responding when deliberate control was minimized.

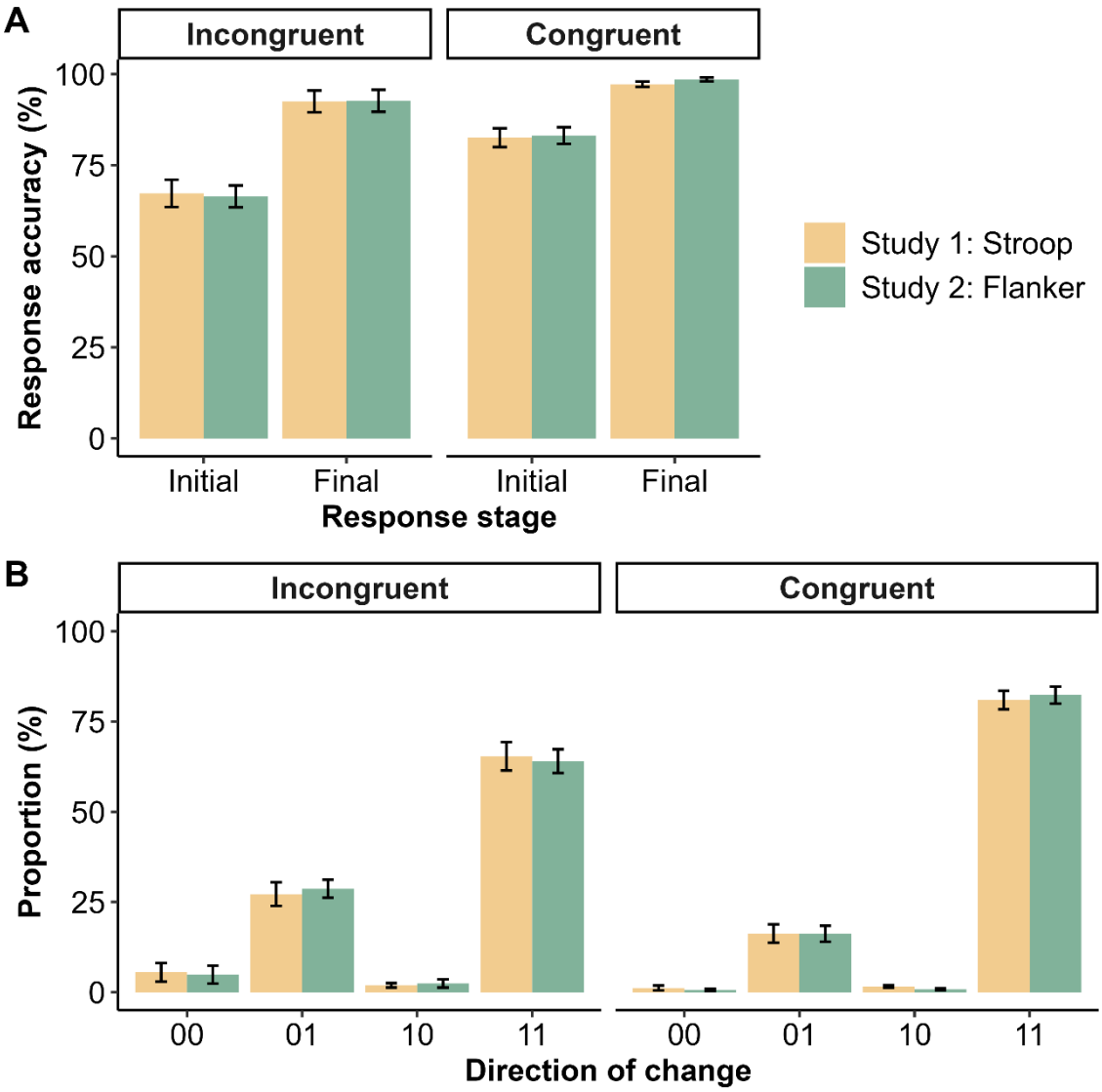

Figure 2. Accuracy and Direction of change in Study 1 (Stroop task) and Study 2 (Flanker task). **A)** Response accuracy at incongruent and congruent trials as a function of response stage. **B)** Proportion of each direction of change category at

incongruent and congruent trials. The error bars represent the Standard Error of the Mean. "00" = incorrect initial and incorrect final response; "01" = incorrect initial and correct final response; "10" = correct initial and incorrect final response; "11" = correct final and correct initial response.

Stability index

We also calculated a stability index for the initial responses of the critical, incongruent trials. Specifically, for each participant, we calculated on how many out of their initial responses in the incongruent trials they showed the same dominant accuracy (i.e., "0" or "1"; e.g., if out of 100 trials 60 were incorrect, the stability index would be 60%; similarly, if 60 trials were correct, the stability index would be 60% etc.). The average stability index was 76.1% (*SD* = 12.1%). If initial responding was prone to systematic guessing, we would expect more inconsistency in participants' initial responses across trials.

Direction of Change

To get a more precise picture of how participants changed their responses after deliberation, we also conducted a direction of change analysis (Bago & De Neys, 2017, 2019a). More specifically, we looked into how the accuracy changed (or did not change) from the initial to the final stage on every trial. In every stage, participants can either have an accuracy of "1" (i.e., correct response) or an accuracy of "0" (i.e., incorrect response). This way, we end up with four possible response patterns in each trial: "00" (incorrect initial and incorrect final response), "01" (incorrect initial and correct final response), "10" (correct initial and incorrect final response) and "11"(correct initial and correct final response).

Regarding the critical incongruent trials, as Figure 2B shows, the vast majority had a "11" pattern (65.4%). This high "11" proportion was also accompanied by a low "00" proportion (5.5%), and a low "10" proportion (1.9%). Critically, the proportion of "01" responses (27.2%) is lower than that of "11" responses. This indicates that, although deliberate correction occurs, in the majority of trials with correct final responses, the correct response was generated already from the initial stage. This so-called non-correction rate (i.e., proportion 11/11+01) reached 70.6%.

For completeness, as Figure 2B shows, a similar pattern was observed for congruent trials. In the vast majority of cases, correct responses were generated

intuitively. The non-correction rate reached 83%. Again, since intuitive, automatic processing is expected to cue the correct response on these trials, this pattern is not surprising.

Response mapping

A potential difficulty that arises from the specific Stroop task version we adopted is that participants may have struggled to apply the four-option color-response key mapping during the initial stage. In order to respond, participants first need to identify the color and then translate it into a button press. Despite the time and load constraints during the initial stage, participants likely had enough time to identify the colour. However, the complex four-response mapping may have interfered with translating the colour into a button press, which would lead to random guessing. If this was the case, the high accuracy observed in the initial stage could be attributed to guessing.

To examine this further, we looked into the types of errors participants made. Specifically, in the Stroop task, people could make two errors: lure errors (responding with the read word instead of the correct ink color) and non-lure errors (responding with any other incorrect ink color). If participants were responding randomly due to time and load constraints, we would expect non-lure errors to occur as frequently as lure errors (i.e., at a chance level of 66.6% and 33.3% respectively, considering that on each trial participants could make three different wrong button presses; two non-lure and one lure). If, however, participants had sufficient time to press the intended buttons, we would expect primarily lure errors, since participants would be influenced by the read word.

To examine this, we visualized the proportion of lure errors out of all initial errors for each participant separately as a function of initial accuracy (see Figure 3A). We conducted a binomial test for each participant's data, to determine if the lure error proportion exceeded the chance level of 33.3%. In the graph, green dots indicate a significant effect at $p < .05$ (one-tailed), blue dots indicate significance at $p < .01$, and red dots indicate a non-significant effect.

As expected, participants with very high initial accuracy rarely obtained a low p-value since they made very few errors (i.e., they had a low proportion of both lure and non-lure errors). Critically, however, the majority of data points in the upper right corner of Figure 3 are either green or blue. This means that even

among participants with high accuracy, the majority of errors were lure errors. This suggests that their high accuracy cannot be attributed to random guessing.

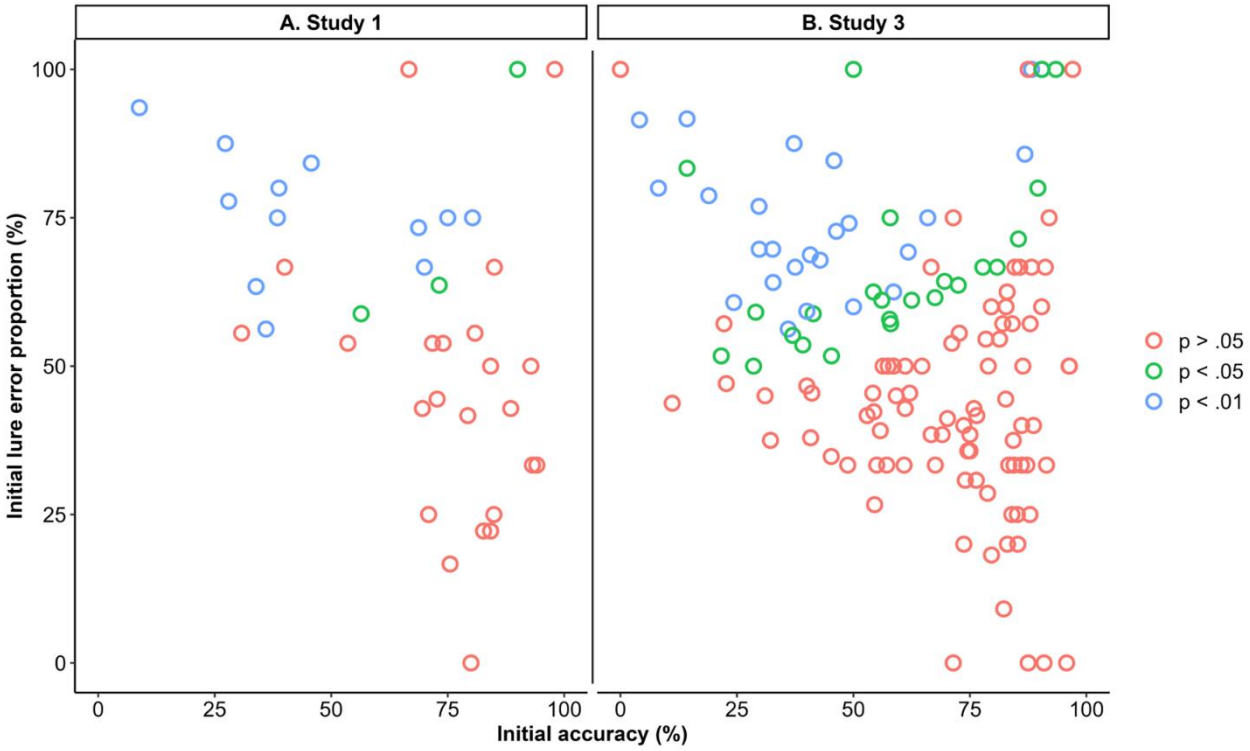

Figure 3. The initial lure error proportion (% of lure errors out of all errors) as a function of initial response accuracy, separately for each participant in Study 1 (left panel) and Study 3 (right panel). A binomial test was conducted for each participant to determine whether the proportion of lure errors exceeded the chance level of 33.3%. Red dots indicate a non-significant effect, green dots indicate a significant effect at $p < .05$ (one-tailed) and blue dots indicate significance at $p < .01$.

Finally, it is worth noting that a few participants had only non-lure errors. This could result from the fact that they were systematically wrong about the translation of colors into button presses. The existence of these participants indicates that the color-key mapping of the Stroop task was not trivial to learn.

Reaction Times

The average reaction time at the initial response stage was 543 ms (*SD* = 104 ms) for the congruent trials and 581 ms (*SD* = 55 ms) for the incongruent trials. This is much faster than the average reaction times usually found in previous Stroop studies (e.g., AÏte et al, 2016; Penner et al., 2012; Strauss et al.,

2005; Wright & Wanley, 2003) and our one-response control study. Together with the high percentage of missed trials, it shows that participants experienced considerable time pressure. Participants spent longer on the final response stage, with an average of 890 ms (*SD* = 873 ms) for congruent trials and 982 ms (*SD* = 744 ms) for incongruent trials. Supplementary Material section B gives a full overview of reaction times according to response accuracy.

Exploratory Analysis

To make maximally sure that participants did not deliberate during the initial response stage, we excluded a considerable amount of trials. In theory this could have artificially boosted the critical non-correction rate. That is, if these excluded trials would be specifically of the "01" type, the true non-correction rate would obviously be lower suggesting that correct intuitive response generation would be much rarer than reported here. To examine this possibility, we re-ran the direction of change analysis while including all missed load and missed deadline trials. Since in the missed deadline trials the initial response was not recorded, we opted for the strongest possible test and coded all these as "0" (i.e., incorrect response). In the missed load trials both initial and final responses were recorded. The analysis (see Supplementary Material section C for full results) pointed to a higher proportion of "01" incongruent trials (47.2%), but the proportion of "11" (41.2%) responses and the non-correction rate remained high (46.6%). Hence, even in this extremely conservative analysis, correct incongruent responses were still generated intuitively about half of the time.

Study 2

Study 1 showed that when participants gave a correct final Stroop response they had typically already generated a correct response in the initial stage. This indicates that correct responding in the Stroop task can occur even when deliberate control is minimized. However, Study 1 was but the first to adopt the two-response paradigm with a classic cognitive control task. Thus, it is important to test the generalizability of these findings to another classic cognitive control task before drawing strong conclusions. Therefore, in Study 2 we designed a two-response version of the Flanker task. Since the Flanker task is a binary-response task, it also allowed us to sidestep the difficulty of the specific Stroop task

response format we adopted in Study 1, namely that participants may have found it challenging to apply the four-option color-response key mapping in the initial stage.[2] As in Study 1, the key question is whether participants can provide correct responses to the critical incongruent Flanker trials when their deliberate control is constrained.

Method

Preregistration and data availability

The study design and hypothesis were preregistered on the Open science Framework (https://osf.io/eqdks). No specific analyses were preregistered. All data are also available on the Open Science Framework (https://osf.io/gkhbm/).

Participants

We recruited our participants online on Prolific Academic (www.prolific.ac). Only native English speakers from Canada, Australia, New Zealand, the United States of America, or the United Kingdom were allowed to take part in the study. Participants were paid £2.40 for their participation (£6 hourly rate).[3] For consistency with Study 1 we recruited 50 adult participants. The mean age of participants was 38.6 years (*SD* = 14.6) and 58% were female. Thirty-eight percent of participants had a high-school degree as their highest education level, 46% had a bachelor's degree, 12% a Master's degree, 2% a doctoral degree, and 2% had not completed high school.

Materials

Flanker stimuli. The stimuli consisted of a row of five arrows. This row included a central arrow flanked by two surrounding arrows on each side, all with arrowheads pointing either to the left or to the right. In congruent stimuli, the surrounding arrows pointed in the same direction as the central arrow (←←←←← or →→→→→). In incongruent stimuli, the surrounding arrows pointed in the opposite direction to the central arrow (←←→←← or →→←→→).

[2] However, note that although both the Flanker task and the Stroop task involve conflict resolution in the incongruent trials, they tap into different aspects of cognitive control, and while the Stroop involves semantic conflict, the Flanker involves a more perceptual conflict (Ridderinkhof et al., 2021; see Method and General Discussion).

[3] The hourly rate in this study is £6 instead of the £5 hourly rate of Studies 1 and 3, as Prolific increased their minimum pay by the time Study 2 was run.

A total of 128 experimental trials, consisting of 64 congruent and 64 incongruent trials, were presented to the participants in a randomized order. The stimuli were presented in the center of the screen on a white background. Participants were instructed to press the "f" key if the central arrow pointed left and the "j" key if it pointed right. Response times were measured from the onset of the stimulus until the button press. Our main results concern the critical incongruent trials, unless otherwise stated.

As we noted, since the Flanker task is a binary-response task, it also allows us to sidestep a potential difficulty of the specific Stroop task response format we adopted in Study 1, namely that participants may have found it challenging to apply the four-option color-response key mapping in the initial stage. However, in theory, the version of the Flanker task that we used may present its own limitations. For example, one may note that in the congruent trials it is not necessary to focus attention on the central arrow, since all items are identical, but in the incongruent trials participants need to focus their attention on the central arrow to produce a correct response. This may invite an alternative strategy that people can use: they can first determine whether all items are the same and, if they are not, they can focus their attention on the central target only. Since focusing takes time this strategy could generate longer reaction times in the incongruent, compared to the congruent trials. In this sense the Flanker task would not necessarily evoke response conflict like the Stroop. However, the evidence in the cognitive control literature with the specific version of the Flanker task (with a 1-to-1 response mapping) we adopted suggests that this alternative account is insufficient to explain the entirety of the flanker effect (e.g., Hübner et al., 2010) and may not even play a significant role in contributing to it (Servant & Logan, 2019). That is because participants focus attention on the central arrow in a similar way in congruent and incongruent trials (Servant & Logan, 2019). This supports the original interpretation of the Flanker, which emphasizes response competition as a key factor in the task (Eriksen & Eriksen, 1974; Eriksen & Hoffman, 1973). Nevertheless, it remains the case that the Stroop and Flanker tasks may tap different aspects of cognitive control (e.g., Friedman & Miyake, 2004; Rey-Mermet et al., 2018; see also General Discussion).

Load task. In the two-response version of the Flanker task, we used the same secondary digit memorization task as in the Stroop task of Study 1 (Lavie 2005;

Lavie et al., 2004), since it has been shown to burden cognitive control in classic control tasks (Lavie et al., 2004).

Procedure

One-response (deliberative-only) pre-test. To obtain a baseline Flanker performance, we ran a pre-test where participants performed a traditional one-response arrow Flanker task, without a digit memorization load or a deadline. As in Study 1, we recruited 25 participants (48% female; mean age = 36.4 years, *SD* = 11.0) online on Prolific Academic (www.prolific.ac). Only native English speakers from Canada, Australia, New Zealand, the United States of America, or the United Kingdom were allowed to take part in the study. Participants were paid £0.70 for their participation. A total of 40% of the participants reported a bachelor's degree as their highest education level, while 28% reported a Master's degree and 28% a high school degree.

The deadline of the initial response stage in our two-response design was based on the average response time of the one-response pretest (e.g., Bago & De Neys, 2017, 2020). Thus, the one-response pretest was similar to the main study in terms of stimuli and amount of trials, but participants were instructed to provide a single response on each trial and to answer "as fast and as accurate as possible". The average response time for the congruent trials was 428 ms (*SD* = 52.2 ms) and for the incongruent trials it was 458 ms (*SD* = 47.2 ms).[4] Based on these values we decided to set the maximum response deadline for the initial response to 420 ms (i.e., approximately the mean of congruent trials which do not require controlled processing to answer correctly).

To confirm that participants were under time pressure in the initial stage, we compared response times for critical incongruent trials between the one-response pre-test and the initial responses in the main two-response study. We first excluded all trials with missed load memorization or missed deadlines in the initial stage of the two-response study. The results showed that participants responded much faster in the initial response stage of the main study (incongruent trials: 314.6 ms, *SD* = 43.7 ms), compared to that of the one-response study (incongruent trials: 457.6 ms, *SD* = 47.2 ms; i.e., responses were on average

[4] Before computing the average reaction times all trials with reaction times higher than 2 SDs above the general mean were removed from the analysis.

more than 2.5 SDs faster than in the one-response study). A Welch Two Sample t-test indicated that this difference was significant**,** $t(3222.88) = 55.80$, $p < .001$.

The one-response pre-test also allowed us to check for a potential consistency confound in our main two-response study, which could potentially underestimate the correction rate from initial to final responses. To test for this confound, we compared the accuracy of incongruent trials between the final two-response stage of our main study ($M = 92.7\%$, $SD = 20.6\%$) and the pretest ($M = 97.5\%$, $SD = 2.5\%$). Although a Welch Two Sample t-test revealed a significant difference, $t(2792.13) = 5.77$, $p < .001$, this difference was small. Even if we factor in a possible 5% extra correction trials (i.e., "01" trials) in our results, the non-correction rate conclusions remain unaffected (i.e., 65.5% with the extra correction trials vs. 69% without). Therefore, a potential consistency confound cannot explain the low correction rates.

Two-response Flanker task. The experiment was run online on Gorilla Experiment Builder (gorilla.sc). Participants were informed that the study would take 20 minutes and that it required their full attention. They were told that they would be presented with an arrow at the center of the screen, and that they had to press the button that matched the arrow's direction. Specific instructions about the key mapping were provided. Participants were then told that the central arrow would always appear along with four other arrows, and that their task was to identify the direction of the central arrow (for literal instructions see Supplementary Material section A).

To familiarise themselves with the key mappings, participants first practiced with 6 trials (3 congruent and 3 incongruent). They were given feedback after each response and in case of an incorrect answer they were reminded of the correct key pairs.

At the end of this practice round, participants were introduced to the two-response paradigm. They were told that we were first interested in their initial, intuitive response to the direction of the central arrow and wanted them to answer as fast as possible with the first response that came to mind. They were told that after this first response, they would have more time to reflect before providing their final answer. Participants were introduced to the deadline of the initial response, and were shown an example of an initial trial. Then, they were presented with 6 two-response trials.

Following the two-response practice round, participants were presented with the load task, with the same instructions as in Study 1. They were then presented with five load memorisation practice trials.

After the load practice round, participants were reminded that they had to memorise the numbers while responding to the direction of the central arrow. They were instructed to first focus on the memorisation task, and then on the arrow task. They were then presented with 12 two-response practice trials (with load and deadline). Critically, the first 6 practice trials had a looser initial response deadline (670 ms instead of 420 ms). This was done to familiarise participants with the two-response format. For the last 6 practice trials the actual 420 ms deadline was applied.

After this practice session, participants started the experimental trials. The main task was composed of 128 trials which were grouped in 3 blocks. Participants were told that after each block they could take a short break. Before each new block started, they were reminded of the response key mapping. At the end of the experiment, they completed standard demographic questions and were presented with a debriefing message.

Exclusion criteria

Like in Study 1 and following our preregistration, we discarded from all analyses participants who scored lower than 50% on both their initial congruent and initial incongruent trials. Based on this, 1 out of the 50 participants was excluded. We were thus left with a sample of 49 participants (57% female) with a mean age of 38.6 years (*SD* = 14.6).

In addition, we excluded the trials in which participants failed the load and/or the deadline. Participants failed to answer before the deadline on 39.0% of incongruent initial trials (1222 out of 3136) and 29.3% of congruent initial trials (919 out of 3136). In addition, participants failed the load task on 8.2% of incongruent initial trials (257 out of 3136) and 12.1% of congruent initial trials (381 out of 3136). Overall, we kept 55.7% (3493 out of 6272), by rejecting trials in which participants missed the deadline and failed the load task. On average, each participant contributed 71.3 trials (out of 128 trials, *SD* = 32.0). As in Study 1, the high number of missed trials indicates that meeting the deadline and load constraints was challenging for participants.

Results and Discussion

Accuracy

Figure 2A gives an overview of the initial and final accuracies. Overall the results are very similar to that of the Stroop task of Study 1. Participants generally performed better on congruent trials, but they also managed to solve most incongruent trials correctly when deliberate processing was allowed. Initial responses showed high accuracy rates both for congruent (M = 83.2%, SD = 15.8%) and critical incongruent trials (M = 66.5%, SD = 20.5%) and they both differed from 50% chance, $t(47) = 14.58$, $p < .001$ and $t(45) = 5.45$, $p < .001$, respectively. This suggests that participants often produced correct responses even when relying on mere intuitive processing. So, as in the Stroop task, correct responding in the Flanker task does not necessarily require deliberate controlled processing. A two-way within-subjects ANOVA on the effect of response stage and congruency status on response accuracy, revealed that accuracy was higher for congruent trials, $F(1, 45) = 22.21$, $p < .001$, $\eta^2g = 0.10$, and that accuracy at the final stage was higher than that at the initial stage, $F(1, 45) = 100.38$, $p < .001$, $\eta^2g = 0.29$. This difference between initial and final accuracy was higher for incongruent compared to congruent trials, as indicated by the response stage by congruency interaction, $F(1, 45) = 10.70$, $p < .01$, $\eta^2g = 0.02$.

In sum, these results align with the Stroop results of Study 1 and show that correct responding in the incongruent trials of the Flanker task is possible when deliberate control is minimized.

Stability index

We also calculated a stability index for the initial responses of the critical, incongruent trials. More specifically, for each participant we again calculated on how many out of the their initial responses in the incongruent trials they showed the same dominant accuracy (i.e., "0" or "1"). The average stability index was 71.8% (SD = 14.5%). If initial responding was prone to systematic guessing, we would expect more inconsistency in participants' initial responses across trials.

Direction of Change

To get a more precise picture of how participants changed their responses after deliberation, we again conducted a direction of change analysis (Bago & De Neys, 2017, 2019a). Regarding the critical incongruent trials, as Figure 2B shows, the vast majority had a "11" pattern (64.0%). This high "11" proportion was also accompanied by a low "00" proportion (4.9%), and a low "10" proportion (2.4%). Critically, the proportion of "01" responses (28.7%) was lower than that of "11" responses. This indicates that, although deliberate correction occurs, in the majority of trials with correct final responses the correct response was generated already from the initial stage. The non-correction rate (i.e., proportion 11/11+01) reached 69%. As it was expected and as Figure 2B shows, in the vast majority of congruent trials correct responses were generated intuitively and the non-correction rate reached 83.6%.

Reaction Times

The average reaction time at the initial response stage was 305.6 ms (*SD* = 58.3 ms) for the congruent trials and 314.6 ms (*SD* = 43.7 ms) for the incongruent trials. This is much faster than the average reaction times found in previous Flanker studies with similar amount of trials (e.g, Abutalebi et al., 2012; Fan et al., 2005) and our one-response control study. Together with the high percentage of missed trials, it shows that participants experienced considerable time pressure. Participants spent longer on the final response stage, with an average of 515.2 ms (*SD* = 225.1 ms) for congruent trials and 543.2 ms (*SD* = 260.2 ms) for incongruent trials. Supplementary Material section B gives a full overview of reaction times according to response accuracy.

Exploratory Analysis

To ensure that participants did not deliberate during the initial response stage, we excluded a considerable amount of trials, which could have potentially inflated the non-correction rate. To examine this possibility, we re-ran the direction of change analysis while including all missed load and missed deadline trials. As in Study 1, we opted for the strongest possible test and coded all missed deadline trials as "0" (i.e., incorrect response). In the missed load trials both initial and final responses were recorded. The analysis (see Supplementary Material

section C for full results) pointed to a higher proportion of "01" incongruent trials (52.3%), but the proportion of "11" (39.2%) responses and the non-correction rate remained high (42.8%). Hence, even in this extremely conservative analysis, correct incongruent responses were still generated intuitively about 43% of the time.

Study 3

Studies 1 and 2 showed that in both the Stroop and Flanker tasks, when participants provided a correct final response, they had typically already generated a correct response in the initial intuitive stage. This indicates that correct responding in cognitive control tasks is possible even when deliberate control is minimized. The first aim of Study 3 was to replicate the Stroop findings of Study 1 on a larger scale. The second aim was to explore whether individual performance in the Stroop task, both at the initial and final stage, correlates with performance in classic heuristics-and-biases tasks.

Study 3 comprised two parts: a Colour-Word Stroop task followed by a Reasoning task consisting of a battery of heuristics-and-biases reasoning problems. We used a two-response paradigm (Thompson et al., 2011) for both the Stroop and the Reasoning task.

Method

Preregistration and data availability

The study design and hypothesis were preregistered on the Open science Framework (https://osf.io/dm7h9). No specific analyses were preregistered. All data and material are also available on the Open Science Framework (https://osf.io/yqkm7/).

Participants

We recruited our participants online on Prolific Academic (www.prolific.ac). Only native English speakers from Canada, Australia, New Zealand, the United States of America, or the United Kingdom were allowed to take part in the study. Participants were paid £4.50 for their participation (£5 hourly rate). Based on the Raoelison et al. (2020, Study 2) correlational two-response study, we aimed to recruit 160 participants. Due to a software error, the Reasoning task data of one

participant could not be recovered, so we ended up with 159 participants (69.2% female), with a mean age of 33.1 years (*SD* = 13.6). This allowed us to pick up small to medium size correlations (.22) between the Stroop and Reasoning task performance with a power of 80%. The majority of participants (45%) had a high-school degree as their highest education level, 37% had a bachelor's degree, 15% a Master's degree, and 3% had not completed high school.

Materials

The Stroop task was run on Gorilla Experiment Builder (gorilla.sc) and the Reasoning task was run on the Qualtrics (www.qualtrics.com) software server. We first ran an initial batch of 10 participants that was identical to the main study. This was done to ensure that no technical problems would occur during the transition from Gorilla Experiment Builder to the Qualtrics platform. Data from one participant of this first batch could not be analysed (see above). We then ran the main study batch, which consisted of the remaining 150 participants.

The Colour-Word Stroop task that was used in this study was identical to the Stroop task described in Study 1.

The Reasoning task included three different types of reasoning problems (i.e., bat-and-ball problems, base-rate problems, and syllogistic reasoning problems). We used the exact same two-response format (response deadlines and load, see below) that was validated for these tasks in previous work (Bago & De Neys 2017, 2019a; De Neys, 2006). To avoid confusion, it is important to stress that the deadline and the concurrent cognitive load of the Reasoning task differs from that of the Stroop task. As a reminder, the goal of these two constraints is to minimize deliberation involvement and enforce intuitive thinking. However, there is no gold standard procedure which can ensure that people will respond intuitively, and the definition of "limited cognitive resources" always depends on the task at hand. For example, heuristics-and-biases tasks are lengthy (e.g., a couple of preamble sentences and response option reading), so deadlines are based on the pretested average reading times which are usually a couple of seconds (participants need to have the minimum time to read the problem before responding). On the contrary, Stroop responding is considerably faster since participants only see a single stimulus (i.e., word), so a strict deadline necessarily cannot be much longer than a single second. The same goes for the cognitive load, whose goal is to burden participants' cognitive resources. The strain on resources

may depend on the specific nature of the task. That is why, for each of our tasks, we opted for a load that has been independently shown in the literature to burden cognitive resources and decrease performance in this specific type of task.

Counterbalancing. Each of the three types of reasoning problems was composed of eight incongruent and eight congruent items. For every type of problem, we created two sets of items. In each set, the congruency status of the items was counterbalanced. More specifically, all the incongruent items of the first set appeared in their congruent version in the second set, and all the congruent items in the first set appeared in their incongruent version in the second set. Half of the participants were presented with the first set while the other half were presented with the second set. This way, the same item content was never presented more than once to a participant and, at the same time, everyone was exposed to the same items, which minimized the possibility that mere item differences influence the results (e.g., Bago & De Neys, 2017). The presentation order of the items within each task was randomized. Each participant was randomly allocated to one of six potential task orders. More specifically, each participant was first randomly allocated to task 1/3 (i.e., either bat-and-ball, base-rates or syllogisms), and then they were randomly allocated to one of the two potential task order combinations for the second and third task (e.g., if a given participant had the bat-and-ball as their first task, they could continue with base-rates as their second task and syllogisms as their final task, or the inverse).

Bat-and-ball problems (BB). Each participant was presented with eight multiple-choice bat-and-ball items (four incongruent and four congruent) taken from Bago and De Neys (2019a). The prices and the names of the objects varied between items, but all the items shared the same structure with the classic bat-and-ball problem. Participants were always presented with four response options: the logical option ("5 cents" in the original bat-and-ball), which is considered correct, the heuristic option ("10 cents" in the original bat-and-ball), and two foil options. The two foil options were always the sum of the correct and heuristic answer (e.g., "15 cents" in original bat-and-ball units) and their second greatest common divisor (e.g., "1 cent" in the original). An example of the problems is presented below:

A pencil and an eraser cost $1.10 in total.

The pencil costs $1 more than the eraser.

How much does the eraser cost?

- *5 cents*
- *1 cent*
- *10 cents*
- *15 cents*

The congruent versions were constructed by removing the "more than" statement from the incongruent versions ("A pencil and an eraser cost $1.10 in total. The pencil costs $1. How much does the eraser cost?"). Each problem was presented serially. First, the first sentence, which always stated the two objects and their total cost (e.g., A pencil and an eraser cost $1.10 in total.) was presented for 2000 ms. Afterward, the second sentence along with the question and the answer options was added under the first sentence (which remained on screen). The problem remained on screen until a response was given or until the deadline. As in Bago and De Neys (2019a), the deadline for the initial response was 5000 ms.[5]

Base-Rate problems (BR). Each participant was presented with eight base-rate items (four incongruent and four congruent) taken from Bago and De Neys (2017). Each item consisted of a sentence describing the composition of a sample (e.g., "This study contains scientists and assistants."), a sentence with a stereotypical description of a random person from the sample (e.g., "Person 'C' is intelligent."), and a sentence with the base-rate information (e.g., "There are 4 scientists and 996 assistants."). Participants had to indicate to which group the random person most likely belonged to. The answer option that was considered correct was always the one that corresponded to the largest group in the sample. The presentation of all items was based on Pennycook et al.'s (2014) rapid-response paradigm. Each sentence was presented serially and the amount of text presented on the screen was minimized. An example of the problems is presented below:

[5] The specific deadlines in each type of problem were based on pilot reading and one-response pretests (see respective subsections) and have been shown to create substantial time pressure.

This study contains scientists and assistants.

Person 'C' is intelligent.

There are 4 scientists and 996 assistants.

Is Person 'C' more likely to be:

- *A scientist*
- *An assistant*

The congruent versions were constructed by reversing the base-rates of the incongruent versions. For example in its congruent version, the second sentence of the above problem would read "There are 996 scientists and 4 assistants". Each problem was presented in three stages. First, the first sentence was presented for 2000 ms. Then, the second sentence was added under the first sentence (which remained on screen) for another 2000 ms. Finally, the critical base-rate information along with the question and the answer options were added until a response or until the deadline. As in Bago and De Neys (2017), the deadline for the initial response was 3000 ms.

Syllogistic reasoning problems (SYL). Each participant was presented with eight syllogistic reasoning items (four incongruent and four congruent), taken from Bago and De Neys (2017). Each item consisted of a major premise (e.g., "All things made of wood can be used as fuel."), a minor premise (e.g., "Trees can be used as fuel.") and a conclusion (e.g., "Trees are made of wood."). Participants were told to always consider the premises as true and were asked to say if the conclusion followed logically from the premises or not. A conclusion was considered logical only when it was valid. An example of the problems is presented below:

All things made of wood can be used as fuel

Trees can be used as fuel

Trees are made of wood

Does the conclusion follow logically?

- *Yes*
- *No*

In the incongruent items, the believability and the validity of the conclusion conflicted. More specifically, the conclusion of the incongruent items was either valid-unbelievable or invalid-believable. For instance, in the above example of an incongruent problem the syllogism is believable, but invalid. For the congruent

items, the validity of their conclusion was in accordance with their believability. Meaning that the conclusion was either valid-believable or invalid-unbelievable. For example, in its congruent version, with a valid-believable conclusion, the above problem would read: "All things made of wood can be used as fuel. Trees are made of wood. Trees can be used as fuel." Each problem was presented in three stages. First, the first sentence of the problem was presented for 2000 ms. Then, the second sentence was added under the first sentence (which remained on screen) for 2000 ms. Finally, the conclusion along with the question and the answer options were added until a response was given or until the deadline. As in Bago and De Neys (2017), the deadline for the initial response was 3000 ms.

Load task. For the Stroop task, we used the same digit memorization task (Lavie, 2005; Lavie et al., 2004) as in Study 1. For the Reasoning task, the load memorization task that was used was a complex visual pattern (i.e., 4 crosses in a 3 × 3 grid, see Bago & De Neys, 2017, 2019a; Raoelison & De Neys, 2019), which was briefly presented before each reasoning problem (Miyake et al., 2001). After providing an initial response to the reasoning problem, participants were presented with four different load patterns (i.e., with different cross placings) and had to identify the one that they had been asked to memorize. Miyake et al. (2001) showed that this task burdens cognitive resources, and previous studies have shown that it hampers sound deliberating and decreases reasoning accuracy on the specific types of reasoning problems we adopted (e.g., De Neys, 2006, Franssens & Neys, 2009; Johnson et al., 2016).

Composite reasoning measure. For simplicity and to maximize power, our analyses focused on the composite incongruent accuracy across the three different reasoning problem types (i.e., bat-and-ball, base-rates, syllogisms). To calculate the composite performance, we averaged for each participant the proportion of correct initial and final responses, separately for each problem type. Then we averaged across all problem types (separately for initial and final trials). For completeness, we calculated the composite performance also for congruent trials.

For the main correlational analysis between the Stroop and the Reasoning task, we first calculated the z-scores separately for each participant, each problem type, each response stage (i.e., initial, final), and each direction of change category (see further). Then, we averaged the z-scores across the three problem types, separately for each response stage and each direction of change category.

It is important to clarify that because of practical limitations, we did not have a composite cognitive control measure. Thus, we tested whether the composite reasoning measure correlated with performance at the Stroop task only.

Procedure

Participants were informed that the study would take 55 minutes to complete and that it demanded their full attention. They were told that the experiment was divided into two parts (i.e., the Stroop task and the Reasoning task). All participants begun the experiment with the Stroop task, and once they finished, they were redirected to the Reasoning task. The Stroop task's procedure was identical to the one described in Study 1. Once participants started the Reasoning task, they were told that it consisted of three different types of reasoning problems (i.e., bat-and-ball, base-rates and syllogisms). Then, they were told that they would have to provide two consecutive responses to various items. They were instructed to first answer with the very first answer that came to their mind and then reflect on the problem before providing their final response (see Raoelison et al., 2020, for literal instructions).

Afterwards, participants were presented with instructions specific to each problem type. Each problem type made up a block of the task and the three different types were presented in a pseudorandomized order (see Counterbalancing). Every problem type was introduced with a short transition text which indicated the participant's progress (e.g., "You are going to start task 1/3. Click on Next when you are ready to start task 1."). Then, the presentation format of the respective problem type was explained, an example problem was shown, and the deadline of the initial response was introduced. After these instructions, participants solved two practice items (without a concurrent load task) to familiarize themselves with the presentation format. Next, they solved two practice matrix recall items (without a concurrent reasoning problem). Finally, they solved the two earlier practice items with a concurrent load task.

Each trial started with a fixation cross that was shown for 1000 ms. Next, the target pattern for the memorization task was presented for 2000 ms. Then the first part of the problem was presented (for more details see Materials subsections for each problem type). Afterwards, the whole problem was presented along with the question and the answer options. Participants could provide their

initial response by clicking on one of the answer options. One second before the deadline, the screen turned yellow to remind participants of the upcoming deadline. If they did not respond within the deadline, they were presented with a message asking them to try and respond within the deadline on the next trials. If they responded within the deadline, they were asked to rate their confidence in the correctness of their initial response on a scale from 0 (absolutely not confident) to 100 (absolutely confident).[6] After entering their confidence, participants were shown four matrix patterns and were asked to recall the correct, to-be-memorized pattern. They were then given feedback on whether their recall was correct or not. Finally, participants viewed the full problem again and were asked to provide their final answer. Next, they were asked to report their confidence in the correctness of their final response. After responding to all the items of a problem type, a transition message appeared to indicate participants' progress (e.g., "You are going to start task 2/3. Click on Next when you are ready to start task 2."). At this point the next problem type was introduced.

After participants had responded to all three problem types, they were shown the classic bat-and-ball problem and were asked whether they had seen or read about this specific problem before (Yes/No). Immediately afterwards they were asked to provide an answer to the problem ("What do you think the correct answer is? Please enter it below"). Finally, participants were asked to complete standard demographic questions and were shown a debriefing message.

Exclusion criteria

As in Study 1, we discarded from all analyses participants who scored lower than 50% on both their initial congruent and initial incongruent Stroop trials. As a result, 13 out of the 159 participants were excluded. We were thus left with a sample of 146 participants (59% female), with a mean age of 36.6 years (*SD* = 14.1).

Stroop task. Participants did not respond within the deadline on 18.1% congruent initial trials (1690 out of 9344) and 29.2% of incongruent initial trials (2728 out of 9344). In addition, participants failed the load recall on 15.2% of congruent initial trials (1416 out of 9344) and 10.5% of incongruent initial trials (979 out of

[6] The confidence was recorded both at the initial and the final responses simply for a comparison with previous reasoning findings (see Supplementary Material section F).

9344). By rejecting the trials with a missed deadline and an incorrect load recall, we kept 63.5% of all trials (11875 out of 18688). On average, each participant contributed 81.3 trials (out of 128 trials, *SD* = 31.5).

Reasoning task. The trials in which participants failed the load and/or the deadline were excluded from subsequent analyses. Participants failed to answer before the deadline on 5.4% of incongruent initial trials (103 out of 1908) and 2.7% of congruent initial trials (52 out of 1908). In addition, participants failed the load recall on 12.5% of incongruent initial trials (239 out of 1908) and 14.7% of congruent initial trials (281 out of 1908). By rejecting the trials with a missed deadline and an incorrect load recall, we kept 82.3% of all trials (3141 out of 3816). On average, each participant contributed 19.8 trials (out of 24 trials, *SD* = 3.1).

Since the bat-and-ball problem has become very popular, some participants may have been previously exposed to the correct "5 cents" answer. If this is the case, they would not need to override an initially incorrect, heuristic response in order to arrive at the correct answer when solving the problem, which could distort our results. Following Raoelison et al. (2020), we therefore asked participants whether they had seen/solved the bat-and-ball problem before or if they had read about it (see "Procedure"). We also asked them to provide an answer to the problem ("What do you think the correct response is? Please enter it below."). The bat-and-ball trials were excluded for all participants that reported having seen the original bat-and-ball problem and that were able to provide the correct "5 cents" response.[7] Their trials for the other tasks were included in the analysis. In total, we excluded from the analysis an additional 440 bat-and-ball trials (i.e., 5.8% of all trials) from 32 participants. Note that, 56 of the bat-and-ball trials of these participants were already excluded because of missed deadline or load.

Results and Discussion

Stroop Task

In Study 3, we replicated the main findings observed in the Stroop task of Study 1, with a much larger sample. Specifically, we found that participants can

[7] The answer to this question was in free-response format. The responses that were considered as correct were: 5 cents, 5 CENTS, 5c, 5, $0.05, 0.05, .05, 0.5.

typically provide correct Stroop responses, even when deliberate control is minimized. The mean accuracy at congruent trials was 63.6% (*SD* = 24.3%) in the initial response stage and 91.5% (*SD* = 17.9%) in the final stage, while the non-correction rate (i.e., proportion 11/11+01) reached 66.5%. These results again suggest that, more often than not, correct Stroop responses are generated in the absence of deliberate controlled correction. For brevity, the full results of the Stroop task of Study 3 are reported in Supplementary Material section D.

Reasoning Task

Accuracy. Figure 4 gives an overview of the initial and final Reasoning task accuracies. Although we focus our analysis on the composite reasoning performance, individual task trends are reported in the graphs for completeness. The overall pattern is very similar to what was observed in previous two-response studies. First, people perform well on congruent trials both at the initial (*M* = 90.0%, *SD* = 7.2%) and the final stage (*M* = 92.6%, *SD* = 7.1%), while incongruent trials typically have low initial (*M* = 35.2%, *SD* = 20.5%) and final (*M* = 40.6%, *SD* = 22.6%) accuracies. This indicates that even after deliberation, the majority of reasoners remain biased (Bago & De Neys, 2017, 2019a; Raoelison & De Neys, 2019; Raoelison et al., 2020). As it can be seen in Figure 4, these composite level trends were also observed for each individual task separately.

In addition, note that consistent with previous findings, reasoners' accuracy at the incongruent trials is typically below or near 50%, (and close to guessing accuracy). However, the high accuracy on the congruent trials confirms that participants are not merely guessing throughout the study. Instead, they are simply lured by the heuristic cue when solving the incongruent items.

To examine whether there was an effect of the response stage (initial; final) and congruency status (incongruent; congruent) on response accuracy, a two-way within-subjects ANOVA was conducted. As Figure 4 shows, the accuracy at the congruent trials was higher than that at the incongruent trials, $F(1, 158) = 402.54$, $p < .001$, $\eta^2g = 0.518$, and the accuracy at the final stage was higher than that at the initial stage, $F(1, 158) = 29.91$, $p < .001$, $\eta^2g = 0.008$, showing that accuracy improved after deliberation. Finally, this difference between initial and final accuracy was higher for incongruent compared to congruent trials, as indicated by the response stage by congruency interaction, $F(1, 158) = 4.08$, $p < .05$, $\eta^2g = 0.001$.

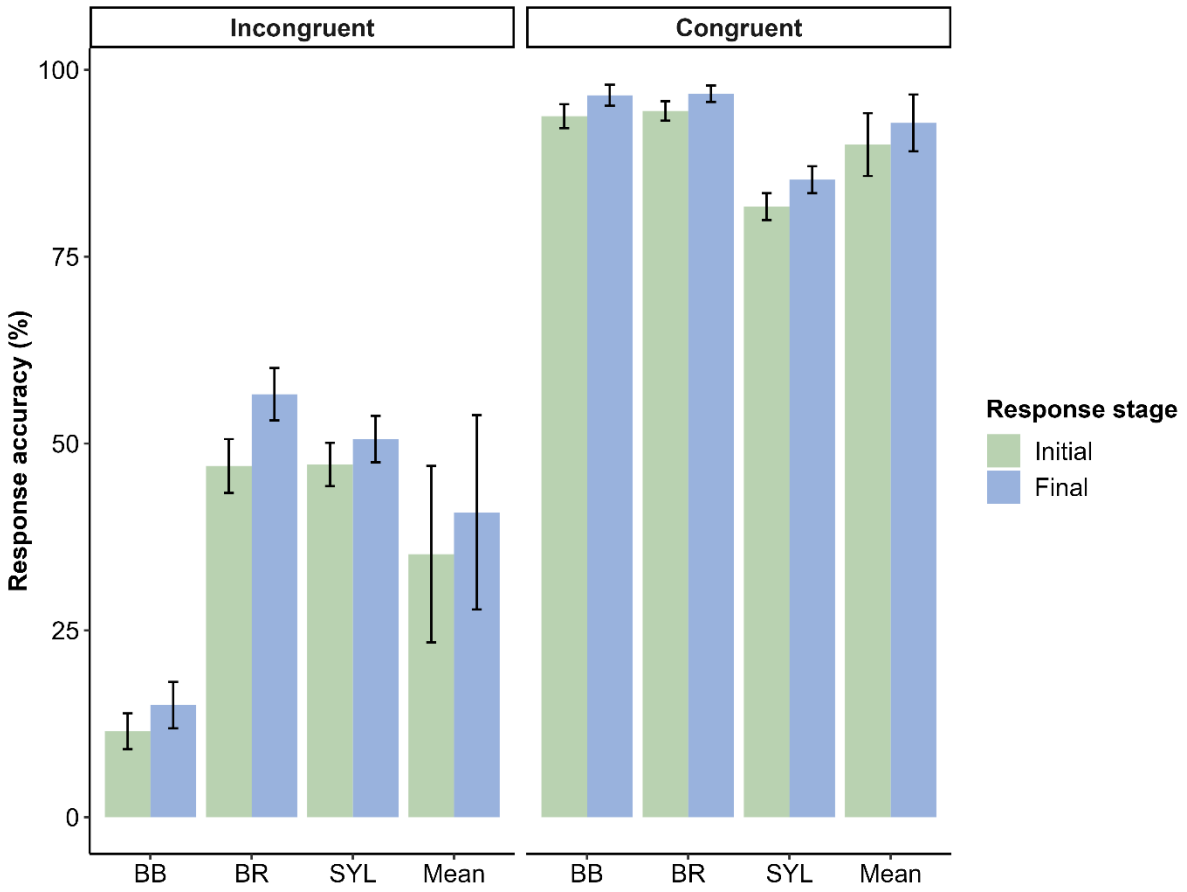

Figure 4. Response accuracy at incongruent and congruent trials of the Reasoning task in Study 3 for initial and final responses, separately for each problem type and for the mean across the three problem types. The error bars represent the Standard Error of the Mean. BB = Bat-and-ball; BR = Base-rates; SYL = Syllogisms; Mean = the mean across the four problem types.

Stability index. Like in the Stroop task, we calculated a stability index for the initial responses of the critical, incongruent trials. For each participant, we calculated on how many out of their initial responses in the incongruent trials they showed the same accuracy (i.e., "0" or "1"). The average stability index was 95.5% (*SD* = 11.1%) in the bat-and-ball task, 93.0% (*SD* = 14.6%) in the base-rate task and 81.3% (*SD* = 19.6%) in the syllogistic reasoning task. If initial responses were susceptible to systematic random guessing, we would observe more inconsistency in response patterns.

Direction of Change. To get a more precise picture of how participants changed their responses after deliberation we also conducted a direction of change analysis

(Bago & De Neys, 2017, 2019a). As Figure 5A shows, at the composite level, the majority of the critical, incongruent trials had a "00" pattern (52.2%) which confirms that reasoners are easily lured by the heuristic response when solving reasoning items. Critically, in the incongruent trials, the proportion of "11" responses (35.2%) is higher than that of the "01" responses (9.2%). The mean composite non-correction rate (i.e., proportion 11/11+01) reached 79.3%. Hence, as in the Stroop task, although there is some accuracy increase after deliberation, correct responses are, for the most part, already generated intuitively.

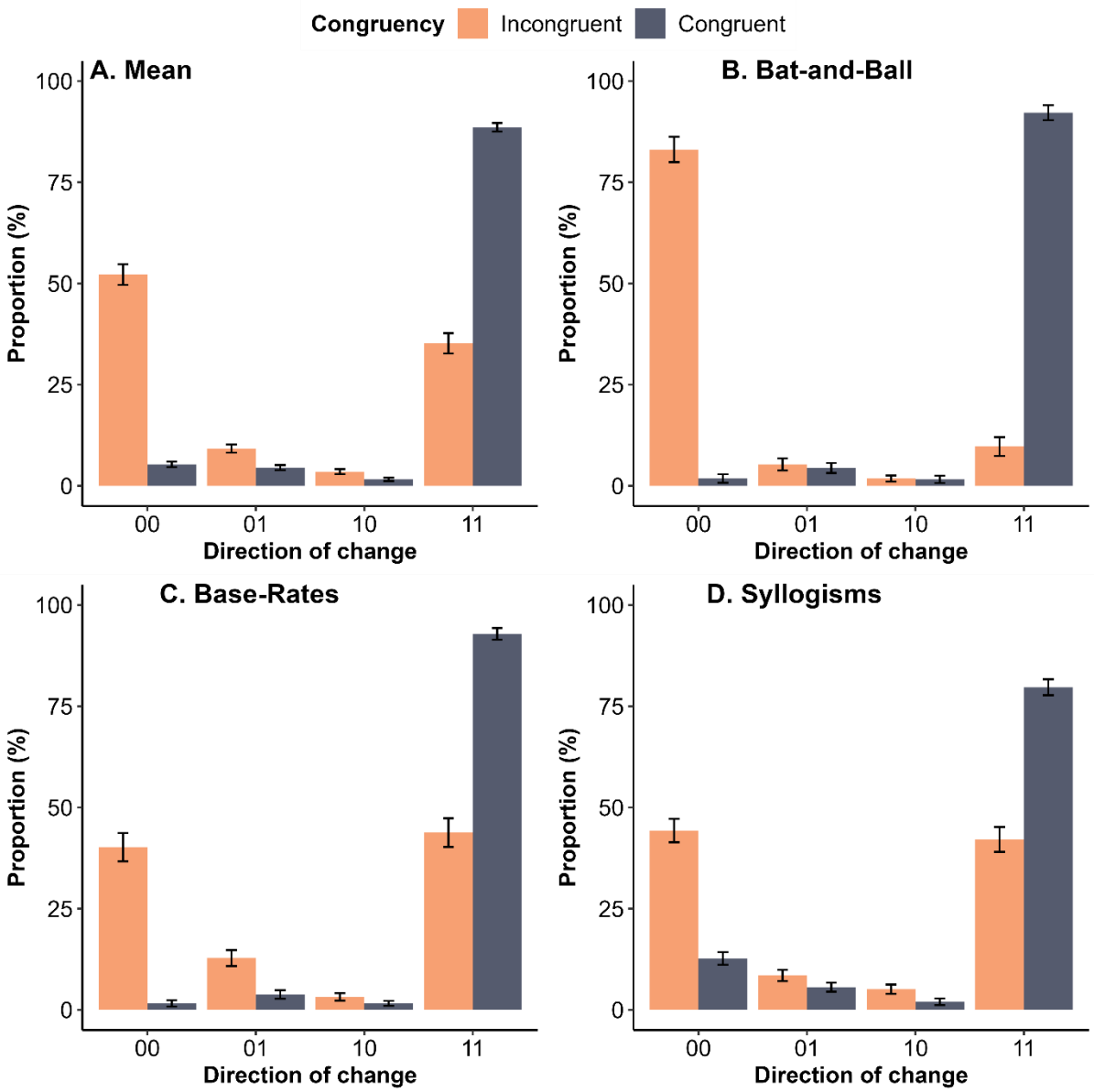

Figure 5. Direction of change in the Reasoning task of Study 3 separately for incongruent and congruent trials. **A)** Proportion of each direction of change category at the composite level. **B)** Proportion of each direction of change category at the Bat-and-Ball trials. **C)** Proportion of each direction of change

category at the Base-Rates trials. **D)** Proportion of each direction of change category at the Syllogistic reasoning trials. The error bars represent the Standard Error of the Mean. "00" = incorrect initial and incorrect final response; "01" = incorrect initial and correct final response; "10" = correct initial and incorrect final response; "11" = correct final and correct initial response.

Correlation between Stroop and Reasoning

We now turn to the main analysis of Study 3 in which we explore the relationship between participants' performance on the Stroop task and Reasoning task. For simplicity, we always use the Stroop task as the predictor of the Reasoning performance when interpreting the results.

For completeness, we also computed the split-half reliability of incongruent trials in both the Stroop task and the Reasoning task, separately for initial and final responses. The split-half reliability in the Stroop task was 0.94 for initial responses and 0.97 for final responses. In the bat-and-ball task, the split-half reliability was 0.78 for initial and 0.98 for final responses, in the base-rate task it was 0.91 for initial and 0.89 for final responses, and in the syllogistic reasoning task it was 0.66 for initial and 0.65 for final responses. For the composite reasoning measure, the split-half reliability was 0.82 for initial and 0.86 for final responses.

Accuracy. As a first step, we looked into whether the individual accuracies of participants in the Stroop task and the Reasoning task were related. As Table 1 shows, although there was a slight trend towards a positive association between the final Stroop performance and the initial and final Reasoning performance, all correlations were weak and typically did not reach significance.

Table 1

Pearson's product-moment correlation between the average accuracy of each individual on the Stroop task, and the accuracy of that individual on the Reasoning task. Correlations are reported both at the composite level and for each type of reasoning problem, separately for each Response stage (initial response; final response).

Reasoning Accuracy		Stroop Accuracy			
		Initial		Final	
		r	p	r	p
Initial	BB	−0.05	0.584	0.14	0.142
	BR	0.07	0.426	0.14	0.092
	SYL	0.08	0.343	0.09	0.277
	Composite	-0.06	0.569	0.14	0.145
Final	BB	−0.09	0.327	0.14	0.154
	BR	0.04	0.674	**0.18**	0.037
	SYL	0.07	0.432	0.10	0.266
	Composite	-0.12	0.234	0.16	0.109

Note. BB = Bat-and-ball; BR = Base-rates; SYL = Syllogisms. Significant correlations ($p < .05$) are in bold.

Direction of Change. In order to obtain a more detailed picture of the relationship between the two tasks, we focused on the direction of change patterns (i.e., "00", "01", "10", "11"). This allowed us to examine whether the tendency to change one's response after deliberation (or not) was related in the two tasks. More specifically, for each direction of change category we examined whether the proportion of trials of each category in the Stroop task was correlated with the proportion of trials of this same category in the Reasoning task. Table 2 shows the main results, but a full cross-tabulation table can also be found in Supplementary Material section E.

Table 2

Pearson's product-moment correlation between the proportion of each direction of change (i.e., "00", "01", "10", "00") of each individual in the Stroop task, and the proportion of each direction of change of that individual in the Reasoning task. Correlations are reported both at the composite level and separately for each type of reasoning problem.

	BB		BR		SYL		Composite	
	r	p	r	p	r	p	r	p
00	0.14	0.157	**0.19**	0.029	0.06	0.521	**0.17**	0.040
01	**0.20**	0.033	0.04	0.622	0.10	0.260	**0.17**	0.044
11	−0.02	0.817	0.04	0.654	0.14	0.092	0.12	0.149
10	−0.03	0.769	−0.05	0.598	**0.19**	0.022	0.12	0.161

Note. BB = Bat-and-ball; BR = Base-rates; SYL = Syllogisms; "00" = incorrect initial and incorrect final response; "01" = incorrect initial and correct final response; "10" = correct initial and incorrect final response; "11" = correct final and correct initial response. Significant correlations ($p < .05$) are in bold.

As Table 2 shows, at the composite level, there was overall evidence for a weak positive association between the direction of change patterns of each task. The more a participant showed a specific change pattern in the Stroop task, the more they tended to show this pattern in the Reasoning task. At the composite level, the correlation reached significance for the "00" and "01" pattern. Hence, the more a reasoner tended to provide entirely incorrect responses in the Stroop task (i.e., both their initial and final responses were incorrect), the more they tended to do so in the Reasoning task. Likewise, the more a reasoner tended to correct an initial incorrect response after deliberation in the Stoop task, the more they tended to show this change pattern in the Reasoning task. For the "11" and "10" patterns the composite correlations did not reach significance. At the individual task level, the trends were more diffuse (see Table 2).

To test whether the tendency to generate a correct final response through deliberation (i.e., a "01" pattern) showed a stronger link between the two tasks than the tendency to generate a correct response through intuitive processing (i.e., a "11" pattern) we also contrasted the composite "01" ($r = .18$) and "11" ($r = .12$) correlations directly. The difference between these correlations did not reach significance, $p = 0.61$.

In sum, the correlational analyses indicated that there is evidence for a weak association between Stroop and Reasoning performance. However, there was no clear indication that initial Stroop performance would be a better predictor than the final Stroop performance or vice versa.

General Discussion

In the present paper we were inspired by recent two-response findings that show evidence for correct intuitive responding in reasoning problems (e.g., Bago & De Neys, 2017, 2019a) and tested whether they could be generalized to low level cognitive control tasks. For this purpose, we examined whether people who respond accurately to the classic Stroop and Flanker tasks, could also do so when their deliberate control was minimized. We used the two-response paradigm to test the accuracy of both initial responses (given under limited deliberation conditions) and final responses. As a second step, we examined how the two-response Stroop performance was related to the performance on classic reasoning problems.

Concerning our first research question, both our studies showed that in most cases where people provided a correct final response to the Stroop and Flanker tasks, they had already responded correctly in the initial stage. In other words, deliberate control was not always necessary for correct responding in these tasks, which suggests that the two-response reasoning findings can generalize to lower level cognitive control tasks. In general, this fits the claim that popular "fast-and-slow" dual process models need to upgrade their view of the fast and intuitive System 1 (De Neys & Pennycook, 2019). Across a wide range of fields, responses that are traditionally believed to necessitate slow controlled deliberation, often seem to fall within the realm of more intuitive processing (De Neys, 2022).

As mentioned in the Introduction, the idea that control does not always require deliberation and can be exerted automatically, is in line with some existing evidence from the cognitive control field (Abrahamse et al., 2016; Chiu & Aron, 2014; Desender et al., 2013; Jiang et al., 2015, 2018; Linzarini et al., 2017). This evidence shows that participants perform better in an incongruent Stroop trial when it is preceded by an unconsciously presented incongruent trial (compared to an unconsciously presented congruent trial). In this case, participants recruit automatic control during the first, unconscious trial, which is boosting their

performance in the trial that follows. These findings suggest that cognitive control, as we traditionally conceive it, might result from related automatic control processes, which fits with the findings of the present paper. However, we would like to clarify that automatic ("intuitive") control is typically understood as unconscious control (e.g., implying subliminal presentation of trials). Although in the present paper our initial responses are extremely challenging for participants, they clearly fall outside the unconscious processing range. Therefore, we obviously do not argue that our studies provide direct evidence for unconscious control, but that they point in the same direction as the aforementioned evidence of automatic control: control can be exerted faster and more effortlessly than traditionally assumed.

With our second research question, we attempted to explore how performance on cognitive control tasks, like the Stroop, and reasoning tasks are related. More specifically, we wanted to test whether an individual's initial Stroop performance would be a better predictor of their reasoning accuracy than their final Stroop performance. This question was inspired by Raoelison et al.'s (2020) research which showed that cognitive capacity primarily predicts intuitive, rather than deliberate reasoning performance. Under this "smart intuitor" view, smarter people (i.e., people with high cognitive capacity) are better at providing correct responses intuitively, rather than deliberately correcting their erroneous intuitions. Our rationale was that both reasoning tasks and cognitive control tasks might tap into the same automatic control processes. This is why we expected that people who provide correct Stroop responses when their cognitive control is restricted, will also be able to provide correct intuitive responses to reasoning problems. However, we only found a weak association between people's response patterns in the Stroop task and those in the Reasoning tasks. Critically, there was no clear indication in our data to suggest that the initial, "intuitive" Stroop performance could better predict reasoning accuracy compared to the final, deliberate Stroop performance. Below we discuss two main potential reasons for the lack of association between these tasks.

First, for practical reasons, in Study 3 of the present paper we focused on one cognitive control task, namely the Stroop. To our knowledge, our paper is the first to test the corrective assumption in the Stroop task and investigate how it relates to reasoning accuracy. Thereby, it provides critical new insight into the generalization of the two-response findings. However, it is possible that the Stroop

task alone was not an optimal psychometric predictor of cognitive control. While it is not uncommon to use a single task to tap cognitive control, a discussion exists in the literature concerning the task impurity problem (Miyake et al., 2000). More specifically, since no single cognitive control task is a pure measure of cognitive control, there are concerns that the observed results in studies that use only one type of predictor task are tied to the requirements of the task itself (e.g., specific demands and properties) rather than to cognitive control abilities (Gärtner & Strobel, 2021; Miyake et al., 2000). This might also explain why correlations of performance between these different cognitive control tasks are often weak or absent (e.g. Enge et al., 2014; Singh et al., 2018). Relatedly, deliberate control might not represent a single common process, but instead be separated into subtypes which would all be measured by different types of control tasks (e.g., Morra et al., 2018). For example, the Stroop task and the Flanker task that we used in Studies 1 and 2 are sometimes thought to tap into different aspects of cognitive control and measure different inhibition-related functions (e.g., Friedman & Miyake, 2004; Rey-Mermet et al., 2018; but see also Nigg, 2000). While the Stroop task measures prepotent response inhibition (i**.**e., the ability to deliberately supress dominant or automatic responses), the Flanker task measures resistance to distractor interference (i.e., the ability to maintain focused attention and resist interference from distractors that are irrelevant to the task at hand). Although these two inhibitory functions are closely related (Friedman & Miyake, 2004), they are also distinguishable (Kane et al., 2016). One solution to combat these issues in future research would be to use a pool of common cognitive control tasks in order to create a cognitive control composite index. If we assume that our Stroop task is a weak indicator of individual cognitive control, this could explain why it is not strongly or differentially correlated with intuitive and deliberate reasoning performance.

Second, the weak association between the performance on the two tasks could also be due to their different nature. In the present paper we attempt to draw a link between the Stroop task and heuristics-and-biases reasoning tasks, but it might be that these are not necessarily directly comparable. That is, although the same pattern of results (i.e., correct intuitive responding) is present in both tasks, the specific mechanism that gives rise to this pattern might differ between them.

One of the potential differences between the Stroop task and heuristics-and-biases reasoning tasks is that in the reasoning tasks—for those who manage to respond correctly—the correct response might be more dominant because it is based on a rule that has been practiced to automaticity. That is, it has been hypothesised that the origin of people's logical intuitions in reasoning tasks lies in a practice or learning process (De Neys, 2012; De Neys & Pennycook, 2019; Raoelison et al., 2021; Stanovich, 2018). Reasoners have typically already been exposed to the core logical principles and often even practiced them at length in the school curriculum (Raoelison et al., 2020). This repeated exposure would have allowed good reasoners to automatize their application. In others words, for sound reasoners the critical "mindware" (i.e., the knowledge of elementary logical principles) has been fully instantiated (i.e., automatized, Stanovich, 2018) such that its activation strength will outcompete the conflicting heuristic intuition. In the Stroop task, however, the correct response is based on a new (in se trivial) instruction that people have not previously practiced or been exposed to. That is, participants are faced with an automatic, habitual response (i.e., reading the word) which they are told to consider as incorrect, and a competing response (i.e., naming the words' colour) which they should consider as correct according to the task's instructions. Consequently, when responding to the Stroop task, participants always need to recruit cognitive control (automatically or deliberately) in order to inhibit their habitual response and answer correctly. However, because the more instantiated correct logical intuition will already dominate the competing heuristic intuition for good reasoners, correct intuitive responding may no longer require (or require less) control per se in a reasoning task. In sum, contrary to the Stroop task, correct responding to reasoning problems might not always demand engagement of (automatic) cognitive control, as the two potential responses are not always "competing" with each other. This could explain why, in our findings, individual performance at the Stroop and the Reasoning task are not strongly related.

To sum up, we speculate that when solving reasoning problems one can be a sound intuitor either because one's "logical" intuitions are very strong, or because one's competing logical and heuristic intuitions are similar in strength and cognitive control is automatically exerted (i.e., the heuristic response is automatically supressed). Interestingly, it has been argued in the reasoning field that the level of similarity between the alleged intuitions is reflected in response

confidence: the more similar the competing intuitions are, the more conflicted and less certain one would feel about their decision (Bago & De Neys, 2020; De Neys, 2022; De Neys & Pennycook, 2019). This speculatively points to a possible test of this hypothesis. By contrasting initial correct responders who express the most and the least response confidence (i.e., who can be hypothesized to have less and more dominant logical intuitions, respectively), we can test whether the Reasoning-Stroop performance of the less confident (more conflicted) people is more strongly related. Presumably, participants with lower confidence have less dominant logical intuitions, thus they need automatic control to generate correct intuitive responses to reasoning problems (just like in the Stroop task). It is, therefore, in these participants that we may expect to find a clearer relationship between the initial Stroop and Reasoning performance.

Accidentally, we did (for different purposes, see Supplementary Material section F) record response confidence for the Reasoning task in Study 3. We used these to split our group of reasoners in two halves based on the median "11" (i.e., correct final responses that were already generated intuitively) Reasoning task confidence: low "11" and high "11" confidence. We then performed a post-hoc correlational analysis separately for each group. As it can be seen in the Supplementary Material section G, for the people that had a low "11" confidence (high conflict), their "11" Stroop performance clearly correlated with their "11" Reasoning composite performance (r = 0.34, p = .01). However, for the people that had a high "11" confidence (low conflict), their "11" Stroop performance did not correlate with their Reasoning performance (r = 0.07, p = .58). The difference between these correlations was not significant, p = .136. Although this post-hoc analysis should be interpreted with caution, it does lend some credence to the idea that the Stroop and the Reasoning tasks might be only related in the cases where (automatic) cognitive control is required for sound reasoning.

Relatedly, it is worth considering that deliberation per se might play different roles in reasoning and lower level cognitive control tasks. For example, in the reasoning field it has been shown that even when people intuitively arrive at the correct response, they subsequently engage in deliberation to justify that response (Bago & De Neys, 2019a; De Neys & Pennycook, 2019). In other words, although sound reasoners typically generate the correct response intuitively, they often struggle to explain how they arrived at their response (Bago & De Neys, 2019a). However, after the final response stage, in which they are allowed to

deliberate, they readily provide such justifications (e.g., Bago & De Neys, 2019a). Hence, it has been argued that deliberation during reasoning might be primarily required to justify and communicate one's response. Arguably, such justification is less central for lower level cognitive control tasks. Although this hypothesis is speculative, it underscores that deliberation might play different or additional roles in these two domains.

A possible general critique against the present paper is that we can never be sure that all possible deliberation was prevented in the initial, intuitive response stage. For example, it could be that the paradigm still allowed for some minimal deliberation during the initial stage, which could explain the correct responses at that stage. However, note that to minimize the possibility that reasoners engage in deliberate control in the initial stage, we combined three validated procedures which have been shown to reduce deliberation: instructions, time pressure, and concurrent load. One could always argue that a more demanding deadline or load task could have been used. Nevertheless, especially in the case of our low-level control tasks, it is important to consider the substantial number of missed trials both in Study 1 (41.9%), Study 2 (44.3%) and Study 3 (36.5%). These percentages suggest that the tasks were extremely challenging, and that introducing additional load or time pressure would lead to practical and statistical issues (i.e., selection effects due to a large portion of discarded trials, e.g., Bouwmeester et al., 2017).

Nevertheless, the underlying point remains that regardless of how challenging the test conditions are in the initial stage, we can never be entirely certain that participants did not deliberate. The issue here is that dual process theories are underspecified (De Neys, 2021). While these theories suggest that deliberation is slower and more demanding than intuition, they do not provide a definite criterion or threshold for distinguishing between intuitive and deliberate processes (Bago & De Neys, 2019a; De Neys, 2022). So, as long as there are correct initial responses, one can always argue that they would disappear "with just a little bit more load/time pressure". At this point, the corrective assumption becomes unfalsifiable, since any evidence for correct intuiting can always be explained by arguing that the methodological design allowed for deliberation. At the same time, this indicates that the label correct "intuiting" needs to be interpreted within practical boundaries and some caution. Although our results

question the corrective role of deliberation in low-level control tasks, they should always be interpreted with this limitation in mind.

To conclude, although the link between cognitive control and reasoning performance might be complex, our key finding is that successful cognitive control does not necessarily require slow and effortful deliberation. This lends credence to the idea that cognitive control can be exerted automatically. These results point to an interesting generalization of the two-response findings to low-level cognitive control tasks. This further underscores the claim that the popular "fast-and-slow" dual process models of human cognition need to revise and upgrade their view of the fast and intuitive System 1 (De Neys & Pennycook, 2019). We also hope that the study can serve as a proof-of-principle and lead to a deeper integration of the related—but hitherto somewhat isolated—cognitive control and reasoning fields. We believe that such an integration will be indispensable to pinpoint the mechanisms underlying intuitive-automatic responding in higher and lower level cognition.

References

Abrahamse, E., Braem, S., Notebaert, W., & Verguts, T. (2016). Grounding cognitive control in associative learning. *Psychological Bulletin*, *142*(7), 693–728. https://doi.org/10.1037/bul0000047

Abreu-Mendoza, R. A., Coulanges, L., Ali, K., Powell, A. B., & Rosenberg-Lee, M. (2020). Children's discrete proportional reasoning is related to inhibitory control and enhanced by priming continuous representations. *Journal of Experimental Child Psychology*, *199*, 104931. https://doi.org/10.1016/j.jecp.2020.104931

Abutalebi, J., Della Rosa, P. A., Green, D. W., Hernandez, M., Scifo, P., Keim, R., ... & Costa, A. (2012). Bilingualism tunes the anterior cingulate cortex for conflict monitoring. *Cerebral cortex*, *22*(9), 2076-2086. https://doi.org/10.1093/cercor/bhr287

Aïte, A., Cassotti, M., Linzarini, A., Osmont, A., Houdé, O., & Borst, G. (2016). Adolescents' inhibitory control: Keep it cool or lose control. *Developmental Science*, *21*(1), e12491. https://doi.org/10.1111/desc.12491

Algom, D., & Chajut, E. (2019). Reclaiming the Stroop Effect Back From Control to Input-Driven Attention and Perception. *Frontiers in Psychology*, *10*, 1683. https://doi.org/10.3389/fpsyg.2019.01683

Bago, B., Bonnefon, J.-F., & De Neys, W. (2021). Intuition rather than deliberation determines selfish and prosocial choices. *Journal of Experimental Psychology:*

General, 150(6), 1081. https://doi.org/10.1037/xge0000968

Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition*, *158*, 90–109. https://doi.org/10.1016/j.cognition.2016.10.014

Bago, B., & De Neys, W. (2019a). The Smart System 1: Evidence for the intuitive nature of correct responding on the bat-and-ball problem. *Thinking & Reasoning*, *25*(3), 257–299. https://doi.org/10.1080/13546783.2018.1507949

Bago, B., & De Neys, W. (2019b). The intuitive greater good: Testing the corrective dual process model of moral cognition. *Journal of Experimental Psychology: General, 148*(10), 1782. https://doi.org/10.1037/xge0000533

Bago, B., & De Neys, W. (2020). Advancing the specification of dual process models of higher cognition: A critical test of the hybrid model view. *Thinking & Reasoning*, *26*(1), 1–30. https://doi.org/10.1080/13546783.2018.1552194

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624–652. https://doi.org/10.1037/0033-295X.108.3.624

Bouwmeester, S., Verkoeijen, P. P., Aczel, B., Barbosa, F., Bègue, L., Brañas-Garza, P., ... & Wollbrant, C. E. (2017). Registered replication report: Rand, greene, and nowak (2012). *Perspectives on Psychological Science, 12*(3), 527–542. https://doi.org/10.1177/1745691617693624

Braem, S., & Egner, T. (2018). Getting a grip on cognitive flexibility. *Current Directions in Psychological Science*, *27*(6), 470–476. https://doi.org/10.1177/0963721418787475

Burič, R., & Šrol, J. (2020). Individual differences in logical intuitions on reasoning problems presented under two-response paradigm. *Journal of Cognitive Psychology, 32*(4), 460-477. https://doi.org/10.1080/20445911.2020.1766472

Chiu, Y.-C., & Aron, A. R. (2014). Unconsciously Triggered Response Inhibition Requires An Executive Setting. *Journal of Experimental Psychology. General*, *143*(1), 56–61. https://doi.org/10.1037/a0031497

de Fockert, J. W., Rees, G., Frith, C. D., & Lavie, N. (2001). The role of working memory in visual selective attention. *Science, 291*(5509), 1803–1806. https://doi.org/10.1126/science.1056496

De Neys, W. (2006). Dual processing in reasoning: Two systems but one reasoner. *Psychological science, 17*(5), 428–433. https://doi.org/10.1111/j.1467-9280.2006.01723.x

De Neys, W. (2012). Bias and conflict: A case for logical intuitions. Pe*rspectives on Psychological Science, 7*(1), 28–38. https://doi.org/10.1177/1745691611429354

De Neys, W. (2017). Bias, Conflict, and Fast Logic: Towards a hybrid dual process

future? In *Dual Process Theory 2.0*. Routledge.

De Neys, W. (2021). On dual-and single-process models of thinking. *Perspectives on psychological science, 16*(6), 1412-1427. https://doi.org/10.1177/1745691620964172

De Neys, W. (2022). Advancing theorizing about fast-and-slow thinking. *Behavioral and Brain Sciences*, 1-68. https://doi.org/10.1017/S0140525X2200142X

De Neys, W., Cromheeke, S., & Osman, M. (2011). Biased but in doubt: Conflict and decision confidence. *PloS one, 6*(1), e15954. https://doi.org/10.1371/journal.pone.0015954

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition*, *106*(3), 1248–1299. https://doi.org/10.1016/j.cognition.2007.06.002

De Neys, W., & Pennycook, G. (2019). Logic, Fast and Slow: Advances in Dual-Process Theorizing. *Current Directions in Psychological Science*, *28*(5), 503–509. https://doi.org/10.1177/0963721419855658

Desender, K., Lierde, E. V., & Bussche, E. V. den. (2013). Comparing Conscious and Unconscious Conflict Adaptation. *PLOS ONE*, *8*(2), e55976. https://doi.org/10.1371/journal.pone.0055976

Diamond, A. (2013). Executive functions. *Annual review of psychology, 64*, 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Enge, S., Behnke, A., Fleischhauer, M., Küttler, L., Kliegel, M., & Strobel, A. (2014). No evidence for true training and transfer effects after inhibitory control training in young healthy adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(4), 987. https://doi.org/10.1037/a0036165

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & psychophysics, 16*(1), 143-149. https://doi.org/10.3758/BF03203267

Eriksen, C. W., & Hoffman, J. E. (1973). The extent of processing of noise elements during selective encoding from visual displays. *Perception & Psychophysics, 14*(1), 155-160. https://doi.org/10.3758/BF03198630

Evans, J. St. B. T. (2002). Logic and human reasoning: An assessment of the deduction paradigm. *Psychological Bulletin*, *128*(6), 978–996. https://doi.org/10.1037/0033-2909.128.6.978

Evans, J. St. B. T. (2008). Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition. *Annual Review of Psychology*, *59*(1), 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Evans J. St. B. T. (2019). Reflections on reflection: the nature and function of type 2 processes in dual-process theories of reasoning. *Thinking & Reasoning, 25*(4),

383–415. https://doi.org/10.1080/13546783.2019.1623071

Evans, J. St. B. T., & Over, D. E. (1996). Rationality in the selection task: Epistemic utility versus uncertainty reduction. *Psychological Review*, *103*(2), 356–363. https://doi.org/10.1037/0033-295X.103.2.356

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science, 8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Fan, J., McCandliss, B. D., Fossella, J., Flombaum, J. I., & Posner, M. I. (2005). The activation of attentional networks. *Neuroimage*, *26*(2), 471-479. https://doi.org/10.1016/j.neuroimage.2005.02.004

Franssens, S., & De Neys, W. (2009). The effortless nature of conflict detection during thinking. *Thinking & Reasoning*, *15*(2), 105–128. https://doi.org/10.1080/13546780802711185

Friedman, N. P., & Miyake, A. (2004). The relations among inhibition and interference control functions: a latent-variable analysis. *Journal of experimental psychology: General, 133*(1), 101. https://doi.org/10.1037/0096-3445.133.1.101

Gao, Q., Chen, Z., & Russell, P. (2007). *Working Memory Load and the Stroop Interference Effect*. *36*(3), 8. http://hdl.handle.net/10092/2792

Gärtner, A., & Strobel, A. (2021). Individual Differences in Inhibitory Control: A latent Variable Analysis. *Journal of Cognition*, *4*(1), 17. https://doi.org/10.5334/joc.150

Handley, S. J., Capon, A., Beveridge, M., Dennis, I., & Evans, J. S. B. (2004). : Working memory, inhibitory control and the development of children's reasoning. *Thinking & Reasoning*, *10*(2), 175–195. https://doi.org/10.1080/13546780442000051

Handley, S. J., Newstead, S. E., & Trippas, D. (2011). Logic, beliefs, and instruction: a test of the default interventionist account of belief bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(1), 28. https://doi.org/10.1037/a0021098

Hübner, R., Steinhauser, M., & Lehle, C. (2010). A dual-stage two-phase model of selective attention. *Psychological review, 117*(3), 759. https://doi.org/10.1037/a0019471

Jiang, J., Correa, C. M., Geerts, J., & van Gaal, S. (2018). The relationship between conflict awareness and behavioral and oscillatory signatures of immediate and delayed cognitive control. *NeuroImage*, *177*, 11–19. https://doi.org/10.1016/j.neuroimage.2018.05.007

Jiang, J., Zhang, Q., & Van Gaal, S. (2015). EEG neural oscillatory dynamics reveal semantic and response conflict at difference levels of conflict awareness. *Scientific Reports*, *5*(1), 12008. https://doi.org/10.1038/srep12008

Johnson, E. D., Tubau, E., & De Neys, W. (2016). The Doubting System 1: Evidence for

automatic substitution sensitivity. *Acta psychologica*, *164*, 56-64. https://doi.org/10.1016/j.actpsy.2015.12.008

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Strauss, Giroux.

Kahneman, D., & Chajczyk, D. (1983). Tests of the automaticity of reading: dilution of Stroop effects by color-irrelevant stimuli. *Journal of Experimental Psychology: Human perception and performance*, *9*(4), 497. https://doi.org/10.1037/0096-1523.9.4.497

Kane, M. J., Meier, M. E., Smeekens, B. A., Gross, G. M., Chun, C. A., Silvia, P. J., & Kwapil, T. R. (2016). Individual differences in the executive control of attention, memory, and thought, and their associations with schizotypy. *Journal of Experimental Psychology: General, 145*(8), 1017. https://doi.org/10.1037/xge0000184

Keele, S. W. (1972). Attention demands of memory retrieval. *Journal of Experimental Psychology*, *93*(2), 245. https://doi.org/10.1037/h0032460

Kessler, J., Kivimaki, H., & Niederle, M. (2017). Thinking fast and slow: generosity over time. *Preprint at https://stanford.edu/~niederle/KKN_ThinkingFastandSlow.pdf*

Lavie, N. (2005). Distracted and confused?: Selective attention under load. *Trends in Cognitive Sciences*, *9*(2), 75–82. https://doi.org/10.1016/j.tics.2004.12.004

Lavie, N., & De Fockert, J. (2005). The role of working memory in attentional capture. *Psychonomic Bulletin & Review*, *12*(4), 669–674. https://doi.org/10.3758/BF03196756

Lavie, N., Hirst, A., de Fockert, J. W., & Viding, E. (2004). Load Theory of Selective Attention and Cognitive Control. *Journal of Experimental Psychology: General*, *133*(3), 339–354. https://doi.org/10.1037/0096-3445.133.3.339

Linzarini, A., Houdé, O., & Borst, G. (2017). Cognitive control outside of conscious awareness. *Consciousness and Cognition*, *53*, 185–193. https://doi.org/10.1016/j.concog.2017.06.014

Mata, A. (2020). Conflict detection and social perception: Bringing meta-reasoning and social cognition together. *Thinking & Reasoning, 26*(1), 140–149. https://doi.org/10.1080/13546783.2019.1611664

Mata, A., & Ferreira, M. B. (2018). Response: Commentary: Seeing the conflict: an attentional account of reasoning errors. *Frontiers in psychology, 9,* 24. https://doi.org/10.3389/fpsyg.2018.00024

Mata, A., Schubert, A. L., & Ferreira, M. B. (2014). The role of language comprehension in reasoning: How "good-enough" representations induce biases. *Cognition, 133*(2), 457–463. https://doi.org/10.1016/j.cognition.2014.07.011

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to

complex "frontal lobe" tasks: A latent variable analysis. *Cognitive psychology, 41*(1), 49–100. https://doi.org/10.1006/cogp.1999.0734

Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. *Journal of Experimental Psychology: General*, *130*(4), 621–640. https://doi.org/10.1037/0096-3445.130.4.621

Morra, S., Panesi, S., Traverso, L., & Usai, M. C. (2018). Which tasks measure what? Reflections on executive function development and a commentary on Podjarny, Kamawar, and Andrews (2017). *Journal of Experimental Child Psychology*, *167*, 246–258. https://doi.org/10.1016/j.jecp.2017.11.004

Newman, I. R., Gibb, M., & Thompson, V. A. (2017). Rule-based reasoning is fast and belief-based reasoning can be slow: Challenging current explanations of belief-bias and base-rate neglect. (7), 1154. *Journal of experimental psychology: learning, memory, and cognition, 43.* https://doi.org/10.1037/xlm0000372

Nigg, J. T. (2000). On inhibition/disinhibition in developmental psychopathology: views from cognitive and personality psychology and a working inhibition taxonomy. *Psychological bulletin, 126*(2), 220. https://doi.org/10.1037/0033-2909.126.2.220

Penner, I. K., Kobel, M., Stöcklin, M., Weber, P., Opwis, K., & Calabrese, P. (2012). The Stroop task: comparison between the original paradigm and computerized versions in children and adults*. The Clinical Neuropsychologist, 26*(7), 1142–1153. https://doi.org/10.1080/13854046.2012.713513

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2014). Cognitive
style and religiosity: The role of conflict detection. *Memory & Cognition*, *42*(1), 1–10.
https://doi.org/10.3758/s13421-013-0340-7

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.
https://doi.org/10.1016/j.cogpsych.2015.05.001

Raoelison, M., Boissin, E., Borst, G., & De Neys, W. (2021). From slow to fast logic: the development of logical intuitions. *Thinking & Reasoning, 27*(4), 599–622. https://doi.org/10.1080/13546783.2021.1885488

Raoelison, M., & De Neys, W. (2019). Do we de-bias ourselves?: The impact of repeated presentation on the bat-and-ball problem. *Judgment and Decision making, 14*(2), 170.

Raoelison, M., Thompson, V. A., & De Neys, W. (2020). The smart intuitor: Cognitive

capacity predicts intuitive rather than deliberate thinking. *Cognition, 204*, 104381. https://doi.org/10.1016/j.cognition.2020.104381

Rey-Mermet, A., Gade, M., & Oberauer, K. (2018). Should we stop thinking about inhibition? Searching for individual and age differences in inhibition ability. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*(4), 501. https://doi.org/10.1037/xlm0000450

Ridderinkhof, K. R., Wylie, S. A., van den Wildenberg, W. P., Bashore, T. R., & van der Molen, M. W. (2021). The arrow of time: Advancing insights into action control from the arrow version of the Eriksen flanker task. *Attention, Perception, & Psychophysics, 83*, 700-721. https://doi.org/10.3758/s13414-020-02167-z

Servant, M., & Logan, G. D. (2019). Dynamics of attentional focusing in the Eriksen flanker task. *Attention, Perception, & Psychophysics, 81*, 2710-2721. https://doi.org/10.3758/s13414-019-01796-3

Singh, K., Ecker, U., Gignac, G., Brydges, C., & Rey-Mermet, A. (2018). *Interference Control in Working Memory*. PsyArXiv. https://doi.org/10.31234/osf.io/fjrnq

Stanovich, K. E. (2018). Miserliness in human cognition: The interaction of detection, override and mindware. *Thinking & Reasoning, 24*(4), 423–444. https://doi.org/10.1080/13546783.2018.1459314

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate?. *Behavioral and brain sciences, 23*(5), 645–665. https://doi.org/10.1017/S0140525X00003435

Stirling, N. (1979). Stroop interference: An input and an output phenomenon. *The Quarterly Journal of Experimental Psychology, 31*(1), 121–132. https://doi.org/10.1080/14640747908400712

Stoffels, E. J., & Van der Molen, M. W. (1988). Effects of visual and auditory noise on visual choice reaction time in a continuous-flow paradigm. *Perception & Psychophysics*, *44*, 7-14. https://doi.org/10.3758/bf03207468

Strauss, G. P., Allen, D. N., Jorgensen, M. L., & Cramer, S. L. (2005). Test-retest reliability of standard and emotional stroop tasks: an investigation of color-word and picture-word versions. *Assessment, 12*(3), 330–337. https://doi.org/10.1177/1073191105276375

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of experimental psychology, 18*(6), 643. https://doi.org/10.1037/h0054651

Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning, 20*(2), 215–244. https://doi.org/10.1080/13546783.2013.869763

Thompson, V. A., Turner, J. A. P., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive psychology, 63*(3), 107–140.

https://doi.org/10.1016/j.cogpsych.2011.06.001

Vega, S., Mata, A., Ferreira, M. B., & Vaz, A. R. (2021). Metacognition in moral decisions: judgment extremity and feeling of rightness in moral intuitions. *Thinking & Reasoning, 27*(1), 124–141. https://doi.org/10.1080/13546783.2020.1741448

Wright, B. C., & Wanley, A. (2003). Adults' versus children's performance on the Stroop task: Interference and facilitation. *British Journal of Psychology, 94*(4), 475–485. https://doi.org/10.1348/000712603322503042

Chapter 4

Conflict detection and temporal stability in reasoning biases

Voudouri, A., Bialek, M., Domurat, A., Kowal, M., & De Neys, W. (2022). Conflict detection predicts the temporal stability of intuitive and deliberate reasoning. *Thinking & Reasoning*, 1-29. https://doi.org/10.1080/13546783.2022.2077439

Supplementary material for this chapter can be found in Supplementary material for Chapter 4.

Abstract

Although the susceptibility to reasoning biases is often assumed to be a stable trait, the temporal stability of people's performance on popular heuristics-and-biases tasks has been rarely directly tested. The present study addressed this issue and examined a potential determinant for answer change. Participants solved the same set of "bias" tasks twice in two test sessions, two weeks apart. We used the two-response paradigm to test the stability of both initial (intuitive) and final (deliberate) responses. We hypothesized that participants who showed higher conflict detection in their initial intuitive responses at session 1 (as indexed by a relative confidence decrease compared to control problems), would be less stable in their responses between session 1 and 2. Results showed that performance on the reasoning tasks was highly, but not entirely, stable two weeks later. Notably, conflict detection in session 1 was significantly more pronounced in those cases that participants changed their answer between session 1 and 2 than when they did not change their answer between sessions. We discuss practical and theoretical implications.

Introduction

Although reasoning has been characterized as the essence of our being, it is often prone to cognitive biases. Decades of research in the reasoning and decision making fields have shown that when faced with simple reasoning tasks, people tend to overlook their underlying logical principles and, as a result, provide incorrect answers (Kahneman, 2011). Consider the following problem:

> *Imagine you are running a race. If you pass the person in second place what place are you in?*

The answer that often pops into mind is "first place". However, if one takes the time to further reflect on the problem, it is clear that the correct answer is in fact "second place". Despite the simplicity of the solution, mistakes in reasoning tasks like the above are very frequent. This is because people often base their answer on mental shortcuts (e.g., "after second comes first" in the above example), instead of providing an answer that agrees with logical norms (e.g., "if you pass the second runner, there is still a person ahead of you"). A prevalent explanation as to why these errors of judgement happen, has been proposed by dual-process theories. These theories view reasoning as an interaction between two systems, System 1 and System 2, which approximately correspond to intuitive and deliberate thinking (e.g., Epstein, 1994; Evans, 2008; Evans & Stanovich, 2013; Kahneman, 2011; Sloman, 1996). The main difference between these systems is that while System 1 is autonomous and does not make use of cognitive resources, System 2 requires cognitive resources to operate. System 1 can be helpful in many cases (e.g., when a decision has to be taken quickly), but it also often cues "heuristic" answers, responses that are based on rules of thumb, stored associations, and stereotypes. Classic dual process theories support that when a problem cues a "heuristic" answer that conflicts with logical considerations, reasoners need to engage in effortful thinking and further contemplate the problem in order to override their "intuitive", erroneous answer and provide a normative response[1] (Evans & Stanovich, 2013; Kahneman, 2011). However, in most cases, in order to minimize effortful thinking reasoners stick to their

[1] When we refer to the "logical", "normative", or "correct" response we are referring to the response that has traditionally been considered to be correct according to standard logic and probability theory.

"heuristic" answer and respond incorrectly (Evans & Over, 1996; Kahneman, 2011).

Heuristic biases have been widely researched in the literature and have been predicted using a range of cognitive tasks (Białek et al., 2020; Šrol & De Neys, 2021; Stupple et al., 2013; Toplak et al., 2014). Nevertheless, it is not completely clear whether the performance of reasoners on bias tasks is stable over time. Although bias susceptibility is generally assumed to be a stable individual trait, in the sense that biased reasoners are thought to remain biased from one moment in time to another, reasoners' response consistency has been rarely directly tested (e.g., Białek & Pennycook, 2018; Meyer et al., 2018; Stango & Zinman, 2020). In the present paper, we will investigate this consistency and discuss a potential determinant for answer change.

The determinant we will focus on is reasoners' detection of conflict between competing responses (e.g., De Neys & Glumicic, 2008; De Neys, 2012; Šrol & De Neys, 2021). Over the last decade, numerous studies have indicated that when people solve classic "bias" tasks in which they are faced with a cued heuristic response that conflicts with logical principles, they often show some sensitivity to this conflict (e.g., Bago & De Neys, 2017; De Neys, 2014; De Neys et al., 2013; Gangemi et al., 2015; Mata, 2020; Pennycook et al., 2015; Stupple et al., 2013; but see also Ferreira et al., 2016; Mata et al., 2017; Pennycook et al., 2012). For example, reasoners typically show lower confidence when answering a classic "bias" task than when solving a control version in which the cued heuristic does not conflict with logical principles (e.g., a no-conflict version of the introductory race problem might read "Imagine you are running a race. If you pass the person in first place, what place are you in?"). This suggest that people detect, to some extent, that there are conflicting responses at play.

In this study, we wanted to explore if conflict detection is related to how often people change their answers on classic bias tasks from one point in time to another. The general idea was that the more conflicted reasoners feel about an answer, the more likely they might be to change this answer at a future time. Evidence for this comes from the two-response paradigm, where participants are asked to provide two consecutive responses to a problem (Thompson et al., 2011). During the first (initial) response stage participants see the problem and are asked to give the very first answer that comes to mind. Then, during the second (final) response stage, they are presented with the problem again and are asked to

reflect on it before providing their final answer. Because of the instruction differences, the initial response is thought to be provided predominantly through System 1 processing with minimal System 2 involvement, while the final response is thought to be given predominantly through deliberate, System 2 processing (Thompson et al., 2011). In an attempt to minimize System 2 engagement during the initial stage, recent studies ask participants to provide their first response under a strict deadline and a cognitive load (e.g., a parallel task taxing their cognitive resources). Since System 2 requires cognitive resources to operate, these constraints force participants to provide their answers intuitively during the initial stage (Bago & De Neys, 2017). Hence, the two-response paradigm allows us to directly compare intuitive and deliberate responses on the same problem.

Studies using this paradigm have shown that the higher the conflict detection at the initial response stage, the more likely participants' answers are to change in the final stage (Bago & De Neys, 2017, 2020; Thompson & Johnson, 2014). In other words, if reasoners feel more conflicted (i.e., less certain) about their initial response, they are more likely to change it after they are given the time to deliberate. This (un)certainty about the initial response is also being referred to as the "Feeling of Rightness" (FOR, Thompson et al., 2011). That is, the lower the feeling of rightness (i.e., the confidence) that reasoners show at the initial answer, the more likely it is for them to reconsider their answer in the final stage (Thompson et al., 2011).

Recent dual process models have presented a new conceptualization to account for the conflict detection and two-response findings (Bago & De Neys, 2017, 2019a; De Neys & Pennycook, 2019; Handley et al., 2011; Pennycook et al., 2015, Newman et al., 2017; see De Neys, 2017, for review). In essence, these models postulate that the "logical" response that has traditionally been considered to be cued by System 2, can also be cued by System 1. The main idea is that System 1 can not only give rise to "heuristic" intuitions, which cue responses that contradict logic, but also to "logical" intuitions, which cue responses that are in line with logical principles. The latter are believed to be based on an intuitive/automated understanding of probabilistic and mathematical rules. The most dominant "type" of intuition (i.e., heuristic or logical) will be the one to eventually prevail. Let's imagine that the two competing intuitions–"heuristic" and "logical"–have a large difference in their activation levels, with one's strength dominating over the other's. In that case, there will be little conflict experienced

when generating an initial response and it will be unlikely that the reasoner engages in deliberation and changes their response. Instead, if the two types of intuitions have very similar activation levels, conflict will be maximal and it will be more likely that the reasoner will engage in deliberation to correct their initial response (Bago & De Neys, 2019a; De Neys & Pennycook, 2019; Pennycook et al., 2015; Trippas & Handley, 2017).

Our rationale in the present study was that the same mechanism that drives answer change from the initial to the final response in a single trial, might also drive answer change across a longer time window, for example at different test occasions. Our aim was to explore whether the conflict detection at the initial, intuitive response of a given test session, is related to the response change at a later re-test session (both at the intuitive and at the deliberate level). The reasoning behind this is similar to the one described above: the more dominant one intuition is compared to its competitor (e.g., say, one is strength "9 out of 10" and the other is strength "2 out of 10"), the less conflict is created, and the more likely it should be that it will keep dominating over the weaker intuition at a future test occasion. The more similar the two intuitions are in strength (e.g., one is strength "5 out of 10" and the other is strength "6 out of 10"), the higher the conflict that is created, and the more likely it is that potential random noise (e.g., 1 unit variability due to participants' concentration, level of tiredness etc.) will reverse the strength ordering and make the other intuition dominate, thus, leading to answer change.[2]

Above we sketched the theoretical background that inspired our rationale. However, we can clarify the core idea with a simple non-theoretical analogy. Imagine one has a choice between two desserts; ice-cream or cupcakes. Person A really likes cupcakes, but dislikes ice-cream, while Person B likes both equally well. When you ask Person A about their decision, they will have little doubt about it given their dominant preference and, if you ask them again next week, it is very likely that they will make the same decision. Person B, however, will presumably face a hard decision since they like both desserts but they have to choose one.

[2] Since the dominance of two intuitions of similar strength can be reversed by random noise, we should note that this reversal can go both ways. More specifically, a participant's heuristic response at the first test session can be turned into a logical response at the re-test session and vice versa (i.e., a logical response at the test session can become a heuristic response at the re-test session).

Whatever the final choice of Person B is, they will presumably be less confident that they made the right decision and it is more likely that they will choose differently if they are asked at another time in the future. It is in this sense that we expect response conflict (or inversely response confidence) to be predictive of response stability. The stronger the preference, the less conflict or doubt there will be about the decision, so the more likely it is that one's choices will remain stable over time.

To test whether conflict detection can be predictive of response stability, we asked participants to solve a set of heuristics-and-biases tasks (test session 1), and re-contacted them again two weeks later to solve the same tasks again (test session 2). We used the two-response paradigm for both test sessions. We hypothesized that participants who showed higher conflict detection in their initial, intuitive response at session 1, would be less stable in their responses between session 1 and session 2 (both at the intuitive and the deliberate level). For the calculation of conflict detection we focused on initial trials, as they offer a purer measure of conflict that is independent of deliberation (Bago & De Neys, 2019a).

Method

Preregistration

The study design and research question were preregistered on the Open Science Framework (https://doi.org/10.17605/OSF.IO/8FN3U). No specific analyses were preregistered.

Participants

We recruited our participants online on Prolific Academic (www.prolific. ac). Only native English speakers from Canada, Australia, New Zealand, the United States of America, or the United Kingdom were allowed to take part in the study. There were two test sessions that were two weeks apart. Participants were re-contacted two weeks after the first test session. The second session was not announced during session 1. Hence, participants were not aware that they were going to be re-tested before they were re-contacted. Participants were paid respectively £1.7 and £2 for their participation in session 1 and 2.

We initially recruited 200 participants of which 132 completed both test sessions. Of these 132, 60 had to be discarded because of a randomization coding

error. We therefore recruited an additional 100 participants of which 79 completed both test sessions. This resulted in a total sample of 151 participants who completed both test sessions as intended. The mean age of these participants was 36.5 years (*SD* = 13.4) and 60.2% of them were female. Thirty-eight percent had a high school degree as their highest education level and 47% had a bachelor's degree. All reported data concern the results of these 151 participants who completed both test sessions.

Materials

Counterbalancing

Participants were presented with four different reasoning tasks (i.e., bat-and-ball, base-rates, syllogisms and conjunction fallacy tasks). Each task was composed of eight conflict and eight no-conflict problems. For every reasoning task two sets of items were created in which the conflict status of each item was counterbalanced. More specifically, all the conflict items of the first set appeared in their no-conflict version in the second set, and all the no-conflict items in the first set appeared in their conflict version in the second set. Half of the participants were presented with the first set of problems while the other half was presented with the second set. This way, the same content was never presented more than once to a participant and everyone was exposed to the same items, which minimized the possibility that mere item differences influence the results. The presentation order of the tasks and the items within each task was randomized.

Bat-and-ball problems (BB)

Each participant was presented with eight bat-and-ball problems in multiple-choice format (four conflict and four no-conflict) taken from Raoelison and De Neys (2019). Although the amounts and the names of the objects varied between items, all items shared the same structure with the classic bat-and-ball problem. Participants were always provided with two answer options; a logical answer ("5 cents" in the original bat-and-ball), which was also considered as correct, and a heuristic answer ("10 cents" in the original bat-and-ball), which was considered as incorrect. An example of the problems is presented below:

A national park has 650 roses and lotus flowers in total.

There are 600 more roses than lotus flowers.

How many lotus flowers are there ?

- *25*
- *50*

The no-conflict versions were constructed by removing the "more than" statement from the conflict versions. For instance, in its no-conflict version the above example would become "A national park has 650 roses and lotus flowers in total. There are 600 roses. How many lotus flowers are there?". Each problem was presented in two stages. First, the first sentence was presented for 2000 ms. Afterward, the second sentence along with the question and the answer options was added until a response was given or until the deadline. As in Bago and De Neys (2019), the deadline for the initial response was 4000 ms.

Base-rate problems (BR)

The base-rates problem presentation format was based on Pennycook et al's (2014) rapid-response paradigm. The sentences of each problem were presented serially and the amount of text that was presented on the screen was minimized. Participants were presented with eight base-rate problems (four conflict and four no-conflict) taken from Pennycook et al. (2014). Each problem consisted of a sentence describing the composition of a sample (e.g., "This study contains businessmen and firemen."), a sentence with a stereotypical description of a random person from the sample (e.g., "Person 'K' is brave.") and a sentence with the base-rate information (e.g., "There are 996 businessmen and 4 firemen."). Participants were then asked to choose the group that the random person most likely belonged to. The answer option that was considered correct was always the one that corresponded to the vast majority of the people in the sample. An example of the problems is presented below:

This study contains businessmen and firemen.

Person 'K' is brave.

There are 996 businessmen and 4 firemen.

Is Person 'K' more likely to be:

- *A businessman*
- *A fireman*

The no-conflict versions were constructed by reversing the base-rates of the conflict versions. For example in its no-conflict version, the second sentence of the above problem would read "There are 4 businessmen and 996 firemen". Each problem was presented in three stages. First, the first sentence was presented for 2000 ms. Then, the second sentence was added for another 2000 ms, and finally the critical base-rate information along with the question and the answer options were added until a response or until the deadline. As in Bago and De Neys (2017), the deadline for the initial response was 3000 ms.

Syllogistic reasoning problems (SYL)

Each participant was presented with eight syllogistic reasoning problems, four conflict and four no-conflict, taken from Bago and De Neys (2017). Each problem consisted of a major premise (e.g., "All fruits can be eaten."), a minor premise (e.g., "Strawberries are fruits.") and a conclusion (e.g., "Strawberries can be eaten."). Participants were told to always consider the premises as true and were asked to say if the conclusion followed logically from the premises or not. A conclusion was considered logical only when it was valid. An example of the problems is presented below:

All fruits can be eaten.

Strawberries can be eaten.

Strawberries are fruits.

Does the conclusion follow logically?

- *Yes*
- *No*

In the conflict problems, the believability and the validity of the problems were in conflict, meaning that a syllogism was either valid and unbelievable or invalid and believable. For instance, in the above conflict problem the syllogism is believable, but invalid. On the contrary, in the no-conflict problems, the syllogisms were either valid and believable or invalid and unbelievable. For example, the valid and believable no-conflict version of the above problem would read: "All fruits can be eaten. Strawberries are fruits. Strawberries can be eaten". Each problem was presented in three stages. First, the first sentence of the problem was presented for 2000 ms. Then, the second sentence was added for 2000 ms., and finally the conclusion along with the question and the answer options were added until a

response was given or until the deadline. As in Bago and De Neys (2017), the deadline for the initial response was 3000 ms.

Conjunction fallacy problems (CONJ)

Each participant was presented with eight conjunction fallacy problems, four conflict and four no-conflict, that were taken from Frey et al. (2018), apart from one item (i.e., the Linda problem) which was adapted from the material of Tversky and Kahneman (1983). Each problem consisted of a stereotypical description of an individual followed by two statements about this individual, and participants were asked to choose the statement that was more likely to be true. The first answer option consisted of a single statement related to the individual (e.g., "Jon plays in a rock band"), while the second response option was a conjunction of the first statement with a second statement (e.g., "Jon plays in a rock band and is an accountant"). One of the two statements had a strong fit with the stereotypical description, while the second one had a lower fit. Since the possibility of a single event occurring is always higher than the possibility of the conjunction, the single statement was always considered as the correct choice. An example of the problems is presented below:

John 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 statement is most likely:

- *John plays in a rock band*
- *John plays in a rock band and is an accountant*

The no-conflict versions were created by replacing the singular option with the statement that showed a strong stereotypical fit to the description. For instance, in the no-conflict version of the above example the two answer options would be : Option 1: "John is an accountant", Option 2: "John is an accountant and plays in a rock band". Each problem was presented in two stages. First, the first part of the problem (description) was presented for 4000 ms. Then the critical question and answer options were added and remained on screen until a response was given or until the deadline. The deadline for the initial response was 5000 ms (see Boissin et al., 2021).

Two-response format

We used the two-response paradigm (Thompson et al., 2011) for the presentation of all items. In this paradigm participants are asked to provide two consecutive responses on every trial (see Procedure). The paradigm's format was based on recent studies in which, during the initial response, participants are asked to perform a load memorization task as well as to respond under a strict deadline, which is pre-tested to be demanding for the respective task (e.g., Bago & De Neys, 2017, 2019a; Boissin et al., 2021; Raoelison et al., 2020). During the final response there is no load or deadline. As already mentioned, System 2 requires cognitive resources to operate, so by restricting the processing time and adding a memorization load during the first stage, System 2 involvement is minimized. As a result, one can be maximally sure that the initial response is provided intuitively (i.e., without deliberation), while in the final response stage reasoners are allowed to deliberate. The load memorization task that we used was a complex visual pattern (i.e., 4 crosses in a 3 × 3 grid) and it was briefly presented before each problem (Miyake et al., 2001). After providing an initial response, participants were presented with four different load patterns (i.e., with different cross placings) and had to identify the one that they had been asked to memorize.

Procedure

The experiment was run online using the Qualtrics platform. Participants were told that the study would take 20 minutes to complete and that it demanded their full attention. They were first presented with a general description of the task, where they were informed that they would have to provide two consecutive responses to various reasoning problems. More specifically, they were told to first answer with the very first answer that came to their mind and then reflect on the problem before providing their final response (see Bago & De Neys, 2017 for literal instructions). In order to familiarize themselves with the two response procedure, they first solved two simple mathematical problems (addition and subtraction) with the two response format. Then, they practiced the load task alone, by solving two memorization trials. Finally, they practiced the two math problems in their full two-response format (problem + deadline and load task on initial response). After the practice, participants started the main task which consisted of four blocks and

32 reasoning problems (eight problems per block). Each block consisted of a single task (i.e., either bat-and-ball, base-rates, syllogisms or conjunction fallacies). At the start of each block participants received specific instructions for the respective task, they were shown an example problem and solved a practice problem. Each trial started with a fixation cross shown for 2000 ms. Then the first part of the problem was presented (for more details see Materials subsections for each reasoning task), followed by the matrix for the cognitive load task which remained on screen for 2000 ms. Then the whole problem was presented, along with the question and the answer options. Participants could provide their initial response by clicking on one of the answer options. One second before the deadline, the screen turned yellow to remind participants that the deadline was approaching. If they did not respond within the deadline, they were presented with a message asking them to try and respond within the deadline on the next trials. If they responded within the deadline, they were asked to rate their confidence in the correctness of their initial response on a scale from 0 (absolutely not confident) to 100 (absolutely confident). Immediately after, participants were shown four matrices and were asked to recall the test matrix. They were then given feedback on the correctness of their recall. Finally, participants viewed the full problem again and were asked to provide their final answer. Next, they were asked their confidence in the correctness of their final response.

Participants were re-contacted after two weeks to complete session 2 of the study, which was fully identical to session 1.

Trial exclusion

The trials in which participants failed the load and/or the deadline were excluded from subsequent analyses, since in these trials we could not ensure that deliberation was minimized during the initial stage. Participants failed to answer before the deadline on 4.6% of conflict initial trials (224 out of 4832) and 3.6% of no-conflict initial trials (175 out of 4832) of both test sessions combined. In addition, participants failed the load task on 14.9% of conflict initial trials (719 out of 4832) and 11.8% of no-conflict initial trials (572 out of 4832) of both test sessions. Overall, by rejecting the missed deadline and missed load trials, we kept 80.5% of conflict initial trials (3889 out of 4832) and 84.5% of no-conflict initial trials (4085 out of 4832) in session 1 and session 2 combined.

Conflict detection index

As mentioned before, conflict detection is typically calculated by subtracting the baseline confidence (i.e., the confidence at the correct no-conflict trials), from the confidence at the conflict trials (De Neys et al., 2013; Frey et al., 2018; Mevel et al., 2015; Pennycook et al., 2015). The higher the difference between the two, the more conflict is thought to be experienced by the participant. However, when reasoners deliberate on a problem, the initial doubt that they might have felt in relation to it can be dissolved (e.g., Bago & De Neys, 2020; De Neys et al., 2013). In this case, their reported confidence will not be a pure measure of the conflict that they initially experienced. To tackle this issue, previous one-response studies discarded correct conflict trials when calculating conflict detection, as in these trials the heuristic response had been overcome (i.e., the conflict associated with it had been resolved, e.g., De Neys & Glumicic, 2008; Pennycook et al., 2015; Šrol & De Neys, 2021). For the same reason, studies that use the two-response paradigm focus on the confidence of the initial responses for the calculation of conflict detection. At this stage deliberation is experimentally minimized. Consequently, conflict detection at this stage gives a purer measure of intuitively experienced conflict, which should more directly reflect the strength of the posited intuitions (Bago & De Neys, 2017).

In addition, by using the confidence at the initial, intuitive trials, one can analyse both incorrect and correct conflict trials, since even correct trials will not be contaminated by deliberation. Note that participants still had to memorize the cognitive load pattern while providing their initial response confidence, which further ensured that their confidence was not affected by post-decision reflection.

Following the above studies and our preregistration, in the present paper we therefore focused on initial conflict detection. Response confidence was recorded both for the initial and the final responses, but we were a priori interested in the initial stage. Likewise, we only used confidence and not reaction times for the calculation of conflict detection, as the latter has been shown to be a less reliable indicator of detection ability (Frey et al., 2018; Šrol & De Neys, 2020), especially in a two-response setting (Bago & De Neys, 2017).

Finally, note that the rare trials in which no-conflict problems were solved incorrectly were discarded for the conflict detection analysis, since it is hard to

interpret these unequivocally (see De Neys & Glumicic, 2008; Pennycook et al., 2015).

Composite Measure

For simplicity and to maximize power, our analyses focused on the composite level across the four individual reasoning tasks. To calculate this composite performance, for each participant, we calculated the proportion of correct initial and final responses for the conflict and no-conflict problems in each of the reasoning tasks and in each session. Then we averaged across all reasoning tasks (separately for each session, each response stage and conflict and no-conflict trials). For completeness, the individual task data is also included in our figures. Overall, the composite trends were reflected in the individual tasks.

Results

Statistical Analysis

The data were processed and analysed using the R software (R Core Team, 2020) and the following packages (in alphabetical order): dplyr (Wickham et al., 2021), ez (Lawrence, 2016), ggplot2 (Wickham, 2016), ggpubr (Kassambara, 2020), Rmisc (Hope, 2013), rstatix (Kassambara, 2021), and tidyr (Wickham, 2021).

Accuracy

To see if there was an effect of the response stage (initial; final) and the session (session 1; session 2) on the accuracy of conflict problems, a two-way within-subjects ANOVA was conducted. As Figure 1 shows, the accuracy at the conflict problems was significantly higher in the final than the initial response stage, $F(1, 150) = 11.07$, $p < .01$, $\eta^2g = 0.003$, which suggests that accuracy improved after deliberation. In addition, the accuracy at the conflict problems was significantly higher in session 2 compared to session 1, $F(1, 150) = 22.65$, $p < .001$, $\eta^2g = 0.01$, indicating that participants slightly improved when given a second chance to solve the problems. This improvement was independent of the response stage, as indicated by the lack of interaction between response stage and session, $F(1, 150) = 2.42$, $p = .12$; $\eta^2g < 0.001$.

As Figure 1 shows, these composite level trends were also observed for each individual task separately, with the exception of the conjunction fallacy problems in which final responses tended to be slightly less accurate than initial responses (see Dujmović et al., 2021, for a similar observation).

As expected, the average accuracy at the no-conflict problems remained at ceiling both for initial (*M* = 90.3, *SD* = 6.6 in session 1; *M* = 89.9, *SD* = 6.7 in session 2) and final responses (*M* = 92.9, *SD* = 7.0 in session 1; *M* = 92.3, *SD* = 7.6 in session 2), showing that participants paid attention throughout the study and refrained from guessing.

To summarize, although deliberation led to a slight improvement in performance, participants remained typically biased when solving classic conflict tasks. Overall, these results are in line with previous two-response studies (e.g., Bago & De Neys 2017, 2019a; Thompson et al., 2011).

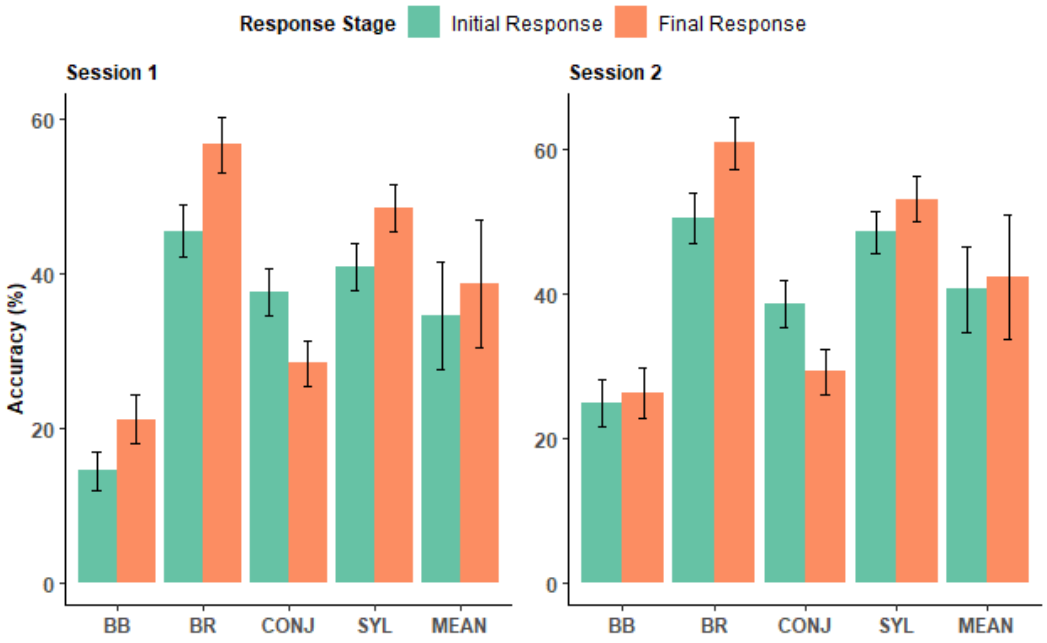

Figure 1. Proportion (%) of correct responses on the conflict problems, separately for each response stage, each session, each reasoning task and for the composite mean across the four tasks. The error bars represent the Standard Error of the Mean. BB = Bat-and-ball; BR = Base-rates; CONJ = Conjunction Fallacies; SYL = Syllogisms; MEAN = the composite mean across the four tasks.

Direction of change

We also conducted a direction of change analysis on the conflict problems to explore whether and how participants changed their responses after deliberation (Bago & De Neys, 2017, 2019a). More specifically, we looked into how their accuracy changed (or did not change) from the initial to the final stage in every trial. At each response stage participants could either have an accuracy of "1" (i.e., correct response) or an accuracy of "0" (i.e., incorrect response). Since participants always provided two responses in a trial, we end up with four possible response patterns: "00" (incorrect initial and incorrect final response), "01" (incorrect initial and correct final response), "10" (correct initial and incorrect final response) and "11" (correct initial and correct final response). The results were consistent with previous findings (Bago & De Neys, 2017, 2019a). As Figure 2 shows, at the composite level, the majority of the conflict trials had a "00" pattern both in session 1 (54.8%) and in session 2 (52.2%), which confirms that reasoners are easily lured by the heuristic response when solving classic heuristics-and-biases tasks. We also note that, in the conflict trials, the proportion of "11" responses (28.0% in session 1; 35.1% in session 2) was higher than that of the "01" responses (10.6% in session 1; 7.2% in session 2). As in previous two-response studies (e.g., Bago & De Neys, 2017, 2019a; Newman et al., 2017), this indicates that correct responses are, for the most part, already generated intuitively and not after deliberation. Finally, the least prevalent response pattern was "10" (session 1: 6.6%; session 2: 5.4%). As Figure 2 shows, these patterns were also observed on each of the individual tasks.

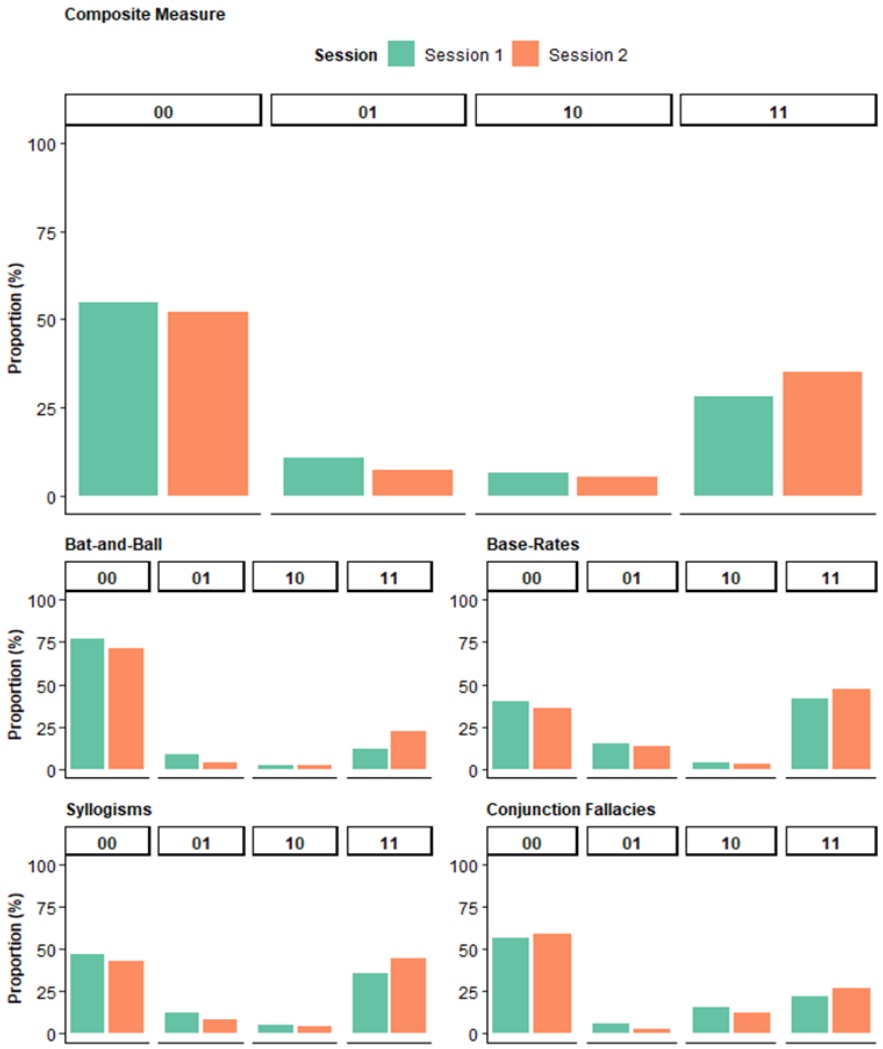

Figure 2. Proportion of each direction of change (i.e., "00" trials, "01" trials, "10" trials and "11" trials) for the conflict trials according to each session, each reasoning task, and the composite measure across the four reasoning tasks. "00" = incorrect initial and final response; "01" = incorrect initial and correct final response; "10" = correct initial and incorrect final response; "11" = correct final and correct initial response.

Accuracy Correlations

Before moving on to the core stability analyses we also examined whether the average accuracy of each individual at session 1 was correlated with the accuracy of that individual at session 2. A Pearson's product-moment correlation test revealed a high, positive accuracy correlation both for initial conflict trials, $r = 0.77$, $t(149) = 14.65$, $p < .001$, and for final conflict trials, $r = 0.84$, $t(149) = 18.73$, $p < .001$. The same pattern was observed for the individual tasks (see Supplementary Material Section A). Hence, this indicates that those individuals

who scored best the first time around, remained scoring well at the-retest. In this sense, the heuristics-and-biases tasks had a high test-re-test reliability.

Stability Index

Next, we investigated the stability of responses from session 1 to session 2. Stability is an inherently different measure of participants' responding than accuracy. To illustrate, consider an example of an exam with yes/no responses consisting of 20 items. The expected accuracy of an unprepared student is 50%. Now imagine that this student, still unprepared, had retaken the exam in the second term and always selected the opposite response compared to the first term. Their accuracy would still be 50%, but their stability would be 0%.

We separately calculated the stability of initial and final responses. Note that with respect to final responses, there are four possible patterns of (in)stability from session 1 to session 2: : "s00" (incorrect final response at both sessions), "s01" (incorrect final response at session 1 and correct final response at session 2), "s10" (correct final response at session 1 and incorrect final response at session 2), and "s11" (correct final response at both sessions). If the final response pattern of an individual item was "s00" or "s11", this item was categorized as "stable", whereas if the pattern was "s01" or "s10", the item was characterized as "unstable". The same stability classification was made for initial responses. These patterns should not be confused with the aforementioned direction of change patterns, hence the added "s", which stands for "stability". While the direction of change deals with the accuracy change from the initial to the final response of a trial, the direction of (in)stability deals with the accuracy change of a response (initial or final) from session 1 to session 2.

After all individual items were categorized as either stable or unstable, the average stability was calculated for each participant. As Figure 3A shows, we observed a very high stability both at the composite level and for each individual task, for the initial and final responses (initial response composite: M = 78.7%, SD = 17.2%; final response composite: M = 83.1%, SD = 14.6%). For completeness, note that we also observed the same pattern at the no-conflict trials (initial response composite: M = 90.2%, SD = 11.5%; final response composite: M = 93.3%, SD = 10.5%). This indicates that overall people's performance is highly stable after two weeks and reasoners rarely change their answers.

Direction of Change Stability

After establishing the stability of initial and final responses, we took a step further and examined the stability of the direction of change patterns from session 1 to session 2. More precisely, if a participant's trial had the same direction of change both in session 1 and session 2, this trial was coded as having a stable direction of change, and vice versa. We found that the stability of the direction of change category was high, both for conflict (M = 70.7%, SD = 19.7%) and no-conflict (M = 86.5%, SD = 15.1%) problems, which confirms that participants' response patterns were very consistent in time. More specifically, this finding indicates that, for the vast majority of the trials, the way people changed (or did not change) their initial responses after deliberation in session 1, was typically the way they changed them when re-tested two weeks later. As Figure 3B shows, the same trends were observed for the individual tasks.

However, at the same time it is clear that neither the responses nor the direction of change categories remained 100% stable from session 1 to 2, and we can still notice some response variability, especially so on the conflict problems. Our main aim was to see if conflict detection could explain this variability.

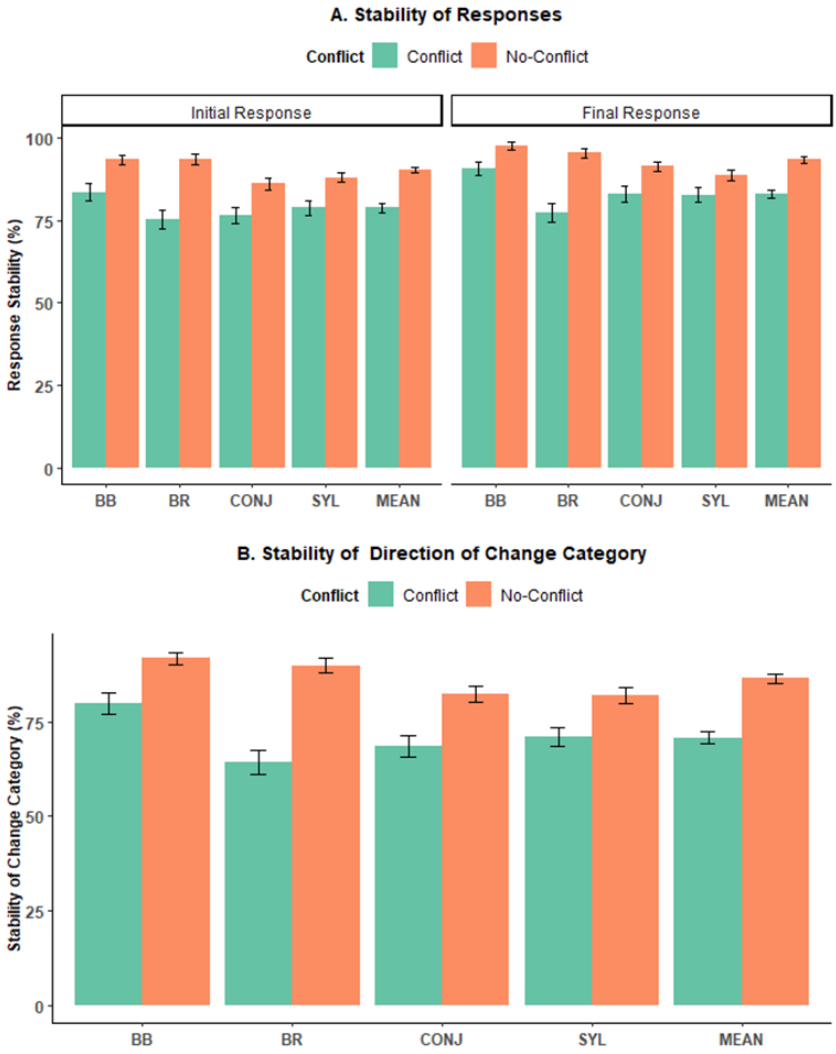

Figure 3. Panel A shows the proportion of responses that remained stable from session 1 to session 2, separately for conflict and no-conflict problems, for each response stage, each reasoning task and for the composite mean across the four tasks. Panel B shows the proportion of trials that had a stable direction of change category (i.e., "00" trials, "01" trials, "10" trials and "11" trials) from session 1 to session 2, separately for each response stage, each reasoning task and for the composite mean across the four tasks. The error bars represent the Standard Error of the Mean. BB = Bat-and-ball; BR = Base-rates; CONJ = Conjunction Fallacies; SYL = Syllogisms; MEAN = the composite mean across the four tasks.

Conflict Detection

As a reminder, the conflict detection was calculated from the confidence ratings at the initial responses in the following manner: Confidence $_{conflict}$ – Confidence $_{no\text{-}conflict\_correct}$. For comparison with previous studies, we first wanted to check whether we observed an overall lower confidence on conflict versus no-

conflict trials, pointing to a group-level conflict detection effect. This was indeed the case across tasks, responses and sessions (see Supplementary Material Section B). In addition, we also wanted to verify whether conflict detection was more pronounced on trials in which reasoners changed their response after deliberation ("01" and "10" trials), compared to trials in which reasoners did not change their response after deliberation ("00" and "11" trials, e.g., see Bago & De Neys, 2017, 2020; Thompson et al., 2011)[3]. As Figure 4 shows, this pattern was consistently observed across tasks, responses and sessions. As in previous work, these results show that the higher the conflict experienced during an initial response, the more likely for this response to change in the final stage. Hence, both with respect to response accuracy and conflict detection, our findings are in line with previous two-response studies.

[3] Note that we used only the dominant no-conflict "11" category for this contrast, as responses in the other no-conflict direction of change categories cannot be interpreted unequivocally.

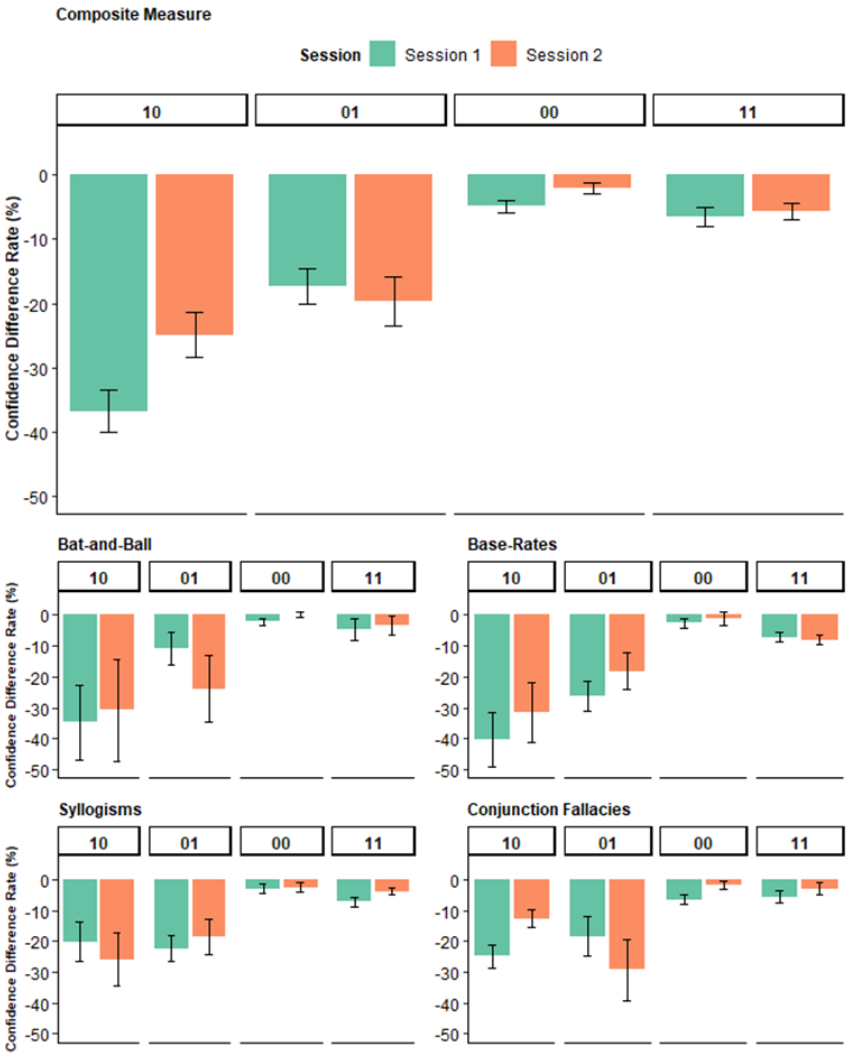

Figure 4. The mean confidence difference rate (%) according to the direction of change category (i.e., "01" trials and "10" trials represent the "change" categories, while "00" and "11" trials represent the "no change" categories), separately for each session, each reasoning task, and the composite measure across the four reasoning tasks. Negative values point to an overall successful conflict sensitivity. The error bars represent the Standard Error of the Mean. BB = Bat-and-ball; BR = Base-rates; CONJ = Conjunction Fallacies; SYL = Syllogisms; MEAN = the composite mean across the four tasks.

As one reviewer suggested, for comparison with previous one-response studies (e.g., De Neys & Glumicic, 2008; Pennycook et al., 2015; Šrol & De Neys, 2021), we re-ran this analysis by discarding the correct conflict trials when calculating conflict detection (see Supplementary Material Figure S2 for the conflict detection means and Figure S3 for the conflict detection means for each

direction of change category). Overall, the patterns and conclusions were consistent.

Predictive Conflict Detection

We now turn to the test of our main research question, in which we examine whether (initial) conflict detection at session 1 can predict response stability two weeks later, both at the intuitive and the deliberate level. In order to calculate conflict detection for every item of each participant, we first categorized the items of each participant as either "stable" or "unstable". More specifically, if a participant's accuracy at a given Item 1 was the same in session 1 and session 2, Item 1 would be classified as "stable" and vice versa. Once we classified all items as either "stable" or "unstable", we calculated, for each participant, the average conflict detection at all their stable items combined, and at all their unstable items combined. This way, each participant had two conflict detection indices: one for their stable and one for their unstable items. Inevitably, there were some participants whose items were all stable or all unstable throughout the study. Since these participants only had one conflict detection index (either for their stable or for their unstable items), they were examined separately. In the analyses below we were mainly interested in the composite measure and not the differences between the reasoning tasks. For completeness, we also report the data for each individual task. However, these individual task level analyses often have low sample sizes so they should be interpreted with some caution.

Note that as suggested by one reviewer, we also ran this analysis using the absolute confidence values at the initial conflict problem responses, also known as the feeling of rightness (Thompson et al., 2011), instead of the conflict detection indices (see Supplementary Material Figure S5 for the mean confidence values). As Supplementary Material Figure S6 shows, this type of analysis yielded the same pattern of results (see Supplementary Material Table S3 for the significance tests). In addition, we ran the same analysis after discarding the correct conflict trials when calculating conflict detection. As Figure S4 in the Supplementary Material shows, this analysis revealed the same results (see Supplementary Material Table S2 for the significance tests).

Initial Detection and Final Stability

By calculating the grand mean of conflict detection, we found that the initial detection was overall higher for the items that had unstable final responses (M = −9.9, SD = 11.6), compared to the initial detection of the items that had stable final responses (M = –7.3, SD = 9.6). This trend agrees with our hypothesis and, as Figure 5A shows, it is observed in all individual reasoning tasks. To test the statistical significance of these results we compared participants' composite conflict detection index at their stable and at their unstable items. Evidently, we only included the subjects that had both stable and unstable items (n = 114). Any participants with solely stable items were discarded from this analysis (there were no participants with only unstable items). A paired-samples t-test revealed a significant difference in the conflict detection indices between stable (M = –5.6, SD = 11.3) and unstable (M = –12.0, SD = 22.1) items; $t(113) = 3.05$, $p < .01$. As expected, the unstable items had a higher conflict detection compared to the stable ones. It is worth noting that participants with only stable items (n = 37), had a very low average conflict detection (M = –3.8, SD = 6.9).

Initial Detection and Initial Stability

Next, we performed the same analysis as above, but now we focused on how initial conflict detection impacted the initial, intuitive responses at session 2. Consistent with the above results, we found that the grand mean of the composite conflict detection index was overall higher for the items with unstable initial responses (M = –19.1, SD = 20.7), compared to the conflict detection of the items with stable initial responses (M= –6.0, SD = 8.8). As Figure 5B shows, this trend was observed on all individual reasoning tasks. To test the statistical significance of these results we compared participants' composite conflict detection index at their stable and at their unstable items. Evidently, we only included the subjects that had both stable and unstable items (n = 122). Any participants with solely stable items were discarded from this analysis (there were no participants with only unstable items). A paired samples t-test revealed a significant difference in the conflict detection scores between stable (M = –3.5, SD = 7.8) and unstable (M = –18.5, SD = 26.1) items, $t(121) = 6.19$, $p < .001$. It is worth noting that participants that had only stable items (n = 29), had a low average conflict detection (M = –3.2, SD = 7.2).

Our main a priori conflict detection measure concerned the detection at the initial, intuitive response level. For exploratory purposes, we repeated the analysis, this time using the conflict detection at the final responses as a predictor of (initial and final) response stability. Supplementary Material Figure S7 shows the results. Although the trends tended to be slightly weaker, overall the same pattern was observed, in that unstable trials showed a more pronounced conflict detection than stable trials.

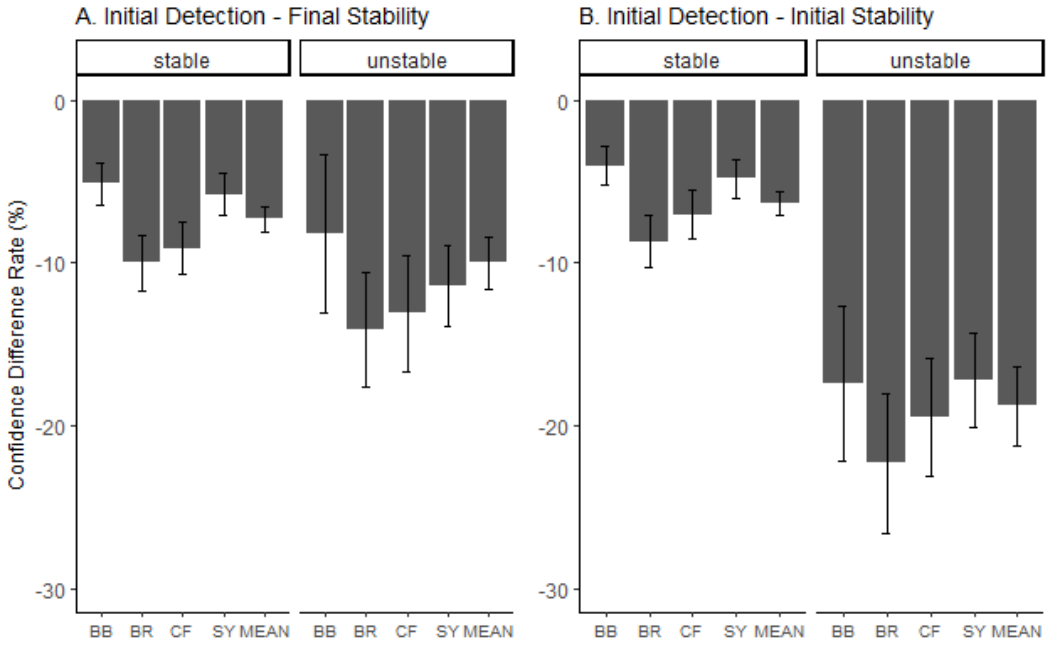

Figure 5. The grand means of the initial conflict detection index (i.e., Confidence $_{conflict}$ – Confidence $_{no\text{-}conflict\_correct}$) according to stability (stable; unstable). Panel A shows the average initial conflict detection according to the stability of the final responses and Panel B shows the average initial conflict detection according to the stability of the initial responses, separately for each reasoning task and for the composite mean across the four tasks. Negative values point to an overall successful conflict sensitivity. The error bars represent the Standard Error of the Mean. BB = Bat-and-ball; BR = Base-rates; CF = Conjunction Fallacies; SY = Syllogisms; MEAN = the composite mean across the four tasks.

Discussion

In the present paper we focused on the temporal stability of reasoning performance and examined a potential determinant for answer change. Participants solved the same tasks twice in two test sessions, two weeks apart.

We used the two-response paradigm to test the stability of both initial (intuitive) and final (deliberate) responses. We hypothesized that participants who showed higher conflict detection in their initial, intuitive responses at session 1, would be less stable in their responses between session 1 and session 2. Conflict detection was operationalized as the confidence difference for initial responses on classic conflict problems versus control no-conflict problems.

Results point to two main conclusions. First, people's responses to classic "bias" tasks are highly stable. In general, participants rarely changed their intuitive and deliberate answers two weeks after they were first tested. This result is in line with the findings by Białek and Pennycook (2018) and Stango and Zinman (2020) who—in one of the rare direct tests of the stability of heuristics-and-biases tasks—also observed that individual biases remained highly stable over time. From a psychometric perspective, the high stability of the performance on heuristics-and-biases tasks is obviously excellent news. This is particularly important as the performance on these tasks is frequently used in the literature as a predictor of a wide range of variables (e.g., Baron et al., 2015; Białek & Sawicki, 2018; Shenhav et al., 2012; Toplak et al., 2017; West et al., 2008). If people's task performance would not be stable, this would undermine its use as a predictor. In this sense, the findings validate the popular use of these tasks by showing that they exhibit an adequate test-retest reliability.

Second, despite the high stability, there was still some variability in initial and final responses after the first test. By directly comparing the conflict detection for items that had a stable accuracy to those that had an unstable accuracy, we found that the initial conflict detection was significantly higher in the unstable items. In other words, the higher the initial conflict detection participants experienced on an item, the more likely they were to change their responses to this item two weeks later. This finding indicates that the variability of responses over time is not entirely random, but can be predicted (Stango & Zinman, 2020).

At the methodological level, we believe that the current findings further underline the potential of the two-response paradigm (Thompson et al., 2011), which has become increasingly popular in the past years (e.g., Bago & De Neys, 2017, 2019a, 2021; Burič & Konradova, 2021; Burič & Srol, 2020; Dujmovic et al., 2021; Vega et al., 2021). As we have mentioned, previous work showed that conflict detection can predict answer change on an intra-trial level. Conflict detection during the initial stage is much more pronounced in the cases that

participants change their initial answers during the final response stage (Bago & De Neys 2017, 2019a; Thompson & Johnson, 2014). With the present study, we show that conflict detection at the initial stage does not only predict answer change in the short, intra-trial term, but also in the longer term, between separate test sessions. The generalization of the conflict detection and answer change coupling over a longer time window points to an interesting new application of the paradigm.

At the theoretical level, conflict detection (or a lowered feeling of rightness in the conceptualization of Thompson et al., 2011) is often conceived as a triggering mechanism that allows a reasoner to switch from System 1 intuiting to System 2 deliberation (e.g., De Neys, 2012; Pennycook et al., 2015; Thompson et al., 2011). One consequence of engaging in deliberation is that people might revise their intuitively generated answer (Thompson et al., 2011). With respect to the stability of final responses, this suggests that conflict experienced at time 1 will make it more likely that the reasoner engages in deliberation at time 1, but also at time 2, two weeks later. Because deliberation increases the probability of answer change, it will be more likely that reasoners give a different final response at time 1 and time 2.

But interestingly, our findings not only concerned the final but also at the initial responses. By definition, in the initial response stage deliberation is minimized and, hence, answer change cannot be driven by differential deliberation per se. So why does conflict detection predict initial answer stability? Our hypothesis was inspired by recent advances in dual process theorizing in which the intuitive reasoning performance is determined by the strength interplay of competing intuitions (e.g., Bago & De Neys, 2020; De Neys & Pennycook, 2019; Pennycook et al., 2015). As we noted, these models postulate that the "logical" response that has traditionally been considered to be cued by System 2, can also be cued by System 1. Hence, it is assumed that when reasoners are faced with a traditional heuristics-and-biases task, System 1 will not only give rise to the traditionally postulated "heuristic" intuition, but also to a "logical" intuition (which is assumed to be based on automatically activated learned mathematical and probabilistic rules, e.g., De Neys, 2012). Whichever intuition is strongest will be selected as initial response. The more similar the strength of the competing intuitions, the more conflict will be experienced. If one intuition clearly dominates over the other, the dominant intuition will be generated with little or no

experienced conflict. We reasoned that any accidental noise at different test sessions will be more likely to affect (revert) the strength ordering of competing intuitions that showed little differentiation to start with. Going back to our introductory analogy, the clearer your preference for one dessert over another, the more likely that you will make the same choice repeatedly. Hence, a highly dominant intuition (indexed by low conflict detection) will be more likely to remain dominant at re-test than a less dominant intuition (indexed by high conflict detection). Consequently, conflict detection will also predict answer stability of the intuitive response.

Obviously, this theoretical account remains speculative. The strength of competing intuitions is a hypothetical construct and was not directly measured. We also acknowledge that this construct can be defined in various ways (e.g., processing "fluency" or "speed"). At present, the specific processes underlying the relationship between logical and heuristic intuitions have not been specified, and we do recognise the need for their precise implementation.

It is worth noting that the current findings are also relevant for the discussion on Individual Differences in conflict detection. Previous studies have shown that, although most people might detect the conflict in their answers, not everyone does (e.g., Frey et al., 2018; Pennycook et al., 2015; Šrol & De Neys, 2021). The high response stability in our study and its relation to a low conflict detection, suggests that there are some participants who always remain biased and unaware of their errors. In other words, some reasoners consistently provide incorrect answers (i.e., they do not change their erroneous responses at time 2) and they have low or no conflict detection at time 1.

One may also note that the observed high stability of participants' responses, both on the intra-trial level and between the separate test sessions, suggests that most participants respond on an intuitive basis even when they are given the time to deliberate. However, we would like to highlight that this does not imply that deliberation is never used or needed when it comes to sound reasoning. Although response change was rare in our study, there were still cases in which people engaged in deliberation to correct their intuitive answers (i.e., "01" cases). In addition, recent studies have suggested that deliberation might be helpful to provide explicit justifications for an intuitive insight (see Bago & De Neys, 2019a; De Neys & Pennycook, 2019).

It is clear that the approach we introduced here can be further developed and fine-tuned. For example, for practical reasons (e.g., attrition) the present study focused on a two week time window. This presents a dramatic departure from the millisecond intra-trial time-scale that two-response studies typically focus on to study answer change. But, obviously, one could further expand the timeline and test the predictability of answer stability at time points that are months or even years apart. Likewise, the present study has focused on heuristics-and-biases tasks only. The two-response paradigm has been used to explore answer change in different domains (e.g., moral reasoning, Bago & De Neys, 2019b; Vega et al., 2021; or prosocial reasoning in economic settings, Bago et al., 2021). In theory, the present approach can be adopted to test the predictability of long-term answer change in all these fields.

To conclude, the present study showed that people's responses to heuristics-and-biases tasks are highly stable. The rare cases in which answers are nevertheless changed seem to be driven by the detection of conflict between competing intuitions. We believe that the results point to the potential of the approach and hope that it can inspire new applications in the reasoning and decision-making fields.

References

Bago, B., Bonnefon, J. F., & De Neys, W. (2021). Intuition rather than deliberation determines selfish and prosocial choices. *Journal of Experimental Psychology: General*, *150*(6), 1081–1094. https://doi.org/10.1037/xge0000968

Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition*, *158*, 90–109. https://doi.org/10.1016/j.cognition.2016.10.014

Bago, B., & De Neys, W. (2019a). The Smart System 1: Evidence for the intuitive nature of correct responding on the bat-and-ball problem. *Thinking & Reasoning*, *25*(3), 257–299. https://doi.org/10.1080/13546783.2018.1507949

Bago, B., & De Neys, W. (2019b). The intuitive greater good: Testing the corrective dual process model of moral cognition. *Journal of Experimental Psychology: General*, *148*(10), 1782–1801. https://doi.org/10.1037/xge0000533

Bago, B., & De Neys, W. (2020). Advancing the specification of dual process models of higher cognition: A critical test of the hybrid model view. *Thinking & Reasoning*, *26*(1), 1–30. https://doi.org/10.1080/13546783.2018.1552194

Baron, J., Scott, S., Fincher, K., & Metz, S. E. (2015). Why does the Cognitive Reflection

Test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory and Cognition*, *4*(3), 265–284. https://doi.org/10.1016/j.jarmac.2014.09.003

Białek, M., Muda, R., Stewart, K., Niszczota, P., & Pieńkosz, D. (2020). Thinking in a foreign language distorts allocation of cognitive effort: Evidence from reasoning. *Cognition, 205*, 104420. https://doi.org/10.1016/j.cognition.2020.104420

Białek, M., & Pennycook, G. (2018). The cognitive reflection test is robust to multiple exposures. *Behavior research methods, 50*(5), 1953-1959. https://doi.org/10.3758/s13428-017-0963-x

Białek, M., & Sawicki, P. (2018). Cognitive reflection effects on time discounting. *Journal of Individual Differences. 39*(2), 99–106 https://doi.org/10.1027/1614-0001/a000254

Boissin, E., Caparos, S., Raoelison, M., & De Neys, W. (2021). From bias to sound intuiting: Boosting correct intuitive reasoning. *Cognition*, *211*, 104645. https://doi.org/10.1016/j.cognition.2021.104645

Burič, R., & Konrádová, Ľ. (2021). Mindware Instantiation as a Predictor of Logical Intuitions in the Cognitive Reflection Test. *Studia Psychologica*, *63*(2), 114–128. https://doi.org/10.31577/sp.2021.02.822

Burič, R., & Šrol, J. (2020). Individual differences in logical intuitions on reasoning problems presented under two-response paradigm. *Journal of Cognitive Psychology, 32*(4), 460–477. https://doi.org/10.1080/20445911.2020.1766472

De Neys, W. (2012). Bias and Conflict: A Case for Logical Intuitions. *Perspectives on Psychological Science*, *7*(1), 28–38. https://doi.org/10.1177/1745691611429354

De Neys, W. (2014). Conflict detection, dual processes, and logical intuitions: Some clarifications. *Thinking & Reasoning*, *20*(2), 169–187. https://doi.org/10.1080/13546783.2013.854725

De Neys, W. (2017). Bias, conflict, and fast logic: Towards a hybrid dual process future? In W. De Neys (Ed.), *Dual Process Theory 2.0* (pp. 47–65). Oxon, UK: Routledge.

De Neys, W. (2021). On dual-and single-process models of thinking. *Perspectives on psychological science, 16*(6), 1412-1427. https://doi.org/10.1177/1745691620964172

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition*, *106*(3), 1248–1299. https://doi.org/10.1016/j.cognition.2007.06.002

De Neys, W., & Pennycook, G. (2019). Logic, Fast and Slow: Advances in Dual-Process Theorizing. *Current Directions in Psychological Science*, *28*(5), 503–509. https://doi.org/10.1177/0963721419855658

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. https://doi.org/10.3758/s13423-013-0384-5

Dujmović, M., Valerjev, P., & Bajšanski, I. (2021). The role of representativeness in reasoning and metacognitive processes: An in-depth analysis of the Linda problem. *Thinking & Reasoning*, *27*(2), 161–186. https://doi.org/10.1080/13546783.2020.1746692

Epstein, S. (1994). Integration of the Cognitive and the Psychodynamic Unconscious. *American Psychologist*, *49*(8), 709–724. https://doi.org/10.1037/0003-066X.49.8.709

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol., 59*, 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Evans, J. S. B., & Over, D. E. (2013). *Rationality and reasoning.* Psychology Press. https://doi.org/10.4324/9780203027677

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Ferreira, M. B., Mata, A., Donkin, C., Sherman, S. J., & Ihmels, M. (2016). Analytic and heuristic processes in the detection and resolution of conflict. *Memory & Cognition*, *44*(7), 1050–1063. https://doi.org/10.3758/s13421-016-0618-7

Frey, D., Johnson, E. D., & De Neys, W. (2018). Individual differences in conflict detection during reasoning. *Quarterly Journal of Experimental Psychology*, *71*(5), 1188–1208. https://doi.org/10.1080/17470218.2017.1313283

Gangemi, A., Bourgeois-Gironde, S., & Mancini, F. (2015). Feelings of error in reasoning—In search of a phenomenon. *Thinking & Reasoning*, *21*(4), 383–396. https://doi.org/10.1080/13546783.2014.980755

Handley, S. J., Newstead, S. E., & Trippas, D. (2011). Logic, beliefs, and instruction: A test of the default interventionist account of belief bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(1), 28–43. https://doi.org/10.1037/a0021098

Hope, R. M. (2013). Rmisc: Rmisc: Ryan Miscellaneous. R package version 1.5. https://CRAN.R-project.org/package=Rmisc

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

Kassambara, A. (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr

Kassambara, A. (2021). rstatix: Pipe-Friendly Framework for Basic Statistical Tests. R package version 0.7.0. https://CRAN.R-project.org/package=rstatix

Lawrence, M. A. (2016). ez: Easy Analysis and Visualization of Factorial Experiments. R

package version 4.4-0. https://CRAN.R-project.org/package=ez

Mata, A. (2020). Conflict detection and social perception: bringing meta-reasoning and social cognition together. *Thinking & Reasoning*, *26*(1), 140-149. https://doi.org/10.1080/13546783.2019.1611664

Mata, A., Ferreira, M. B., Voss, A., & Kollei, T. (2017). Seeing the conflict: An attentional account of reasoning errors*. Psychonomic Bulletin & Review, 24*(6), 1980-1986. https://doi.org/10.3758/s13423-017-1234-7

Mevel, K., Poirel, N., Rossi, S., Cassotti, M., Simon, G., Houdé, O., & De Neys, W. (2015). Bias detection: Response confidence evidence for conflict sensitivity in the ratio bias task. *Journal of Cognitive Psychology*, *27*(2), 227–237. https://doi.org/10.1080/20445911.2014.986487

Meyer, A., Zhou, E., & Shane, F. (2018). The non-effects of repeated exposure to the Cognitive Reflection Test. *Judgment and Decision making, 13*(3), 246.

Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. *Journal of Experimental Psychology: General*, *130*(4), 621–640. https://doi.org/10.1037/0096-3445.130.4.621

Newman, I., Gibb, M., & Thompson, V. (2017). Rule-Based Reasoning Is Fast and Belief-Based Reasoning Can Be Slow: Challenging Current Explanations of Belief-Bias and Base-Rate Neglect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(7), 1154–1170. https://doi.org/10.1037/xlm0000372

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2014). Cognitive style and religiosity: The role of conflict detection. *Memory & Cognition*, *42*(1), 1–10. https://doi.org/10.3758/s13421-013-0340-7

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2012). Are we good at detecting conflict during reasoning? *Cognition*, *124*(1), 101–106. https://doi.org/10.1016/j.cognition.2012.04.004

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. https://doi.org/10.1016/j.cogpsych.2015.05.001

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raoelison, M., & De Neys, W. (2019). Do we de-bias ourselves?: The impact of repeated presentation on the bat-and-ball problem. *Judgment and Decision Making, 14*(2), 170–178.

Raoelison, M., Thompson, V. A., & De Neys, W. (2020). The smart intuitor: Cognitive capacity predicts intuitive rather than deliberate thinking. *Cognition*, *204*,

104381. https://doi.org/10.1016/j.cognition.2020.104381

Shenhav, A., Rand, D. G., & Greene, J. D. (2012). Divine intuition: Cognitive style influences belief in God. *Journal of Experimental Psychology: General*, *141*(3), 423–428. https://doi.org/10.1037/a0025391

Sloman, S. A. (1996). The Empirical Case for Two Systems of Reasoning. *Psychological bulletin, 119*(1), 3–22. https://doi.org/10.1037/0033-2909.119.1.3

Šrol, J., & De Neys, W. (2021). Predicting individual differences in conflict detection and bias susceptibility during reasoning. *Thinking & Reasoning*, *27*(1), 38–68. https://doi.org/10.1080/13546783.2019.1708793

Stango, V., & Zinman, J. (2020). Behavioral Biases are Temporally Stable*. Unpublished working paper.* https://doi.org/10.3386/w27860

Stupple, E. J. N., Ball, L. J., & Ellis, D. (2013). Matching bias in syllogistic reasoning: Evidence for a dual-process account from response times and confidence ratings. *Thinking & Reasoning*, *19*(1), 54–77. https://doi.org/10.1080/13546783.2012.735622

Stupple, E. J., Gale, M., & Richmond, C. R. (2013). Working memory, cognitive miserliness and logic as predictors of performance on the cognitive reflection test. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual conference of the cognitive science society* (pp. 1396–1401). Austin, TX: Cognitive Science Society.

Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning*, *20*(2), 215–244. https://doi.org/10.1080/13546783.2013.869763

Thompson, V. A., Turner, J. A. P., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive psychology, 63*(3), 107–140. https://doi.org/10.1016/j.cogpsych.2011.06.001

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning*, *20*(2), 147–168. https://doi.org/10.1080/13546783.2013.844729

Toplak, M. E., West, R. F., & Stanovich, K. E. (2017). Real-World Correlates of Performance on Heuristics and Biases Tasks in a Community Sample: Heuristics and Biases Tasks and Outcomes. *Journal of Behavioral Decision Making*, *30*(2), 541–554. https://doi.org/10.1002/bdm.1973

Trippas, D., & Handley, S. (2017). The parallel processing model of belief bias: Review and extensions. In W. De Neys (Ed.), *Dual process theory 2.0* (pp. 28–46). Oxon, UK: Routledge.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–

315. https://doi.org/10.1037/0033-295X.90.4.293

Vega, S., Mata, A., Ferreira, M. B., & Vaz, A. R. (2021). Metacognition in moral decisions: Judgment extremity and feeling of rightness in moral intuitions. *Thinking & Reasoning*, *27*(1), 124–141. https://doi.org/10.1080/13546783.2020.1741448

West, R. F., Toplak, M. E., & Stanovich, K. E. (2008). Heuristics and biases as measures of critical thinking: Associations with cognitive ability and thinking dispositions. *Journal of Educational Psychology*, *100*(4), 930–941. https://doi.org/10.1037/a0012842

Wickham, H. (2016*). ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H. (2021). tidyr: Tidy Messy Data. R package version 1.1.3. https://CRAN.R-project.org/package=tidyr

Wickham, H., François, R., Henry, L., & Müller, K. (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.6. https://CRAN.R-project.org/package=dplyr

General Discussion

Summary of findings

The current thesis aimed at investigating the nature of sound responding and error sensitivity across a range of fields beyond heuristics-and-biases tasks. This work builds upon recent studies in the field of logical reasoning, the evidence of which compels us to revise the traditional views on intuitive and deliberate thinking (Bago & De Neys, 2017; De Neys & Pennycook, 2019; Handley et al., 2011; Newman et al., 2017; Pennycook et al., 2015; see De Neys, 2017, for a review). This line of research has demonstrated that the response that has traditionally been assumed to be cued after deliberation is frequently generated intuitively (De Neys, 2022). Furthermore, it has also shown that even when people remain biased, they typically show sensitivity to their errors (Białek & De Neys, 2016; De Neys, 2017; Frey et al., 2018; Gangemi et al., 2015; Mata, 2020; Pennycook et al., 2015) and this error sensitivity can often operate intuitively (Bago & De Neys, 2017; Białek & Neys, 2017; Burič & Konrádová, 2021; Burič & Šrol, 2020; Thompson & Johnson, 2014; Trippas et al., 2016). In sum, these studies propose that alongside the traditional "slow" deliberate route, there also exists a "fast" intuitive route for correct responding and error detection.

In the main Axis 1 (Chapters 1-3) of this thesis I explored the generalizability of these findings beyond logical reasoning. To cover a range of fields where correct responding is commonly believed to be a result of deliberate correction, I focused on risky choice, high-level semantic tasks, and low-level cognitive control tasks. Across these domains, I found that responses that were traditionally thought to necessitate slow, effortful deliberation were often generated intuitively and that even when people remained biased, they showed an intuitive sensitivity to their errors.

First, in risky decision making, people frequently selected the alleged deliberate choice (i.e., the expected value maximizing choice) after mere intuitive processing. Also, regardless of whether people opted for an expected value or a loss averse choice, they often showed intuitive sensitivity to the alternative choice as well. In other words, even when they were loss averse, reasoners processed

the expected value principles on some level. Second, in high-level semantic tasks, avoiding semantic illusions typically required deliberation, but participants still provided correct intuitive answers in a significant number of trials. So, although less prevalent than in the reasoning domain, sound intuiting served as an alternative route to accurate responding. Additionally, when people succumbed to the illusion and provided an incorrect response, they often showed intuitive sensitivity to the erroneous nature of their answer. Third, in low-level cognitive control tasks, in the majority of cases participants managed to provide correct responses even when their deliberate control was constrained.

In supplementary Axis 2 (Chapter 4) of this thesis I tested the stability of biases in logical reasoning and the impact of conflict detection on long-term answer change. The main finding was that people's intuitive and deliberate responses to classic heuristics-and-biases tasks were highly stable over time. However, despite the high stability, there was still some variability in intuitive and deliberate responses after two weeks, and this variability could be explained by conflict sensitivity. More specifically, the more conflicted people felt about their responses when solving a problem, the more likely they were to change their response to this problem over time.

In sum, in Axis 1 (Chapters 1-3), although there was some variation across tasks, results demonstrated a common trend: responses that were traditionally believed to result from effortful deliberation often arose from intuitive processing, in both high level and low level tasks, as well as in decisions under risk. Furthermore, Chapters 1 and 2 showed evidence for automatic conflict detection within these tasks. Chapter 4 (Axis 2), showed that intuitive conflict sensitivity predicted answer change in the long-term. These findings suggest a broad generalization of the two response findings across different domains and timeframes and prompt dual process theorists to reevaluate popular "fast-and-slow" dual process models of cognition, particularly regarding the nature of the fast and intuitive System 1 (De Neys, 2022; De Neys & Pennycook, 2019).

Overall, the results of this thesis help to better understand the nature of intuitive-automatic responses beyond heuristics-and-biases tasks and suggest that, across various domains, decision making can be better understood as an interplay between various "fast" intuitions, rather than a dichotomy between "fast" and "slow" thinking per se.

General implications

Correct intuiting

Although in this thesis correct intuitive responding was always present, the extend of sound intuiting varied across the domains I examined. In low-level cognitive control tasks correct intuiting was very common. In risky choice tasks, while slightly less common, it still remained the dominant pattern of correct responding. Finally, in high-level semantic illusions, sound intuiting was less common and deliberate correction was the predominant route to correct responding. These variations highlight the importance of investigating a range of domains to gain a comprehensive understanding of the role of intuition in human cognition. Below I briefly discuss the potential reasons underlying these differences.

To begin with, the particularity of risky choice tasks, similar to logical reasoning tasks, is that they can be solved using a universal algorithm (De Neys, 2012). For instance, to determine whether a gamble has a positive expected value, one can solve the equation "Expected value = Outcome1 * Probability1 + Outcome2 * Probability2". Since this solution strategy can be applied to all variations of the problem, it can be easily automatized. In fact, in the heuristics-and-biases literature, it is believed that this automatization drives correct intuitive responding (De Neys, 2012; Purcell et al., 2021; Raoelison et al., 2020; Stanovich, 2018). More specifically, most adults have typically been exposed to core logical principles and have extensively practiced them within the school curriculum (De Neys, 2012; De Neys & Pennycook, 2019; Evans, 2019; Raoelison et al., 2020, 2021; Stanovich, 2018). This repeated exposure enables good reasoners to automatize their application. This hypothesis fits risky decision making, since probabilistic and expected value principles are often taught throughout formal education.

On the contrary, in the case of semantic illusions, there is no general algorithm one could apply. Instead, one must carefully search their semantic memory and this search will be unique for each problem. For instance, consider the following two semantic illusions: "What country was Margaret Thatcher president of for several years?" and "In the tale, who found the glass slipper left at the ball by Snow White?". The semantic search for each of these illusions differs

significantly, since the first example requires factual knowledge about Margaret Thatcher's political role, while the second one requires familiarity with the Cinderella fairy tale. Thus, because of their uniqueness, semantic search strategies are less automatized. For instance, someone well-versed in politics would automatically identify that Margaret Thatcher was not a president, but might not be very familiar with the story of Cinderella. This can explain the lower prevalence of correct intuiting and the higher prevalence of deliberate correction in these tasks.

Finally, in lower-level cognitive control tasks, correct intuitive responding was very common compared to the domains I explored in the other chapters. Unlike risky choice tasks and semantic illusions, these tasks explicitly instruct participants what the correct response is and are less related to the reasoner's mathematical or factual knowledge. The correct response is based on a new and seemingly trivial instruction that participants have not encountered or practiced before. The difficulty rather lies in that participants need to suppress conflicting information (e.g., the written word in the Stroop task or the direction of the flanker arrows in the Flanker task). So, even though the correct answer is available to them, they need to consistently recruit cognitive control to suppress the competing response.

This could explain why correct intuitive responses are significantly more common than deliberate correction in cognitive control tasks. In these tasks, both the correct and incorrect cues are strong, and if participants successfully assert automatic control they will always suppress the incorrect cue and respond correctly. The underlying mechanism for risky choice and semantic illusion tasks is different. In these tasks, the intuition (correct or heuristic) with the highest activation level will be selected as the intuitive response (De Neys, 2022; Stanovich, 2018). The relative difference between these intuitions will determine the experienced conflict (i.e., the smaller the difference, the highest the experienced conflict, De Neys, 2022). Let's consider the case when a reasoner provides a heuristic intuitive response. If the "heuristic intuition" clearly dominates over the "logical" one, there will be little or no experienced conflict, and the reasoner will not engage in deliberation (De Neys, 2022). However, if the "heuristic intuition" is similar in strength with the "logical" one, the resulting conflict will be high and will prompt deliberation (De Neys, 2012; Pennycook et al., 2015; Thompson et al., 2011). This deliberation, in turn, may successfully suppress the

"heuristic intuition" and correct the erroneous intuitive response (i.e., "01" case). This may explain why deliberate correction is more frequent in tasks that cue strong heuristics and do not explicitly mention the correct response in the instructions.

In sum, in low-level cognitive control tasks, knowing the correct answer and being able to recruit control even under cognitive constraints makes intuiting a more common path of correct responding than deliberate correction. In risky choice tasks the correct response is not explicitly instructed, but the problems can be solved using a universal algorithm. Therefore, reasoners that have automatized this algorithm can intuitively apply it to all variations of the problem (De Neys, 2012; De Neys & Pennycook, 2019; Raoelison et al., 2020). Finally, in semantic illusions participants are not explicitly instructed about the correct response and, in addition, each problem is unique. Therefore, even if reasoners have automatized the semantic search for one illusion, an illusion with a different semantic content might require deliberation. This may make deliberate correction more common than correct intuiting. Although correct intuitive responding is found to different extends across domains, the main conclusion is that intuiting always remains a viable route to correct responding.

Conflict sensitivity

In addition to correct intuitive responding, further support for the intuitive processing of the alleged deliberate response across domains comes from conflict detection findings. Both in risky choice tasks and semantic illusions I observed that participants were sensitive to their errors, as they reported lower response confidence in conflict trials compared to their baseline confidence in control, no-conflict trials.[1] While in semantic illusions the effect of conflict detection was clear and in line with what is typically observed in heuristics-and-biases tasks (e.g., Bago & De Neys, 2017, 2020), in risky decision making the effect was smaller and less consistent. This may be explained by the relatively low baseline initial no-

[1] In low-level cognitive control tasks I decided to not include a confidence question as it would have made the two-response design overly complex given the task's tight millisecond-level deadline and high number of presented trials. However, in the cognitive control field, there is already evidence for error-related brain activity, often referred to as the error-related negativity (ERN; Falkenstein et al., 1991). This ERN has also been found to be related to a subjective error awareness in cognitive control tasks (e.g., Scheffers & Coles, 2000; Wessel, 2012; but see also Nieuwenhuis et al., 2001, for a null relationship).

conflict confidence in risky choice tasks (on average around 70%) compared to other tasks (around 90% in semantic illusions). This discrepancy might suggest that the no-conflict items we constructed for risky choice tasks were not entirely devoid of conflict and may have, at times, induced conflicts between different intuitions. For instance, in one of our no-conflict risky choice items, participants could adopt the strategy of consistently selecting the lottery with the highest value within the set. Such a strategy would lead them towards a choice that contradicted the correct option. While we did control for this on a general level, as we ran a control analysis excluding participants who often followed this strategy, we cannot completely rule out its potential impact on confidence levels, or the potential effect of other strategies in generating conflicts. Moving forward, it is crucial that for a precise measurement of conflict detection, the no-conflict items genuinely avoid inducing any conflict. Admittedly, ensuring this in risky choice tasks might be more challenging than in other domains.

The conflict detection findings also fit well into the framework of competing intuitions proposed by recent models in dual process theorizing (e.g., Bago & De Neys, 2020; De Neys & Pennycook, 2019; Pennycook et al., 2015). As mentioned in the Introduction, these models suggest that intuitive reasoning is characterized by the interplay between two main types of intuitions: "logical intuitions", which cue responses that are in line with logico-mathematical principles, and "heuristic intuitions", which cue responses that conflict with logico-mathematical principles (De Neys, 2012, 2022; Evans, 2019; Stanovich, 2018). The intuitive response that is eventually selected by the reasoner is determined by the strongest, most activated intuition. If one intuition clearly dominates in strength over the other, it will prevail with little or no experienced conflict. However, if the activation levels of the two intuitions are similar in strength, the experienced conflict will be high and participants will be more likely to engage in deliberation and change their response (Bago & De Neys, 2019a; De Neys & Pennycook, 2019; Pennycook et al., 2015; Trippas & Handley, 2018). Two of the conflict detection findings of this thesis are in line with the competing intuitions framework.

The first finding concerns the relationship between item difficulty and experienced conflict, which has also been previously shown in the reasoning field (e.g., Bago & De Neys, 2020; Pennycook et al., 2015). To clarify, in both semantic illusions (Chapter 2) and risky choice tasks (Chapter 3, Study 3), conflict items

had two levels of difficulty: "easy" and "hard".[2] In both easy and hard items conflict was created between the correct and heuristic cues. However, in easy items the correct cue was more obvious, making it easier for participants to detect the conflict.

Let's first consider the case where people provide a heuristic response (i.e., the "heuristic intuition" dominates). In this case, the "logical intuition" is also activated to some extent, as indicated by the intuitive conflict detection findings. However, the strength of this intuition differs between easy and hard items. In easy items, the "logical intuition" is stronger, so reasoners are expected to experience conflict and have low confidence in their heuristic response. In hard items, the "logical intuition" is weaker, so people are not expected to experience high conflict in their heuristic answers (see Bago & De Neys, 2020 for a similar observation in the reasoning field).[3]

Now, let's consider the case where people provide a correct response (i.e., the "logical intuition" dominates). Here, the competing intuitions account would make the opposite prediction regarding item difficulty and confidence. When the item is hard, although the "logical intuition" dominates, it is not very strongly activated. As a result, people are expected to feel conflicted about their correct answers. Conversely, in easy items, the "logical intuition" is much more strongly activated than the "heuristic" one, so people are expected to have a high response confidence in their correct response.

In sum, in the case of a heuristic intuitive response, the competing intuitions account predicts that people will have lower confidence in easy items, as opposed to hard ones. In the case of a correct intuitive response, it predicts that they will have lower confidence in hard items, compared to easy ones (Bago & De Neys, 2020; De Neys & Pennycook, 2019). In Chapters 1 and 2 there was general evidence supporting both of these patterns, although in Chapter 1 the trend was purely observational and not statistically tested. This suggests that both in risky decisions and semantic illusions, confidence and performance in the initial intuitive

[2] In Chapter 2 the difficulty levels are referred to as "strong impostor" (i.e., hard items) and "weak impostor" (i.e., easy items), respectively. This is because a strong impostor word makes the illusion harder to spot (e.g., "What country was Margaret Thatcher *president* of for several years?"), while a weak impostor word makes it easier to spot (e.g., "What country was Margaret Thatcher *queen* of for several years?").

[3] The "heuristic intuition" is assumed to remain the same in both easy and hard items as the heuristic cue does not change.

stage depend on the strength of the alleged "logical intuition" (as operationalized by item difficulty).

The second finding consistent with the competing intuitions account, relates to conflict detection and its impact on answer change (Chapter 4). Conflict detection is often viewed as a mechanism that prompts people to change their responses after deliberation (De Neys, 2012; Pennycook et al., 2015; Purcell et al., 2023; Thompson et al., 2011; Trippas & Handley, 2018). In Chapter 4, I showed that this effect persists over the long-term: participants who experience greater intuitive conflict about a problem are more likely to change their response to that problem two weeks later. Critically, intuitive conflict detection had a long-term effect not only on deliberate, but also on intuitive responses. This effect could be potentially explained by the change in the strength order of two competing intuitions which show little differentiation to begin with. For instance, if the difference in strength between a "heuristic" and a "logical intuition" is small (i.e., intuitive conflict detection is high), random noise over time may easily reverse the strength hierarchy (i.e., the initially "weaker" intuition will now be slightly stronger than its competitor and be selected). Put differently, the more distinct one's preference for a particular intuition (be it heuristic or logical) over another, the less conflict one will experience, and the more likely one will consistently make the same choice over the long-term.

General reflections and future directions

Role of deliberation

Given that this thesis showed that correct responses are typically provided intuitively across domains, one might wonder what is the role of deliberation in reasoning. To begin with, although the findings show that deliberation is often not necessary for correct responding, there are still some trials where people engaged in deliberation to correct an erroneous intuitive answer (i.e., "01" cases). Critically, in high-level semantic tasks, deliberate correction was the modal response pattern and was more frequent than correct intuiting. This suggests that in tasks where automatizing a generic strategy for correct responding might be complicated, deliberation plays a more corrective role. Similarly, there are times when deliberation is necessary for reasoners to detect conflict in a task. Therefore, the findings do not imply that correct responses and conflict sensitivity are only

possible via a fast, intuitive route; rather they propose that a fast intuitive route operates along with a slower, deliberate one. The specific route a reasoner takes to arrive at a correct response depends, among other things, on the task's characteristics, the extent to which the correct strategy has been practiced in the past as well as the reasoner's cognitive capacities (Burič & Konrádová, 2021; Raoelison et al., 2020; Stanovich, 2018; Thompson, 2021; Thompson et al., 2018).

A promising area for future research is to explore the potential additional functions that deliberation may serve in reasoning. An idea that has recently gained empirical support within heuristics-and-biases tasks, posits that deliberation plays an important role in providing justifications and arguments for intuitive decisions (Bago & De Neys, 2019a; De Neys & Pennycook, 2019; Evans, 2019; Pennycook et al., 2015; Wason & Evans, 1974). Bago and De Neys (2019a) showed that when participants were presented with the bat-and-ball problem, they often generated correct responses intuitively, but could better justify their answers after deliberation. So, although intuitive thinking often leads to correct responses, people struggle to clearly explain the rationale behind these intuitive responses without deliberating. Therefore, a primary role of deliberation appears to be justification, which is critical for argumentation and persuasion in social contexts (Mercier & Sperber, 2011).

However, it is important to recognize that deliberation might serve different or additional roles in different decision domains (Evans, 2019). For instance, in lower-level cognitive control tasks deliberation is mainly acting as a response inhibitor, aiming to suppress competing information, while its role in justification is less important. Since in these tasks the correct answer is explicitly provided in the instructions, justifying it does not add much value (e.g., "the correct answer is "blue" because the ink is blue"). However, in tasks like the base-rate neglect, which involves stereotypes, or risky choice, where personal preferences and financial status play an important role, offering justifications can be important both for increasing confidence in one's response and for potentially persuading others of the response's validity (Pennycook et al., 2015). Empirically exploring the potential roles of deliberation across various fields can enrich dual process models and improve our understanding of deliberation in reasoning.

Finally, if we assume that "logical intuitions" arise from a process of learning and practice, deliberation will always be necessary in the early stages of learning

(e.g., before automatizing a logical rule or a semantic association) to make this knowledge effortlessly accessible (Shiffrin & Schneider, 1977). Hence, the prediction is that early in development deliberation will be crucial. However, as people become more acquainted with logical principles though the school curriculum or real-life experiences, they will be able to automatize these principles and eventually apply them intuitively (De Neys, 2022).

Development

Practicing correct rules to automaticity is not a new concept in cognitive psychology (LaBerge & Samuels, 1974; Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977), and it has even been integrated into traditional dual process theories (Evans & Stanovich, 2013; Kahneman & Klein, 2009). However, the automatization hypothesis plays a more central role in in the new, revised dual process models, as it suggests that "logical intuitions" may be accessible not only to experts, but also to the average, modal reasoner (De Neys & Pennycook, 2019). Despite the prevalence of automatization in recent dual-process models, there is limited empirical evidence to support it (see Purcell et al., 2021; Raoelison et al., 2021). In one of the few direct tests of this hypothesis, Raoelison et al. (2021) showed that older (12th grade) reasoners were more likely to deliberately correct an initial erroneous intuition (i.e., "01" cases) compared to their younger (7th grade) counterparts, when solving classic heuristics-and-biases tasks. Importantly, however, the older students were also far more likely to intuitively generate accurate responses (i.e., "11" cases; Raoelison et al., 2021). These results show that the improvement in reasoning accuracy with age can be partially attributed to improved intuitive accuracy. Hence, they support the automatization hypothesis by implying that deliberation plays a vital role during the initial stages of learning, but as students get more familiar with these logical principles through the school curriculum, they start automatizing them and can eventually apply them intuitively. Nevertheless, more empirical developmental research is necessary. Exploring a broader range of ages and encompassing a wider array of mathematical concepts, as well as non-mathematical concepts like semantic associations, could provide a comprehensive understanding of the origins of "logical intuitions".

Boosting sound intuiting

An effective way to better understand, but also enhance, sound reasoning is through interventions that target cognitive biases. Research has shown that brief, one-shot interventions which draw attention to biases or explain why a response is considered incorrect, often make reasoners provide more logical responses (Hoover & Healy, 2017; Morewedge et al., 2015; Trouche et al., 2014). Critically, recent studies using heuristics-and-biases tasks, have revealed that a brief intervention explaining the rationale behind the problem's solution did not only make deliberate responses, but also intuitive ones, more logical; after receiving the explanation, reasoners were able to provide more logical intuitive answers to structurally similar problems (Boissin et al., 2021, 2022). Therefore, rather than training people to deliberately correct erroneous intuitions, we can help them boost their "logical intuitions", so that they can use mere intuiting processing in their favor (Milkman et al., 2009).

In addition to the debiasing effect of such interventions, they can also allow us to study "logical intuitions" in a wider range of cases. More specifically, in tasks that cue strong heuristics, correct responses remain scarce even after deliberation. Therefore, when comparing correct intuitive responses (i.e., "11" cases) to responses that only become correct after deliberation (i.e., "01" cases), we focus on a minority of cases which might not be representative of the average reasoner. By providing a short "debiasing" explanation and observing whether it leads to more correct responses after intuitive or deliberate processing, we can test "logical intuitions" in a larger sample and include reasoners who may not have provided any correct responses to begin with, but manage to do so after a simple explanation.

For instance, in Chapter 1, I found that expected-value-maximizing choices in risky choice tasks were more often generated intuitively than after deliberation. However, even in the final deliberate stage, expected-value-based responses remained rare. One way to increase the focus group would be to provide a brief explanation of expected value and afterwards assess whether people become less loss averse, both intuitively and after deliberation. If, post-explanation, most reasoners can intuitively make expected-value-maximizing choices, this would imply that they possess "logical intuitions" about expected value.

The core idea here is that some biased reasoners possess overlearned, intuitive knowledge of the correct principles (e.g., expected-value maximization), but this knowledge is not as strongly activated as their heuristic cues. This situation allows them to detect conflict and recognize that their heuristic answer may not be fully warranted (Stanovich, 2018), but their response is still determined by the more strongly activated "heuristic intuition". In essence, these biased reasoners may have "logical intuitions", yet these intuitions are less potent than their "heuristic intuitions". A simple explanation could help highlight the relevance of the underlying principle in question, thereby increasing its strength and potentially allowing it to be selected intuitively as the response (Boissin et al., 2021).

Domain composites and interconnections

While both this thesis and prior studies have examined intuitive logic and conflict sensitivity across various domains (Bago et al., 2021; Bago & De Neys, 2019b; Kessler et al., 2017; Vega et al., 2021), there are two interesting potential avenues for improving these findings. Firstly, instead of examining specific tasks in isolation, it could be beneficial to construct composite indices for each domain of interest, like cognitive control, risky choice, or moral reasoning. Although this approach may introduce practical challenges (e.g., very lengthy experiments), it can provide us with a more robust and comprehensive view of correct intuiting and conflict sensitivity (see Gärtner & Strobel, 2021, for a similar point about inhibitory control studies). Secondly, exploring the interconnectedness of correct intuiting across fields can also shed light on the current findings. For instance, in Chapter 3, I examined the correlation of each individual's intuitive accuracy between heuristics-and-biases tasks and the Stroop task. The results revealed an, at best, weak association, which prompted hypotheses about the ways correct intuiting operates differently in these two tasks. By further exploring the similarities (or lack thereof) of correct intuiting and conflict sensitivity across fields, we could gain valuable insights into their underlying mechanisms (e.g., fields that allow an easy automatization of the correct solution strategy might be more related than those involving alternative mechanisms).

Applications

Finally, an important area for future research involves exploring and improving intuitive accuracy in applied contexts, which could eventually help tackle pressing societal challenges (Osman, 2023; see Pennycook, 2023 for an overview). Among the domains explored in this thesis, risky decision making could be particularly relevant for real-world applications. An interesting area to explore is how training intuitive and deliberate reasoning skills can improve investment decisions. For example, a brief intervention explaining fundamental probability principles could potentially have a positive impact on retirement savings or insurance plan choices (Banks & Oldfield, 2007; Clark et al., 2006). Another relevant area is that of medical prognosis and diagnosis (Hoffrage & Gigerenzer, 1998). Training both physicians and patients to better interpret probabilities and natural frequencies can substantially improve the understanding of disease risks (Galesic et al., 2009; Hoffrage & Gigerenzer, 1998). Moreover, determining whether people can hold "logical intuitions" about these concepts can lead to the development of more effective interventions and policies. In general, irrespective of the applied context, the main principle remains the same: equipping people with the tools to make decisions in line with logical principles when it is advantageous to do so and testing whether these decision-making processes can be automatized.

Concluding remarks

This thesis has demonstrated that intuitive processing constitutes a reliable route to correct responding across a wide range of fields. Moreover, people exhibit an intuitive sensitivity to their errors and this sensitivity serves as a predictor for long-term response change. While the extent of correct intuitive responding and conflict sensitivity varies according to task-specific characteristics, a consistent body of evidence validates their presence. These findings align with previous studies in the field of logical reasoning and show that in domains involving risky choices, semantic associations, and cognitive control, responses traditionally believed to result from deliberation can also stem from mere intuitive processing. This suggest that deliberation's primary role is not necessarily corrective in nature and that intuition plays a more substantial role in sound thinking than previously thought.

References

Bago, B., Bonnefon, J.-F., & De Neys, W. (2021). Intuition rather than deliberation determines selfish and prosocial choices. *Journal of Experimental Psychology: General*, *150*(6), 1081–1094. https://doi.org/10.1037/xge0000968

Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition*, *158*, 90–109. https://doi.org/10.1016/j.cognition.2016.10.014

Bago, B., & De Neys, W. (2019a). The Smart System 1: Evidence for the intuitive nature of correct responding on the bat-and-ball problem. *Thinking & Reasoning*, *25*(3), 257–299. https://doi.org/10.1080/13546783.2018.1507949

Bago, B., & De Neys, W. (2019b). The intuitive greater good: Testing the corrective dual process model of moral cognition. *Journal of Experimental Psychology: General*, *148*(10), 1782–1801. https://doi.org/10.1037/xge0000533

Bago, B., & De Neys, W. (2020). Advancing the specification of dual process models of higher cognition: A critical test of the hybrid model view. *Thinking & Reasoning*, *26*(1), 1–30. https://doi.org/10.1080/13546783.2018.1552194

Banks, J., & Oldfield, Z. (2007). Understanding Pensions: Cognitive Function, Numerical Ability and Retirement Saving*. *Fiscal Studies*, *28*(2), 143–170. https://doi.org/10.1111/j.1475-5890.2007.00052.x

Białek, M., & De Neys, W. (2016). Conflict detection during moral decision-making: Evidence for deontic reasoners' utilitarian sensitivity. *Journal of Cognitive Psychology*, *28*(5), 631–639. https://doi.org/10.1080/20445911.2016.1156118

Białek, M., & Neys, W. D. (2017). Dual processes and moral conflict: Evidence for deontological reasoners' intuitive utilitarian sensitivity. *Judgment and Decision Making*, *12*(2), 148–167. https://doi.org/10.1017/S1930297500005696

Boissin, E., Caparos, S., Raoelison, M., & De Neys, W. (2021). From bias to sound intuiting: Boosting correct intuitive reasoning. *Cognition*, *211*, 104645. https://doi.org/10.1016/j.cognition.2021.104645

Boissin, E., Caparos, S., Voudouri, A., & De Neys, W. (2022). Debiasing System 1: Training favours logical over stereotypical intuiting. *Judgment and Decision Making*, *17*(4), 646–690. https://doi.org/10.1017/S1930297500008895

Burič, R., & Konrádová, ľubica. (2021). Mindware Instantiation as a Predictor of Logical Intuitions in Cognitive Reflection Test. *Studia Psychologica*, *63*(2), 114–128. https://doi.org/10.31577/sp.2021.02.822

Burič, R., & Šrol, J. (2020). Individual differences in logical intuitions on reasoning problems presented under two-response paradigm. *Journal of Cognitive Psychology*, *32*(4), 460–477. https://doi.org/10.1080/20445911.2020.1766472

Clark, R. L., d'AMBROSIO, M. B., McDERMED, A. A., & Sawant, K. (2006). Retirement plans and saving decisions: The role of information and education. *Journal of Pension Economics & Finance*, *5*(1), 45–67. https://doi.org/10.1017/S1474747205002271

De Neys, W. (2012). Bias and Conflict: A Case for Logical Intuitions. *Perspectives on Psychological Science*, *7*(1), 28–38. https://doi.org/10.1177/1745691611429354

De Neys, W. (2017). Bias, conflict, and fast Logic. In W. De Neys (Ed.), *Dual Process Theory 2.0* (pp. 47-65). Routledge. https://doi.org/10.4324/9781315204550-4

De Neys, W. (2022). Advancing theorizing about fast-and-slow thinking. *Behavioral and Brain Sciences*, 1–68. https://doi.org/10.1017/S0140525X2200142X

De Neys, W., & Pennycook, G. (2019). Logic, Fast and Slow: Advances in Dual-Process Theorizing. *Current Directions in Psychological Science*, *28*(5), 503–509. https://doi.org/10.1177/0963721419855658

Evans, J. St. B. T. (2019). Reflections on reflection: The nature and function of type 2 processes in dual-process theories of reasoning. *Thinking & Reasoning*, *25*(4), 383–415. https://doi.org/10.1080/13546783.2019.1623071

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Falkenstein, M., Hohnsbein, J., Hoormann, J., & Blanke, L. (1991). Effects of crossmodal divided attention on late ERP components. II. Error processing in choice reaction tasks. *Electroencephalography and Clinical Neurophysiology*, *78*(6), 447–455. https://doi.org/10.1016/0013-4694(91)90062-9

Frey, D., Johnson, E. D., & De Neys, W. (2018). Individual differences in conflict detection during reasoning. *Quarterly Journal of Experimental Psychology*, *71*(5), 1188–1208. https://doi.org/10.1080/17470218.2017.1313283

Galesic, M., Garcia-Retamero, R., & Gigerenzer, G. (2009). Using icon arrays to communicate medical risks: Overcoming low numeracy. *Health Psychology*, *28*(2), 210–216. https://doi.org/10.1037/a0014474

Gangemi, A., Bourgeois-Gironde, S., & Mancini, F. (2015). Feelings of error in reasoning—In search of a phenomenon. *Thinking & Reasoning*, *21*(4), 383–396. https://doi.org/10.1080/13546783.2014.980755

Gärtner, A., & Strobel, A. (2021). Individual Differences in Inhibitory Control: A latent Variable Analysis. *Journal of Cognition*, *4*(1), 17. https://doi.org/10.5334/joc.150

Handley, S. J., Newstead, S. E., & Trippas, D. (2011). Logic, beliefs, and instruction: A test of the default interventionist account of belief bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(1), 28–43. https://doi.org/10.1037/a0021098

Hoffrage, U., & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine*, *73*(5), 538.

Hoover, J. D., & Healy, A. F. (2017). Algebraic reasoning and bat-and-ball problem variants: Solving isomorphic algebra first facilitates problem solving later. *Psychonomic Bulletin & Review*, *24*(6), 1922–1928. https://doi.org/10.3758/s13423-017-1241-8

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, *64*(6), 515–526. https://doi.org/10.1037/a0016755

Kessler, J., Kivimaki, H., & Niederle, M. (2017). Thinking fast and slow: generosity over time. Retrieved https://users.nber.org/~kesslerj/papers/KesslerKivimakiNiederle_GenerosityOverTime.pdf

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, *6*(2), 293–323. https://doi.org/10.1016/0010-0285(74)90015-2

Mata, A. (2020). Conflict detection and social perception: Bringing meta-reasoning and social cognition together. *Thinking & Reasoning*, *26*(1), 140–149. https://doi.org/10.1080/13546783.2019.1611664

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57–74. https://doi.org/10.1017/S0140525X10000968

Milkman, K. L., Chugh, D., & Bazerman, M. H. (2009). How Can Decision Making Be Improved? *Perspectives on Psychological Science*, *4*(4), 379–383. https://doi.org/10.1111/j.1745-6924.2009.01142.x

Morewedge, C. K., Yoon, H., Scopelliti, I., Symborski, C. W., Korris, J. H., & Kassam, K. S. (2015). Debiasing Decisions: Improved Decision Making With a Single Training Intervention. *Policy Insights from the Behavioral and Brain Sciences*, *2*(1), 129–140. https://doi.org/10.1177/2372732215600886

Newman, I. R., Gibb, M., & Thompson, V. A. (2017). Rule-based reasoning is fast and belief-based reasoning can be slow: Challenging current explanations of belief-bias and base-rate neglect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(7), 1154–1170. https://doi.org/10.1037/xlm0000372

Nieuwenhuis, S., Ridderinkhof, K. R., Blom, J., Band, G. P. H., & Kok, A. (2001). Error-related brain potentials are differentially related to awareness of response errors: Evidence from an antisaccade task. *Psychophysiology*, *38*(5), 752–760. https://doi.org/10.1111/1469-8986.3850752

Osman, M. (2023). Using the study of reasoning to address the age of unreason. *Behavioral and Brain Sciences*, *46*, e135. https://doi.org/10.1017/S0140525X22002953

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. https://doi.org/10.1016/j.cogpsych.2015.05.001

Purcell, Z. A., Wastell, C. A., & Sweller, N. (2021). Domain-specific experience and dual-process thinking. *Thinking & Reasoning*, *27*(2), 239–267. https://doi.org/10.1080/13546783.2020.1793813

Purcell, Z. A., Wastell, C. A., & Sweller, N. (2023). Eye movements reveal that low confidence precedes deliberation. *Quarterly Journal of Experimental Psychology*, *76*(7), 1539–1546. https://doi.org/10.1177/17470218221126505

Raoelison, M., Boissin, E., Borst, G., & De Neys, W. (2021). From slow to fast logic: The development of logical intuitions. *Thinking & Reasoning*, *27*(4), 599–622. https://doi.org/10.1080/13546783.2021.1885488

Raoelison, M., Thompson, V. A., & De Neys, W. (2020). The smart intuitor: Cognitive capacity predicts intuitive rather than deliberate thinking. *Cognition*, *204*, 104381. https://doi.org/10.1016/j.cognition.2020.104381

Scheffers, M. K., & Coles, M. G. H. (2000). Performance monitoring in a confusing world: Error-related brain activity, judgments of response accuracy, and types of errors. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(1), 141–151. https://doi.org/10.1037/0096-1523.26.1.141

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, *84*(1), 1–66. https://doi.org/10.1037/0033-295X.84.1.1

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, *84*(2), 127–190. https://doi.org/10.1037/0033-295X.84.2.127

Stanovich, K. E. (2018). Miserliness in human cognition: The interaction of detection, override and mindware. *Thinking & Reasoning*, *24*(4), 423–444. https://doi.org/10.1080/13546783.2018.1459314

Thompson, V. A. (2021). Eye-tracking IQ: Cognitive capacity and strategy use on a ratio-bias task. *Cognition*, *208*, 104523. https://doi.org/10.1016/j.cognition.2020.104523

Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning*, *20*(2), 215–244. https://doi.org/10.1080/13546783.2013.869763

Thompson, V. A., Pennycook, G., Trippas, D., & Evans, J. St. B. T. (2018). Do smart people have better intuitions? *Journal of Experimental Psychology: General*, *147*(7), 945–961. https://doi.org/10.1037/xge0000457

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*(3), 107–140. https://doi.org/10.1016/j.cogpsych.2011.06.001

Trippas, D., & Handley, S. J. (2018). The parallel processing model of belief bias: Review and extensions. In *Dual process theory 2.0.* (pp. 28–46). Routledge/Taylor & Francis Group.

Trippas, D., Handley, S. J., Verde, M. F., & Morsanyi, K. (2016). Logic brightens my day: Evidence for implicit sensitivity to logical validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(9), 1448–1457. https://doi.org/10.1037/xlm0000248

Trouche, E., Sander, E., & Mercier, H. (2014). Arguments, more than confidence, explain the good performance of reasoning groups. *Journal of Experimental Psychology: General*, *143*(5), 1958–1971. https://doi.org/10.1037/a0037099

Vega, S., Mata, A., Ferreira, M. B., & Vaz, A. R. (2021). Metacognition in moral decisions: Judgment extremity and feeling of rightness in moral intuitions. *Thinking & Reasoning*, *27*(1), 124–141. https://doi.org/10.1080/13546783.2020.1741448

Wason, P. C., & Evans, J. ST. B. T. (1974). Dual processes in reasoning? *Cognition*, *3*(2), 141–154. https://doi.org/10.1016/0010-0277(74)90017-1

Wessel, J. R. (2012). Error awareness and the error-related negativity: Evaluating the first decade of evidence. *Frontiers in Human Neuroscience*, *6*, 88. https://doi.org/10.3389/fnhum.2012.00088

Supplementary Material

Supplementary material for Chapter 1

A. Instructions

Study 1

Please read these instructions carefully!

This experiment is divided into two parts: a betting game and a lottery game.

In the **betting game** you will be asked to choose either **to take** or **to not take bets**.

In the **lottery game** you will be asked to **choose one out of two lotteries**.

In each game you will have to answer 15 multiple-choice questions and a couple of practice questions. The questions will be presented to you one after the other and **you should not pause between them.**

After you finish the first game, you can take **a short break.**

It is important that you complete the experiment in one sitting and without distractions.

Click on **Next** to continue.

Participants were then either first presented with the instructions for the betting game or with the instructions of the lottery game. For the betting game the instructions were the following:

Welcome to the betting game!
In this game you will be presented with **a different bet on every trial**.
You will be given **both probabilities of winning and losing a certain amount of money**.
Based on these probabilities, you can decide **whether you want to take the bet or not**.

Imagine you are playing for real money and want to make **as much profit as possible.**
Imagine that any money you win, you get to keep and any money you lose you need to pay for.
We are interested in whether or not you would want to take the bet if you were to play for real.
See the example below:

If you take this bet you have:
70%** probability to **WIN €100
30%** probability to **LOSE €15
Do you take the bet?

- ***YES***
- ***NO***

In the actual game, you can choose whether you want to take the bet or not **by clicking on one of the answer options**.

Critically, in this game we want to know what your **initial, intuitive response** to the bets is and **how you respond after you have thought about these bets for some more time.**
First, we want you to respond with the **very first answer that comes to mind**. You don't need to think about it. Just give the first answer that intuitively comes to mind as quickly as possible.
To make sure that you answer as fast as possible, **a time limit was set for the first response**, which is going to be **4.5 seconds**. When there is 1 second left, the background colour will turn to yellow to let you know that the deadline is approaching. Please make sure to **answer before the deadline passes**.
Next, **the bet will be presented again** and you can take all the time you want to actively reflect on it. Once you have made up your mind you give your **final response**.
After you made your choice and clicked on it, you will be automatically taken to the next page.
After you have entered your first and final answer we will also ask you to indicate how confident you are that you made the right decision.
We are going to clarify all of this with a couple of practice questions.

We are going to start with **two practice questions** to familiarise you with the game.

For each question, a fixation cross will appear first. Then, the bet will be presented.
You then enter your first hunch as fast as possible before the deadline.
Next, the bet will be presented again and you can take all the time to reflect on it and enter your final response.
After you have entered your initial and your final answer we will also ask you to indicate how confident you are that you made the right decision.

We will let you practice now.

Click on **Next** when you are ready to start the practice session.

Participants were then given two practice trials, without the concurrent load task. They were then introduced to the load task.

You will also need to **memorize a pattern** while you respond to the bets.

You will see a grid with crosses and you will have to memorize their location.
You will first practice with 2 patterns without a bet.
The pattern will be displayed for **2 seconds** and then you will have to select it among **4 different patterns**.
Click on **Next** to begin.

Participants were then given two practice trials for the cognitive load task, without the bets. They were then presented with the following instructions:

In the actual study you will need to memorize the pattern while you respond to the bet. The pattern is briefly presented before each bet.
The difficulty of the pattern might vary. Always try to memorize as many crosses as possible. Each cross counts!
We know that it is not always easy to memorize the pattern while you are also thinking about the bet. The most important thing is to correctly memorize the pattern.
First, **try to concentrate on the memorization task**, and then try to respond to the bet.
As a next step, you can practice this with two questions.
Click on **Next** to proceed.

After those two last practice trials, participants were presented with the following instructions:

> This is the end of all practice!
> **Remember:**
> In this game you have to answer to 15 questions.
> The questions will be presented to you one after the other and **you should not pause between them.**
> Click on **Next** when you are ready to start with the actual game.

After the first game participants were presented with the instructions for the second game (either the lottery or the betting game). The instructions for the lottery game were the following:

> **Welcome to the lottery game!**
> In this game you will be presented with **two different lotteries** on every trial: Lottery A and Lottery B.
> No matter which lottery you choose, **you can always win some money**.
> For each lottery you will be given **the probabilities of winning**.
> Based on these probabilities, you can decide which lottery to choose.
> Imagine you are playing for real money and want to make **as much profit as possible.**
> Imagine that any money you win, you get to keep.
> Which lottery would you choose if you were to play for real?
> See the example below:

Lottery A	***Lottery B***
80% *probability to* ***win €55***	***80%*** *probability to* ***win €180***
20% *probability to* ***win €380***	***20%*** *probability to* ***win €300***

> *Which lottery do you choose?*
>
> - ***A***
> - ***B***
>
> In the above example, the probabilities that are under Lottery A, correspond to Lottery A.
> The probabilities that are under Lottery B, correspond to Lottery B.
>
> In the actual game, you can choose the lottery you want to take **by clicking on one of the answer options**.

Then participants saw the same instructions, considering the practice trials as in the first game. The instructions were identical with the first game, apart from the deadline which was adjusted accordingly.

Study 2-3

In the general instructions of Studies 2-3 the renumeration information was added. The instructions were the following:

> **Please read these instructions carefully!**
>
> This experiment is divided into two parts: a betting game and a lottery game.
>
> In the **betting game** you will be asked to choose either **to take** or **to not take bets**.
>
> In the **lottery game** you will be asked to **choose one out of two lotteries**.
>
> **In each game** you will have to answer 20 multiple-choice questions and a couple of practice questions.
> After you finish the experiment we will **randomly** select **one bet** from the betting game and **one lottery** from the lottery game.
> Then we will play both this bet and this lottery in our software according to the responses you gave.
>
> Any money you make or lose in this bet and this lottery will be added together.
> Then, your payment will be multiplied by a factor of 0.0013.
> **This means that you can make from 1 pound to 0 pounds extra (depending on your choices)** in addition to your standard payment.
>
> Your goal in both games is to make **as much profit as possible.**
> The questions will be presented to you one after the other and **you should not pause between them.**
>
> After you finish the first game, you can take **a short break.**
>
> **It is important that you complete the experiment in one sitting and without distractions.**
> Click on **Next** to continue.

The rest of the instructions were identical to Study 1. However, participants were simply told "**Remember:** your goal is to make as much profit as possible."

Instead of being told "Imagine you are playing for real money and want to make **as much profit as possible.** Imagine that any money you win, you get to keep."

B. Items

Betting game

All betting game items had the following structure:

> *If you take this bet you have:*
> ***_%** probability to **WIN €_***
> ***_%** probability to **LOSE €_***
> *Do you take the bet?*
> - *Yes*
> - *No*

Table S1 below shows the win and lose sentences of each of the items. Note that Study 1 did not include counterbalancing (all participants viewed the same conflict, no-conflict and filler items). In Studies 2-3 we introduced counterbalancing. To make counterbalancing work in these studies, we added a set B of conflict and filler items and a new set A and B of no-conflict items. In Study 3 we also added the easy-conflict items.

Table S1. The betting game items according to the study (1, 2, 3) they were used in, the type (Conflict, No-Conflict, Filler), the Difficulty level of the conflict items (Hard; Easy), the Items' Number, and the counterbalancing set (A; B).

Study	Type	Difficulty	Number	Set	Win sentence	Lose sentence
1, 2, 3	C	Hard	1	A	5% probability to WIN 110€	95% probability to LOSE 5€
1, 2, 3	C	Hard	2	A	10% probability to WIN 100€	90% probability to LOSE 10€
1, 2, 3	C	Hard	3	A	15% probability to WIN 95€	85% probability to LOSE 15€
1, 2, 3	C	Hard	4	A	20% probability to WIN 90€	80% probability to LOSE 20€
1, 2, 3	C	Hard	5	A	25% probability to WIN 85€	75% probability to LOSE 25€
1	NC		1		95% probability to WIN 105€	5% probability to LOSE 25€
1	NC		2		90% probability to WIN 100€	10% probability to LOSE 20€
1	NC		3		85% probability to WIN 105€	15% probability to LOSE 15€
1	NC		4		80% probability to WIN 110€	20% probability to LOSE 10€
1	NC		5		75% probability to WIN 115€	25% probability to LOSE 5€
1, 2, 3	F		1	A	70% probability to WIN 5€	30% probability to LOSE 20€
1, 2, 3	F		2	A	65% probability to WIN 5€	35% probability to LOSE 15€
1, 2, 3	F		3	A	60% probability to WIN 10€	40% probability to LOSE 20€
1, 2, 3	F		4	A	55% probability to WIN 10€	45% probability to LOSE 20€
1, 2, 3	F		5	A	50% probability to WIN 10€	50% probability to LOSE 15€
2, 3	C	Hard	1	B	5% probability to WIN 120€	95% probability to LOSE 5€
2, 3	C	Hard	2	B	10% probability to WIN 115€	90% probability to LOSE 10€
2, 3	C	Hard	3	B	15% probability to WIN 100€	85% probability to LOSE 15€

2, 3	C	Hard	4	B	20% probability to WIN 85€	80% probability to LOSE 20€
2, 3	C	Hard	5	B	25% probability to WIN 80€	75% probability to LOSE 25€
2, 3	NC		1	A	95% probability to WIN 120€	5% probability to LOSE 5€
2, 3	NC		2	A	90% probability to WIN 115€	10% probability to LOSE 10€
2, 3	NC		3	A	85% probability to WIN 100€	15% probability to LOSE 15€
2, 3	NC		4	A	80% probability to WIN 105€	20% probability to LOSE 20€
2, 3	NC		5	A	75% probability to WIN 110€	25% probability to LOSE 25€
2, 3	NC		1	B	95% probability to WIN 110€	5% probability to LOSE 5€
2, 3	NC		2	B	90% probability to WIN 100€	10% probability to LOSE 10€
2, 3	NC		3	B	85% probability to WIN 95€	15% probability to LOSE 15€
2, 3	NC		4	B	80% probability to WIN 115€	20% probability to LOSE 20€
2, 3	NC		5	B	75% probability to WIN 120€	25% probability to LOSE 25€
2, 3	F		1	B	70% probability to WIN 10€	30% probability to LOSE 30€
2, 3	F		2	B	65% probability to WIN 10€	35% probability to LOSE 25€
2, 3	F		3	B	60% probability to WIN 5€	40% probability to LOSE 15€
2, 3	F		4	B	55% probability to WIN 5€	45% probability to LOSE 15€
2, 3	F		5	B	50% probability to WIN 20€	50% probability to LOSE 25€
3	C	Easy	1	A	5% probability to WIN 290€	95% probability to LOSE 1€
3	C	Easy	2	A	10% probability to WIN 285€	90% probability to LOSE 5€
3	C	Easy	3	A	15% probability to WIN 260€	85% probability to LOSE 10€
3	C	Easy	4	A	20% probability to WIN 220€	80% probability to LOSE 15€
3	C	Easy	5	A	25% probability to WIN 205€	75% probability to LOSE 25€
3	C	Easy	1	B	5% probability to WIN 295€	95% probability to LOSE 1€
3	C	Easy	2	B	10% probability to WIN 290€	90% probability to LOSE 5€
3	C	Easy	3	B	15% probability to WIN 265€	85% probability to LOSE 10€
3	C	Easy	4	B	20% probability to WIN 225€	80% probability to LOSE 15€
3	C	Easy	5	B	25% probability to WIN 210€	75% probability to LOSE 25€

Lottery game

For the Lottery game the hard-conflict and no-conflict items were the same across studies 1-2-3. Regarding the filler items, while Study 1 and 2 did not include counterbalancing (all participants viewed the same items) in Study 3 we added a set B of filler items in order to counterbalance. Study 3 also included easy-conflict items.

Table S2. The lottery game items according to the study (1, 2, 3) they were used in, the type (Conflict, No-Conflict, Filler), the Difficulty level of the conflict items (Hard; Easy), the Items' Number, and the counterbalancing set (A; B).

Study	Type	Difficulty	Number	Set	Lottery A	Lottery B
1, 2, 3	C	Hard	1	A	60% probability to win 180€, 40% probability to win 130€	60% probability to win 330 €, 40% probability to win 1€

1, 2, 3	C	Hard	2	A	65% probability to win 200€, 35% probability to win 155€	65% probability to win 340€, 35% probability to win 5€
1, 2, 3	C	Hard	3	A	70% probability to win 350€, 30% probability to win 10€	70% probability to win 230€, 30% probability to win 160€
1, 2, 3	C	Hard	4	A	75% probability to win 360€, 25% probability to win 20€	75% probability to win 260€, 25% probability to win 165€
1, 2, 3	C	Hard	5	A	80% probability to win 370€, 20% probability to win 45€	80% probability to win 290€, 20% probability to win 170€
1, 2, 3	NC		1	B	60% probability to win 1€, 40% probability to win 330€	60% probability to win 130€, 40% probability to win 180€
1, 2, 3	NC		2	B	65% probability to win 5€, 35% probability to win 340€	65% probability to win 155€, 35% probability to win 200€
1, 2, 3	NC		3	B	70% probability to win 10€, 30% probability to win 350€	70% probability to win 160€, 30% probability to win 230€
1, 2, 3	NC		4	B	75% probability to win 165€, 25% probability to win 260€	75% probability to win 20€, 25% probability to win 360€
1, 2, 3	NC		5	B	80% probability to win 170€, 20% probability to win 290€	80% probability to win 45€, 20% probability to win 370€
1, 2, 3	C	Hard	1	B	60% probability to win 185€, 40% probability to win 135€	60% probability to win 335€, 40% probability to win 5€
1, 2, 3	C	Hard	2	B	65% probability to win 205€, 35% probability to win 160€	65% probability to win 345€, 35% probability to win 10€
1, 2, 3	C	Hard	3	B	70% probability to win 235€, 30% probability to win 165€	70% probability to win 355€, 30% probability to win 15€
1, 2, 3	C	Hard	4	B	75% probability to win 365€, 25% probability to win 25€	75% probability to win 265€, 25% probability to win 170€
1, 2, 3	C	Hard	5	B	80% probability to win 375€, 20% probability to win 50€	80% probability to win 295€, 20% probability to win 175€
1, 2, 3	NC		1	A	60% probability to win 5€, 40% probability to win 335€	60% probability to win 135€, 40% probability to win 185€
1, 2, 3	NC		2	A	65% probability to win 10€, 35% probability to win 345€	65% probability to win 160€, 35% probability to win 205€
1, 2, 3	NC		3	A	70% probability to win 165€, 30% probability to win 235€	70% probability to win 15€, 30% probability to win 355€
1, 2, 3	NC		4	A	75% probability to win 170€, 25% probability to win 265€	75% probability to win 25€, 25% probability to win 365€
1, 2, 3	NC		5	A	80% probability to win 175€, 20% probability to win 295€	80% probability to win 50€, 20% probability to win 375€
1, 2, 3	F		1	A	100% probability to win 200€, 0% probability to win 250€	100% probability to win 350€, 0% probability to win 300€
1, 2, 3	F		2	A	100% probability to win 350€, 0% probability to win 400€	100% probability to win 410€, 0% probability to win 460€

1, 2, 3	F		3	A	50% probability to win 200€, 50% probability to win 50€	50% probability to win 275€, 50% probability to win 125€
1, 2, 3	F		4	A	90% probability to win 350€, 10% probability to win 310€	90% probability to win 260€, 10% probability to win 220€
1, 2, 3	F		5	A	90% probability to win 340€, 10% probability to win 305€	90% probability to win 250€, 10% probability to win 215€
2, 3	F		1	B	100% probability to win 210€, 0% probability to win 250€	100% probability to win 360€, 0% probability to win 310€
2, 3	F		2	B	100% probability to win 355€, 0% probability to win 410€	100% probability to win 415€, 0% probability to win 470€
2, 3	F		3	B	50% probability to win 225€, 50% probability to win 75€	50% probability to win 300€, 50% probability to win 150€
2, 3	F		4	B	90% probability to win 330€, 10% probability to win 290€	90% probability to win 240€, 10% probability to win 200€
2, 3	F		5	B	90% probability to win 350€, 10% probability to win 310€	90% probability to win 260€, 10% probability to win 220€
3	C	Easy	1	A	60% probability to win 160€, 40% probability to win 130€	60% probability to win 480€, 40% probability to win 1€
3	C	Easy	2	A	65% probability to win 170€, 35% probability to win 155€	65% probability to win 470€, 35% probability to win 5€
3	C	Easy	3	A	70% probability to win 180€, 30% probability to win 160€	70% probability to win 440€, 30% probability to win 10€
3	C	Easy	4	A	75% probability to win 450€, 25% probability to win 20€	75% probability to win 210€, 25% probability to win 165€
3	C	Easy	5	A	80% probability to win 460€, 20% probability to win 45€	80% probability to win 250€, 20% probability to win 170€
3	C	Easy	1	B	60% probability to win 165€, 40% probability to win 135€	60% probability to win 485€, 40% probability to win 5€
3	C	Easy	2	B	65% probability to win 175€, 35% probability to win 160€	65% probability to win 475€, 35% probability to win 10€
3	C	Easy	3	B	70% probability to win 185€, 30% probability to win 165€	70% probability to win 445€, 30% probability to win 15€
3	C	Easy	4	B	75% probability to win 455€, 25% probability to win 25€	75% probability to win 215€, 25% probability to win 170€
3	C	Easy	5	B	80% probability to win 465€, 20% probability to win 50€	80% probability to win 255€, 20% probability to win 175€

C. Partial results excluding heuristics

When designing the items, we identified two possible heuristics, one for the betting game and one for the lottery game respectively which, when used, would result in a ceiled conflict accuracy. In the betting game this heuristic was "always taking the bet", while in the lottery game it was "always picking the lottery with the highest value in the set". However, the first heuristic would result in a low filler accuracy in the betting game, while the second would lead to a low no-conflict accuracy in the lottery game. To control for these heuristics and follow our preregistration, we excluded from the betting game participants with an accuracy lower than 50% both in their initial and final filler trials (n_{Study1}= 12; n_{Study2} = 8, n_{Study3} = 10), and from the lottery game participants with an accuracy lower than 50% both in their initial and final no-conflict trials (n_{Study1}= 4; n_{Study2} = 3, n_{Study3} = 2). Here, we report the partial accuracy (Figure S1, Table S3) and direction of change (Figure S2) results after excluding these participants.

As Figures S1 and S2 and Table S3 indicate, all trends regarding the proportion of EV maximizing choices and the direction of change respectively remained the same after the exclusion. Therefore, we conclude that these heuristics are not driving our results.

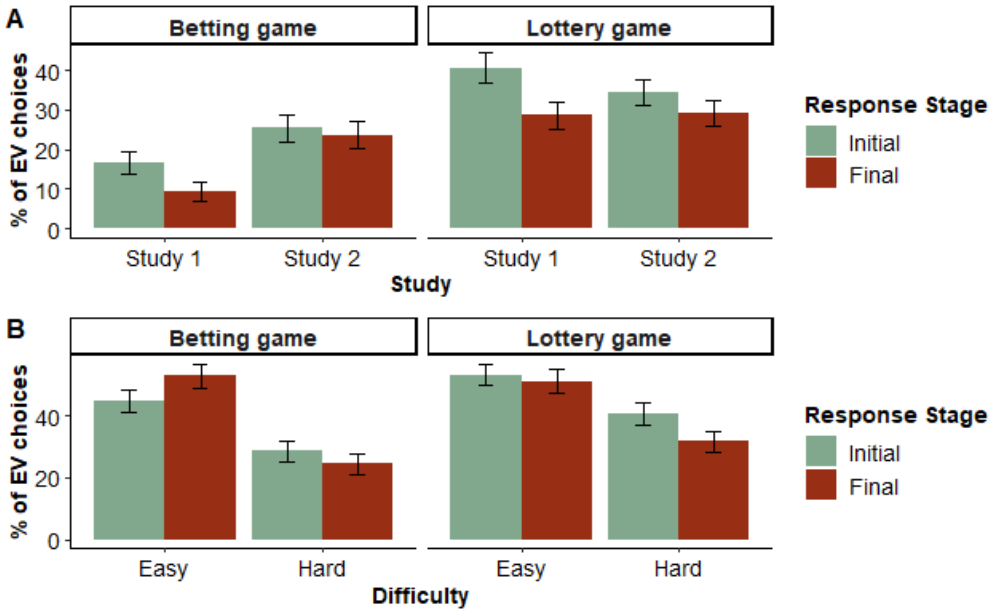

Figure S1. Percentage of Expected Value (EV) maximizing initial and final choices on conflict trials in the betting and lottery game. Panel A shows the means of Study 1 and 2, while Panel B shows those of Study 3, separately for the hard- and easy-conflict items. Error bars are standard errors of the mean.

Table S3. Paired-samples t-tests comparing the proportion of EV maximizing choices between the initial and the final stage, separately for each game, each study and for the hard- and easy-conflict items of Study 3. The mean difference is $M_{initial} - M_{final}$.

			mean difference	*t*	df
Betting game	Study 1		7%	3.07*	87
	Study 2		2%	0.65	91
	Study 3	Hard	4%	1.59	89
		Easy	−8%	−2.74*	89
Lottery game	Study 1		12%	4.46**	95
	Study 2		5%	1.86	96
	Study 3	Hard	9%	2.86*	97
		Easy	2%	0.74	97

*$p < .01$

**$p < .001$

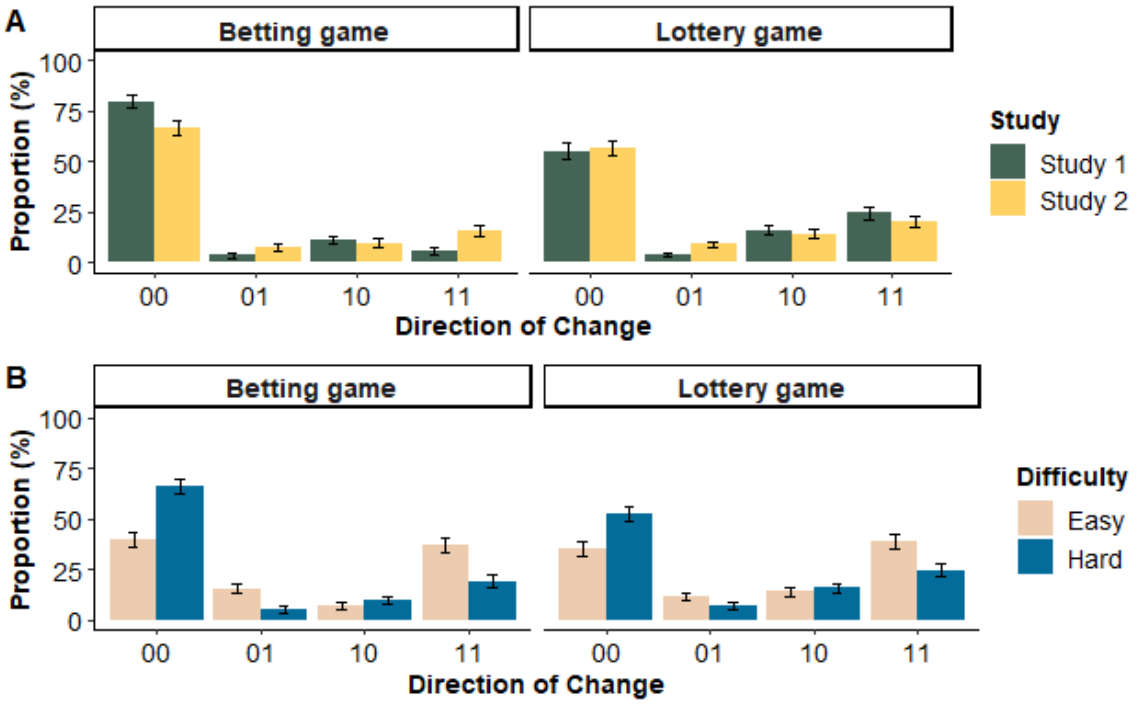

Figure S2. Proportion of each direction of change category in the betting game and the lottery game. Panel A shows the proportions of Study 1 and 2, while Panel B those of Study 3, separately for the hard- and easy-conflict items; "00" = initial and final loss averse response; "01" = initial loss averse response and final Expected Value (EV) maximizing response; "10" = initial EV maximizing response and final loss averse response; "11" = initial and final EV maximizing response. Error bars are standard errors of the mean.

D. Confidence as a function of direction of change

For exploratory purposes, we also analysed the confidence ratings as a function of direction of change. In the logical reasoning field, it has been repeatedly shown that trials where participants change their initial response after deliberation (i.e., "01" or "10" trials) tend to show lower initial response confidence than trials where participants stick to their initial answer (i.e., "11" or "00" trials, e.g., Bago & De Neys, 2017; Thompson et al., 2011). This low confidence (or "Feeling of Rightness") is considered a key determinant of answer change (Bago & De Neys, 2017; Thompson et al., 2011; Voudouri et al., 2022). Figure S3 shows the mean initial confidence ratings for each direction of change category for conflict items across our studies. Table S4 shows the Wilcoxon signed rank tests comparing the initial response confidence between change (i.e., "01" or "10") and no change (i.e., "11" or "00") trials. We replicate the reasoning findings as we find lower initial confidence for change compared to no change trials across our studies.

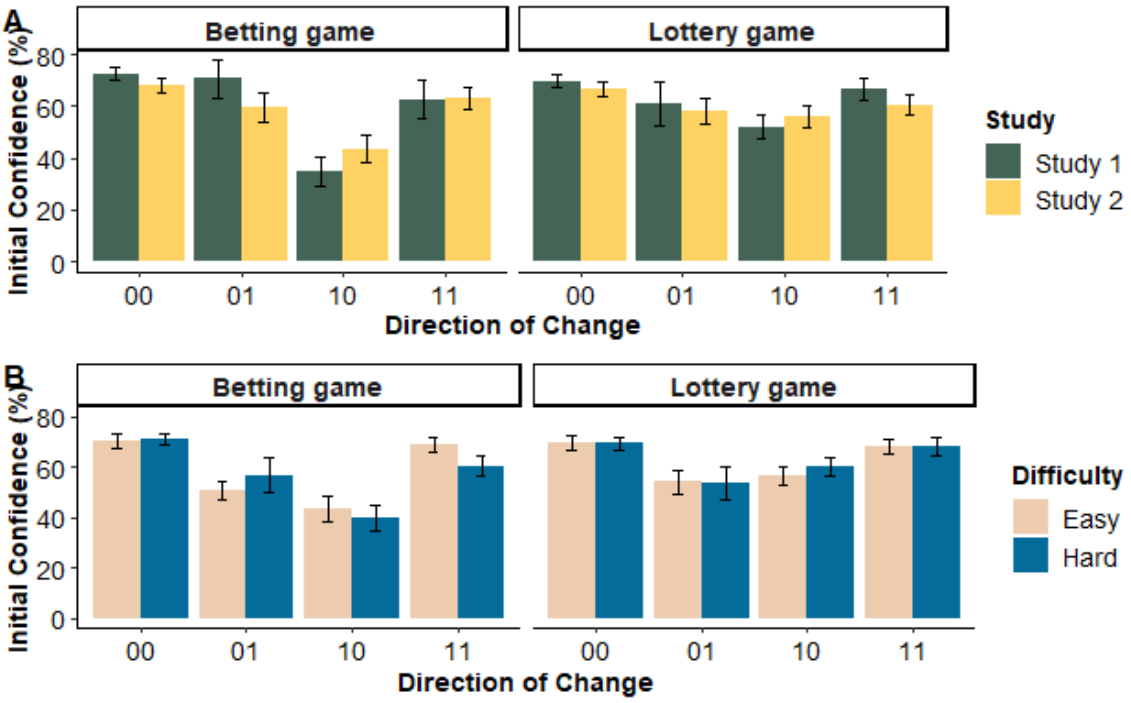

Figure S3. Initial confidence ratings at conflict trials as a function of each direction of change category separately for the betting and the lottery game. Panel A shows the confidence ratings in Study 1 and Study 2, while panel B show the ratings in Study 3, separately for easy- and hard-conflict trials; "00" = initial and final loss averse response; "01" = initial loss averse response and final Expected Value (EV)

maximizing response; "10" = initial EV maximizing response and final loss averse response; "11" = initial and final EV maximizing response. Error bars are standard errors of the mean.

Table S4. Wilcoxon signed rank tests comparing the initial response confidence between change (i.e., "01" or "10") and no change (i.e., "11" or "00") trials, separately for each game, each study and for the hard- and easy-conflict items of Study 3. The mean difference is $M_{confidence\_change} - M_{confidence\_nochange}$.

			mean difference	*t*	df
Betting game	Study 1		−19.49	−6.39**	59
	Study 2		−17.01	−7.58**	72
	Study 3	Hard	−13.79	−3.74**	42
		Easy	−16.60	−4.55**	53
Lottery game	Study 1		−15.57	−5.61**	80
	Study 2		−15.80	−7.37**	87
	Study 3	Hard	−7.64	−3.11*	53
		Easy	−11.15	−3.81**	59

*$p < .01$

**$p < .001$

E. Justifications

In Studies 2 and 3 the last presented item in the betting game and the lottery game was always the same (hard-) conflict item. After responding to this item participants were asked to provide a rationale for their final, deliberate response in an open-response format. This appeared on the screen (the instructions were adapted accordingly for the lottery game):

> We are interested in the reasoning behind your response to the final bet:
> If you take this bet you have:
>
> **25%** probability to **WIN €85**
> **75%** probability to **LOSE €25**
> Do you take the bet?
> - Yes
> - No
>
> Could you please justify, why do you think that your previously entered response is the most advantageous choice for you?

Based on the justifications, the authors defined post-hoc categories. If the justification was given to an EV maximizing response, it was categorized as "Expected Value" or "Gambler" in the betting game and as "Expected Value" or "Partial information" in the lottery game. An "Expected Value" justification referred to the amounts, the probabilities and their relationship (e.g., "The amount you can win is more than the amount you'd statistically lose at the given probabilities. So, say, in 4 bets, you would lose 75 but gain 80, netting 5."). In the betting game, a "Gambler" justification did not refer to the amounts and probabilities, but simply to a preference for gambling (e.g., "I only just preferred to take the risk, it was practically 50/50."). In the lottery game, a "Partial information" justification focused only on the first line of the lotteries (i.e., on the amounts corresponding to the large probabilities, e.g., "80% is a good chance at winning so I think that A will be better for me to maximize my potential winnings.").

If the justification was given to a loss averse response, it was categorized as "Loss aversion" or "Probabilities & Values". A "Loss aversion" justification referred either to the amount or the probabilities, but not to both (e.g., "The loss is too great to justify taking the risk. To lose 25 is too much to gamble away."). A "Probabilities & Values" justification referred both to the amounts and the

probabilities (e.g., "It's quite a low chance to win a small amount with a high chance of losing a sizeable amount.").

After the categories were defined, two coders classified the justifications. In the cases when an agreement was not reached, a third coder provided a classification and the most common category amongst the three coders was chosen. If none of the three coders agreed on a category or if they could not classify the justification to one of the categories, the justification was coded as "Other".

As it can be seen in Table S5, in the betting game, the majority of (the few) participants that chose the EV maximizing option could also explicitly justify it using expected value principles. In the lottery game however, most participants that opted for the EV maximizing choice failed to do this, and only focused on the largest probability when explicitly justifying their choice. In both games, when participants gave the loss averse option, about half the time they could justify it focusing both on the amounts and their probabilities. This indicates that in their explicit justifications about half of the loss averse participants considered all the necessary information (i.e., amounts and probabilities), while the other half only focused on either the amounts or the probabilities.

Table S5. Frequency (and count) of each justification category as a function of response accuracy (correct; incorrect), game (betting game; lottery game), and study (Study 2; Study 3). The justifications for EV maximizing responses are separated into the "Expected value", "Gambler/Partial information" and "Other" categories, while those for the loss averse responses are separated into the "Loss aversion", "Probabilities & Values" and "Other" categories.

Response	Justification category	Betting game		Lottery game	
		Study 2	Study 3	Study 2	Study 3
EV maximizing	Expected value	53.9% (7 out of 13)	42.9% (6 out of 14)	23.7% (9 out of 38)	32.1% (9 out of 28)
	Gambler	7.7% (1 out of 13)	28.6% (4 out of 14)		
	Partial information			57.9% (23 out of 38)	46.4% (13 out of 28)
	Other	38.5% (5 out of 13)	28.6% (4 out of 14)	18.4% (7 out of 38)	21.4% (6 out of 28)
Loss averse	Loss aversion	48.5% (32 out of 66)	50.8% (34 out of 67)	50.0% (28 out of 56)	36.8% (24 out of 57)
	Probabilities & Values	43.9% (29 out of 66)	46.3% (31 out of 67)	28.6% (16 out of 56)	42.1% (21 out of 57)
	Other	7.6% (5 out of 66)	3.0% (2 out of 67)	21.4% (12 out of 56)	21.1% (12 out of 57)

Supplementary material for Chapter 2

A. Trivia questions

Question number	No-anomaly question	Undistorted word	Strong impostor	Weak impostor	Undistorted answer	Filler answer
1	What kind of tree did the later president Washington allegedly chop down?	Washington	Lincoln	Nixon	Cherry	Palm
2	In what movie did Arnold Schwarzenegger go back in time to protect Sarah Connor?	Arnold Schwarzenegger	Sylvester Stallone	Johnny Depp	Terminator 2	Rocky 2
3	What country was Margaret Thatcher prime minister of for several years?	Prime minister	President	Queen	United Kingdom	France
4	In what year did Germany lose the second World War?	Lose	Win	Was the victor of	1945	1918
5	What kind of meat is in the Burger King sandwich known as the Whopper?	Burger King	McDonald's	Taco Bell	Beef	Chicken
6	What season do we associate with football games, starting school, and leaves turning brown?	Brown	Green	Black	Fall	Winter

Question number	No-anomaly question	Undistorted word	Strong impostor	Weak impostor	Undistorted answer	Filler answer
7	What statue given to the U.S. by France symbolizes freedom to immigrants arriving in New York?	France	England	Austria	Statue of Liberty	Christ the Redeemer
8	Who is the video game character and Italian plumber who is Nintendo's mascot?	Nintendo	Sony	Apple	Mario	Sonic
9	In the tale, who found the glass slipper left at the ball by Cinderella?	Cinderella	Snow White	Pocahontas	The prince	The stepmother
10	What is the name of the kimono-clad courtesans who entertain Japanese men?	Japanese	Chinese	French	Geisha	Samurai
11	Which instrument gives the time by measuring the angle of the sun's shadow on a dial?	Time	Temperature	Humidity	Sundial	Oscillator
12	What is the name of the comic strip character who eats spinach to improve his strength?	Strength	Sight	Intelligence	Popeye	Mickey Mouse

Question number	No-anomaly question	Undistorted word	Strong impostor	Weak impostor	Undistorted answer	Filler answer
13	What is the name of the current dictator of North Korea?	North	South	East	Kim Jong-Un	Fidel Castro
14	What is the name of the molten rock coming out of a volcano during an eruption?	Eruption	Earthquake	Tsunami	Lava	Mud
15	How do we call the man in the red suit and white beard who gives Christmas presents from his sleigh?	Christmas	Birthday	Wedding	Santa Claus	Rumpelstiltskin
16	What is the name of the Mexican dip made with mashed-up avocados?	Avocados	Artichokes	Cucumbers	Guacamole	Salsa
17	What is the name of the scary carved pumpkin displayed on Halloween?	Halloween	Thanksgiving	Easter	Jack-o'-lantern	Soul cake
18	When did the Japanese attack Pearl Harbor with their planes during World War II?	Japanese	Germans	Vietnamese	December 7th, 1941	December 7th, 1951
19	What is the name of the New Year festival celebrated on the 31st of December?	December	January	March	New Year's Eve	Carnival

Question number	No-anomaly question	Undistorted word	Strong impostor	Weak impostor	Undistorted answer	Filler answer
20	In the biblical story, how many animals of each kind did Noah take on the Ark?	Noah	Moses	Goliath	Two	Three

B. Instructions

Study 1-2-3

Please read these instructions carefully!

In this experiment you will have to answer 20 multiple-choice trivia questions and 2 practice questions.

For every multiple choice question you will be presented with four answer options but **you can only pick one answer. Please respond as accurately as you can.**

Some of the questions are impossible to answer. In that case, select the answer option: 'This question can't be answered in this form.'

If you don't know the answer to a question, select the response option 'Don't know'.

To clarify the difference between 'Don't know' and 'This question can't be answered in this form.', take a look at the example questions below:

What is the name of former president's Obama's oldest son?

Charles
Jonathan
This question can't be answered in this form.
Don't know.

The above question cannot be answered because Obama doesn't have a son; he only has two daughters. So, the correct answer option to this question is: 'This question can't be answered in this form.'

Here is a different example:

What is the name of former president's Obama's oldest daughter?

Sasha
Malia
This question can't be answered in this form.
Don't know.

In the above example, the question can be answered, since Obama does have an oldest daughter. The correct answer option is 'Malia'. However, if you do not know the answer to this question, you should select 'Don't know'.

Critically, in this study we want to know what your **initial, intuitive response** to the questions is and **how you respond after you have thought about these questions for some more time.**

First, we want you to respond with the **very first answer that comes to mind.** You don't need to think about it. Just give the first answer that intuitively comes to mind as quickly as possible.

To make sure that you answer as fast as possible, a time limit was set for the first response, which is going to be **5 seconds** (Study 1)/**4 seconds** (Study 2 - 3). When there is 1 second left, the background colour will turn to yellow to let you know that the deadline is approaching. Please make sure to **answer before the deadline passes.**

Next, **the question will be presented again** and you can take all the time you want to actively reflect on it. Once you have made up your mind you give your **final response.**

After you made your choice and clicked on it, you will be automatically taken to the next page.

After you have entered your first and final answer we will also ask you to indicate your confidence in the correctness of your response.

We are going to clarify all of this with a couple of practice questions.

First, a fixation cross will appear. Then, the question and the four answer options will appear. You then enter your first hunch as fast as possible before the deadline. Next, the question will be presented again and you can take all the time to reflect

on it and enter your final response.

After you have entered your first and final answer we will also ask you to indicate your confidence in the correctness of your response.

Participants were then given two practice trials, without the concurrent load task. They were then introduced to the load task.

You will also need to **memorize a pattern** while you respond to the trivia questions.

You will see a grid with crosses and you will have to memorize their location.

You will first practice with 2 patterns without a trivia question.

The pattern will be displayed for **2 seconds** and then you will have to select it among **4 different patterns.**

Participants were then given two practice trials for the cognitive load task, without the multiple-choice questions. They were then presented with the following instructions:

In the actual study you will need to memorize the pattern while you respond to the trivia question. The pattern is briefly presented before each question.

The difficulty of the pattern might vary. Always try to memorize as many crosses as possible. Each cross counts!

We know that it is not always easy to memorize the pattern while you are also thinking about the trivia question. The most important thing is to correctly memorize the pattern.

First, **try to concentrate on the memorization task**, and then try to answer the question accurately.

As a next step, you can practice this with two questions.

After those two last practice trials, participants were presented with the following instructions:

Ok, this is the end of practice!

During the experiment, the questions will be presented to you one after the other and you should not pause between them. After the first 10 questions, you can take a short break.

Remember, some of the questions are impossible to answer. In that case, select the answer option: 'This question can't be answered in this form.' If you don't know the answer to a question, select the response option 'Don't know'.

Study 3 pre-test

Please read these instructions carefully!

In this experiment, you will be presented with 25 items. Each item will include an undistorted, correct trivia question (along with the correct answer between brackets), and two distorted versions of this question. In the distorted versions, one or more words of the original question have been replaced by one or more "impostor" words. We know that people often fail to notice such replacements when the impostor word is very similar to the undistorted word.

Your task is to indicate **how similar each distorted sentence is to the original undistorted question on a scale ranging from 0 (Not at all similar) to 100 (Extremely similar).**

To clarify, take a look at the following example:

The undistorted question is: "What is the name of Harry Potter's female best friend, in the famous fantasy NOVEL by J.K. Rowling? (answer: Hermione Granger)".

How similar is each distorted sentence to the original undistorted question?

Please type a number from 0 (Not at all similar) to 100 (Extremely similar) for each sentence.

What is the name of Harry Potter's female best friend, in the famous fantasy POEM by J.K. Rowling?

What is the name of Harry Potter's female best friend, in the famous fantasy SONATA by J.K. Rowling?

In this example, you might think that the first version ("POEM") is more similar to the original version than the second one ("SONATA"), and that here "POEM" can go unnoticed more easily. In this case, you would have to give a higher similarity rating to the first distorted sentence than to the second one.

C. Confidence as a function of direction of change

For exploratory purposes, we also analyzed the confidence ratings as a function of direction of change in Study 1 and Study 2. A classic finding in the reasoning field is that trials where participants change their initial response (i.e., "01" or "10" response categories) tend to show lower initial response confidence than trials where participants stick to their initial answer (i.e., "11" or "00" responses, e.g., Bago & De Neys, 2017; Thompson et al., 2011). This low confidence (or "Feeling of Rightness") is considered as a key determinant of deliberate answer change (Bago & De Neys, 2017; Thompson et al., 2011). Figure S1 shows the mean initial confidence ratings for each direction of change category for anomaly problems. It is clear that we replicate the reasoning pattern and find lower initial confidence for the change ("01" and "10") than no change ("11" and "00") categories both in Study 1 and in Study 2.

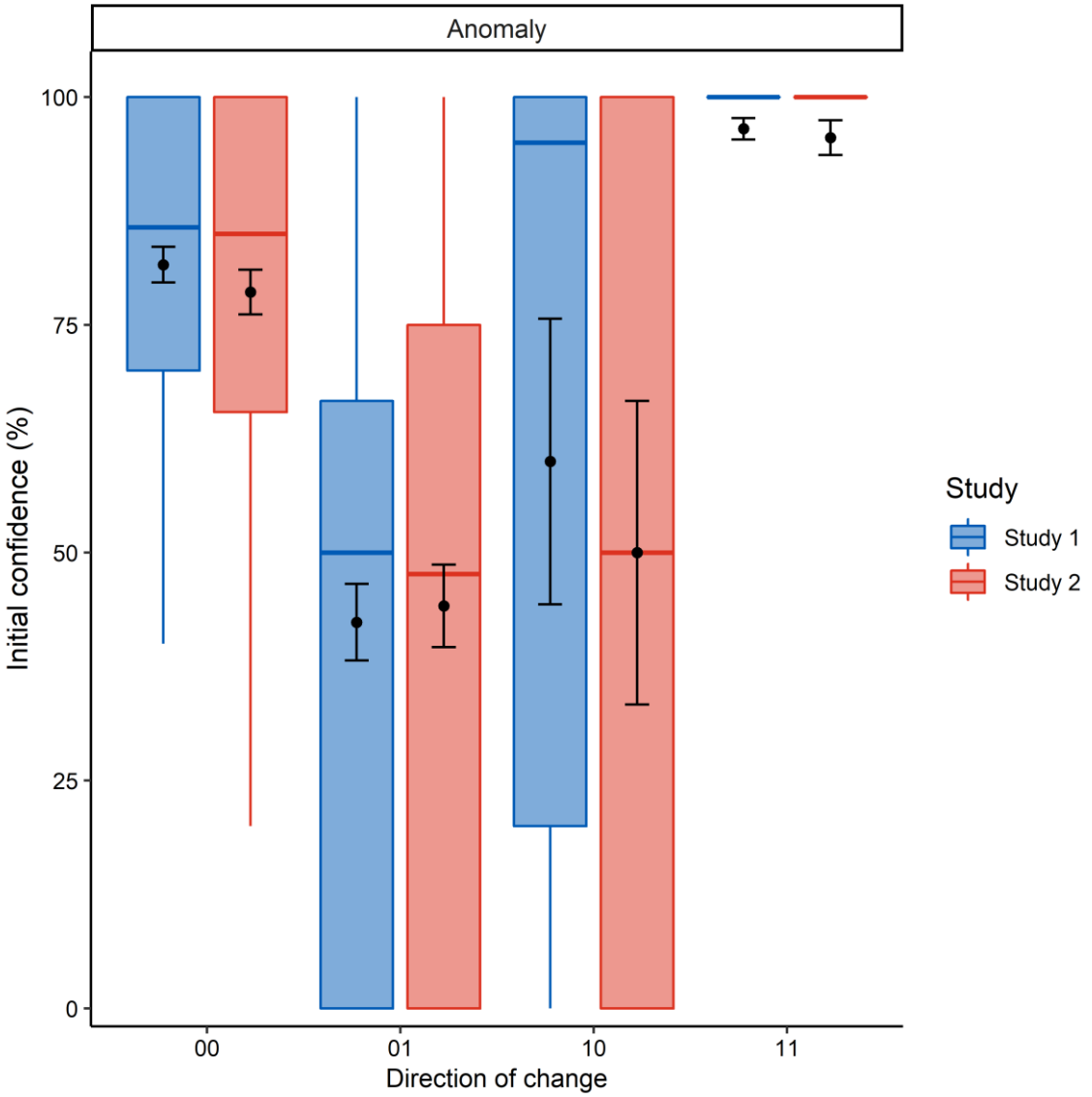

Figure S1. Initial confidence ratings at anomaly trials as a function of each direction of change category in Study 1 and Study 2; "00" = incorrect initial and incorrect final response; "01" = incorrect initial and correct final response; "10" = correct initial and incorrect final response; "11" = correct initial and correct final response. The lower and upper hinges of the boxplot correspond to the first and third quartiles, and the middle line shows the median. The lower (resp. upper) whiskers extend from the hinges to the smallest (resp. largest) value no further than 1.5 times the interquartile range. Overlaid black dots represent the mean and error bars are standard errors of the mean.

To analyze the results statistically, we used linear mixed-effects models. We ran a separate analysis for each direction of change category (see Bago & De Neys, 2017 for a similar analysis). Each model contrasted the initial confidence rating

for "11" control no-anomaly trials (which served as our baseline) to the initial confidence rating of each direction of change for anomaly trials. We will refer to this contrast as the Anomaly factor. We entered the Anomaly factor as fixed effect, as well as random intercepts for participants. We report the four models for Study 1 and Study 2 separately in Table S1 and Table S2 respectively. In both studies, the direction of change categories associated with the biggest confidence decrease compared to the control trials were the "01" and the "10" categories.

Table S1. Regression results contrasting the initial confidence ratings for "11" control no-anomaly trials with anomaly trials for each direction of change category in Study 1.

	00	01	10	11
Intercept (No-anomaly correct)	91.432 **	91.488 **	91.493 **	91.513 **
	(1.141)	(1.029)	(0.960)	(0.902)
Anomaly	-9.518 **	-44.994 **	-31.030 **	3.982 *
	(1.466)	(2.301)	(6.431)	(1.643)
N	1227	870	742	885
R2 (fixed)	0.032	0.304	0.029	0.006
R2 (total)	0.101	0.332	0.139	0.115

** $p < 0.001$; * $p < 0.05$.

Random effects (SD)				
Direction of change	00	01	10	11
Subjects	6.8	4.9	6.6	6.2
Residuals	24.6	24.3	18.5	17.6

Table S2. Regression results contrasting the initial confidence ratings for "11" control no-anomaly trials with anomaly trials for each direction of change category in Study 2.

	00	01	10	11
Intercept (No-anomaly correct)	88.379 **	88.679 **	88.617 **	88.566 **
	(1.659)	(1.541)	(1.470)	(1.444)
Anomaly	-9.162 **	-42.612 **	-40.132 **	6.070 *
	(1.736)	(2.666)	(7.885)	(2.614)
N	1002	737	593	667
R2 (fixed)	0.024	0.252	0.037	0.007
R2 (total)	0.202	0.332	0.225	0.238

** $p < 0.001$; * $p < 0.05$.

Random effects (SD)				
Direction of change	00	01	10	11
Subjects	12.3	9.7	11.1	11.4
Residuals	26.1	28.2	22.5	20.7

D. Illusion strength models

Table S3. Regression results of initial confidence as a function of illusion strength for control no-anomaly correct (baseline), anomaly correct and anomaly incorrect responses. Illusion strength = mean initial no-anomaly accuracy − mean initial anomaly accuracy for each item. The illusion strength variable was centered to show the effect of the response group factor for the mean illusion strength value.

	Study 1	Study 2
Intercept (No-conflict correct)	90.341 ***	86.682 ***
	(1.251)	(1.716)
Conflict correct	3.539	1.954
	(2.796)	(3.974)
Conflict incorrect	-17.312 ***	-16.166 ***
	(1.518)	(1.826)
Illusion strength	0.264 ***	0.252 ***
	(0.057)	(0.076)
Conflict correct * Illusion strength	-0.360 **	-0.455 *
	(0.137)	(0.201)
Conflict incorrect * Illusion strength	0.185 *	-0.007
	(0.085)	(0.107)
N	1542	1270
R2 (fixed)	0.120	0.076
R2 (total)	0.177	0.189

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

Random effects (SD)		
	Study 1	Study 2
Subjects	7.3	11.5
Residuals	27.7	30.8

E. Distribution of individual non-correction rates and initial errors sensitivity measures

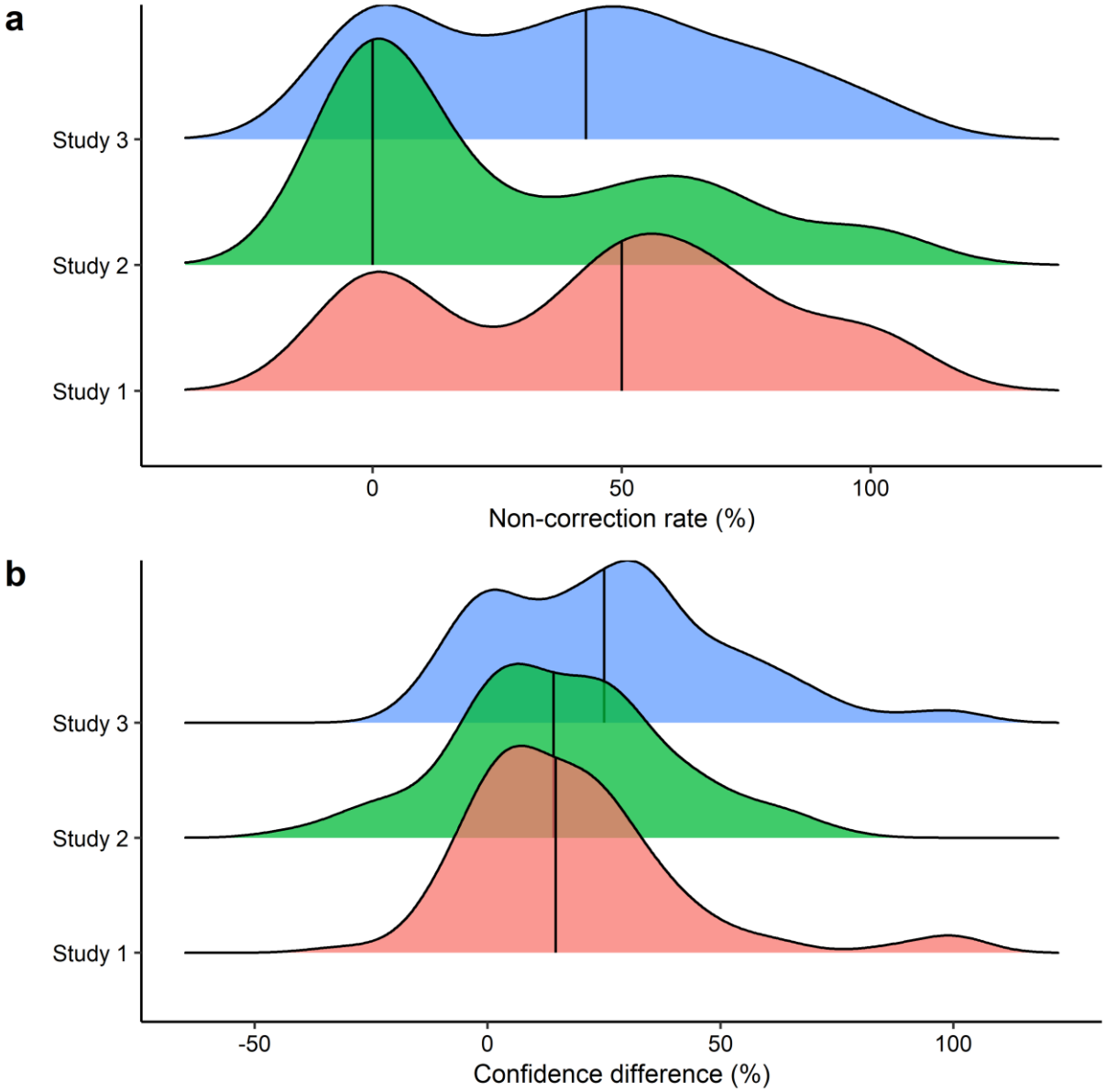

Figure S2. Ridgeline density plots of non-correction rates and initial error sensitivity measures for the three studies. **a**) Individual non-correction rates for anomaly problems. **b**) Initial confidence difference between the correct control no-anomaly trials and the incorrect anomaly trials. Black lines indicate the median.

Supplementary material for Chapter 3

A. Instructions

Stroop task instructions

The literal instructions that were used in the two-response Stroop task stated the following:

> Welcome to the experiment! This experiment will take about 30 minutes to complete and it demands your full attention. You can only do this experiment once. Click on Next to start. Please read these instructions carefully! In this task you will be presented with words, one after the other, to the centre of the screen, and you need to respond to the colour that each word is presented in. Press: d for red; f for blue; j for green; k for yellow. You can see an example of the words below. In this example you would have to press f for blue.
>
> The word "blue" written in blue ink colour was displayed on screen.
>
> We are going to start with a couple of practice problems to familiarise you with the buttons.
> In this practice you will only be presented with colours, not words. First, a fixation cross will appear. Then a colour will appear and you will need to click on the corresponding button. Please respond as fast and as accurately as possible (try to answer as fast as you can while not making mistakes). After you respond, you will be given feedback for your responses. Once you click on the button, you will be automatically taken to the next page. Remember: Press d for red; f for blue; j for green; k for yellow. Prepare yourself by holding the middle and index fingers of your left hand on the "d" and "f" keys and the middle and index fingers of your right hand over the "j" and "k" keys, like it is shown below. Press SPACE to start the practice.
>
> This is the end of this practice. Now you are going to practice with the words. You need to respond to the colour that each word is presented in. Please respond as fast and as accurately as possible (try to answer as fast as you can while not making mistakes). You will be given feedback for your responses. Remember: Press d for red; f for blue; j for green; k for yellow

Prepare yourself by holding the middle and index fingers of your left hand on the "d" and "f" keys and the middle and index fingers of your right hand over the "j" and "k" keys, like it is shown below. In the actual experiment, sometimes the ink color in which the word appears will not match with the word. For example, the following word could appear:

The word “green” written in yellow ink was displayed on screen.

Here the word "green" is written in yellow. We ask you to always respond to the color of the word. So in this example you would need to press the button ‘k’ for ‘yellow’. We will let you practice a couple of these now. You will get feedback for your responses. Press d for red; f for blue; j for green; k for yellow. Prepare yourself by holding the middle and index fingers of your left hand on the "d" and "f" keys and the middle and index fingers of your right hand over the "j" and "k" keys, like it is shown below. Press SPACE when you are ready to start the practice.

This is the end of this practice. In the actual task, you will give two responses to each word. First, we want to know what your initial, intuitive response to the colour of each word is and afterwards we want to see how you respond after you have thought about the colour of each word for some more time. So, for the first response you need to give the very first answer that comes to mind. You don't need to think about it. Just give the first answer that intuitively comes to mind as quickly as possible. To make sure that you answer as fast as possible, a time limit is set for the first response, which is going to be 750 milliseconds (that's less than a second!). Please make sure to answer before the deadline passes. In the next part, you are going to watch an initial trial to get a feel of the deadline. Press Next to see the trial.

This is how fast the word is going to be presented! You need to give a response within this time. You are now going to practice this with some words. First, a fixation cross will appear. Then the word will appear and you will need to click on the button that corresponds to the colour of the word. As we mentioned before, we are first interested in your initial, intuitive response. Next, the word will be presented again and you can take all the time you want to actively reflect on your choice. Once you have made up your mind you enter your final response. After you click on the button, you will be automatically taken to the next page. From here on we will no longer tell you whether the color you picked was correct or not. We will let you know whenever you responded too slowly and missed the deadline. Remember:

Press d for red; f for blue; j for green; k for yellow. Prepare yourself by holding the middle and index fingers of your left hand on the "d" and "f" keys and the middle and index fingers of your right hand over the "j" and "k" keys, like it is shown below. Press SPACE to start the practice.

This is the end of this practice. In the actual task, you will also need to memorise six numbers while you respond to the words. The numbers will be displayed for 2 seconds and then you will view one number with a question mark. You have to press 'd' for yes, the number was part of the set, or press 'k' for no, the number was not part of the set. There is no deadline for your response. You will get feedback after each response. To better understand this, you will first practise with five sets of numbers without the words. You should prepare yourself by holding the index finger of your left hand on the "d" key and the index finger of your right hand over the "k" key. Press SPACE to begin.

In the actual task you will need to memorise the numbers while you respond to the words. The numbers will be briefly presented before each word. We know that it is not always easy to memorise the numbers while you are also thinking about the words. The most important thing is to correctly memorise the numbers. First, try to concentrate on the memorisation task, and then try to solve the colour-word task. The memorization will only be required for your first, intuitive response. For your final response you can take as much time as you want without having to memorize the pattern. You can practice this in this practice round. Remember: Press d for red; f for blue; j for green; k for yellow. Prepare yourself by holding the middle and index fingers of your left hand on the "d" and "f" keys and the middle and index fingers of your right hand over the "j" and "k" keys, like it is shown below. Press SPACE to continue to the practice.

This is the end of all practice rounds! Now you will begin with the task. In the colour-word task there will be a total of 128 trials grouped in 3 blocks. After each block you can take a short break. Within each block one trial will be presented immediately after the other and you should not pause between them. In total the 3 blocks will take approximately 15 minutes. Please make sure to stay maximally focused throughout the study. Remember: Press d for red; f for blue; j for green; k for yellow. Prepare yourself by holding the middle and index fingers of your left hand on the "d" and "f" keys and the middle and index fingers of your right hand over the "j" and "k" keys, like it is shown below. Press SPACE when you’re ready to start with the first block

BREAK You just finished the first block! There are two blocks remaining. Feel free to take a short break. Before you start remember: Press d for red; f for blue; j for green; k for yellow. Prepare yourself by placing the middle and index fingers of your left hand on the 'd' and 'f' keys and the middle and index fingers of your right hand over the 'j' and "k" keys, like it is shown below. Press SPACE when you are ready to continue to the next block.

Flanker task instructions

The literal instructions that were used in the two-response Flanker task stated the following:

Welcome to the experiment!
This experiment will take about 24 minutes to complete and it demands your full attention. You can only do this experiment once. Click on Next to start. Please read these instructions carefully! In this task you will be presented with an arrow at the center of the screen, which will look like the arrows that are shown below.

Two arrows, one pointing to the left and one to the right, were displayed on the screen.

Your task will be to press the button that matches the direction the arrow is pointing to. Click on Next to continue. Press F if the central arrow is pointing Left. Press the correct key to continue. Press J if the arrow is pointing Right. Press the correct key to continue.

Two rows of five arrows were displayed on the screen, one after the other.

The central arrow will always be presented along with four other arrows as it is shown below. Your task is to identify the direction of the CENTRAL arrow. Ignore the peripheral arrows. Remember: Press F if the central arrow is pointing Left. Press J if the central arrow is pointing Right. Click on Next to continue.

We are going to start with 6 practice trials to familiarise you with the buttons. First, a fixation cross will appear. Then five arrows will appear and you should identify the direction of the CENTRAL arrow by clicking on the corresponding button.

Remember: Press F if the central arrow is pointing Left. Press J if the central arrow is pointing Right. You should prepare yourself by holding the index finger of your left hand on the F key and the index finger of your right hand on the J key. After you respond, you will be given feedback for your responses. Once you click on a key, you will be automatically taken to the next trial. Press SPACE to start the practice.

This is the end of this practice. In the actual task, you will give two responses to each trial. First, we want to know what your initial, intuitive response to the direction of the central arrow is and afterwards we want to see how you respond after you have thought about it for some more time. So, for the first response you need to give the very first answer that comes to mind. You don't need to think about it. Just give the first answer that intuitively comes to mind as quickly as possible. To make sure that you answer as fast as possible, a time limit is set for the first response, which is going to be 420 milliseconds (that's less than half a second!). Please make sure to answer before the deadline passes. In the next part, you are going to watch an initial trial to get a feel of the deadline. Press Next to see the trial.

After a fixation cross was shown, a row of five arrows was displayed on the screen.

This is how fast the word is going to be presented! You need to give a response within this time. You are now going to practice this with some trials. First, a fixation cross will appear. Then the arrows will appear and you will need to click on the button that corresponds to the direction of the central arrow. As we mentioned before, we are first interested in your initial, intuitive response. Next, you will see the reminder "Please give your final response". The same arrows will be presented again and you can take all the time you want to actively reflect on the direction of the central arrow. Once you have made up your mind you can enter your final response. After you click on the key, you will be automatically taken to the next trial. We will no longer tell you whether the direction you picked was correct or not. We will only let you know whenever you responded too slowly and missed the deadline. Remember: Press F if the central arrow is pointing Left. Press J if the central arrow is pointing Right. You should prepare yourself by holding the index finger of your left hand on the F key and the index finger of your right hand on the J key. Press SPACE to start this practice session.

This is the end of this practice. In the actual task, you will also need to memorise six numbers while you view the arrows. The numbers will be displayed for 2 seconds and then you will view one number with a question mark. You have to press F for yes, the number was part of the set, or press J for no, the number was not part of the set. There is no deadline for your response. You will get feedback after each response. To better understand this, you will first practise with five sets of numbers without the arrows. You should prepare yourself by holding the index finger of your left hand on the F key and the index finger of your right hand over the J key. Press SPACE to begin.

In the actual task you will need to memorise the numbers while you respond to the direction of the central arrow. The numbers will be briefly presented before the arrows. We know that it is not always easy to memorise the numbers while you are also thinking about the direction of the central arrow. The most important thing is to correctly memorise the numbers. First, try to concentrate on the memorisation task, and then try to solve the arrow task. The memorization will only be required for your first, intuitive response. For your final response you can take as much time as you want without having to memorize the pattern. You can practice this in this practice round.

This is the end of all practice rounds! Now you will begin with the task. There will be a total of 128 trials grouped in 3 blocks. After each block you can take a short break. Within each block one trial will be presented immediately after the other and you should not pause between them. In total the 3 blocks will take approximately 18 minutes. Please make sure to stay maximally focused throughout the study. Remember: Press F if the central arrow is pointing Left. Press J if the central arrow is pointing Right. You should prepare yourself by holding the index finger of your left hand on the F key and the index finger of your right hand over the J key. Press SPACE when you’re ready to start with the first block.

BREAK You just finished the first block! There are two blocks remaining. Feel free to take a short break. Before you start remember: Press F if the central arrow is pointing Left. Press J if the central arrow is pointing Right. You should prepare yourself by holding the index finger of your left hand on the F key and the index finger of your right hand over the J key. Press SPACE when you are ready to continue to the next block.

B. Reaction times

Table S1.

Mean (*SD*) of reaction times in Study 1 (Stroop task), Study 2 (Flanker task) and Study 3 (Stroop task) as a function of congruency status (congruent; incongruent), Response stage (initial response; final response) and response accuracy (correct; incorrect; overall). Reaction times are expressed in milliseconds. The first column ("Overall") refers to both correct and incorrect trials combined.

		Overall		Correct		Incorrect	
		Incongruent	Congruent	Incongruent	Congruent	Incongruent	Congruent
Study 1	Initial	581 (55)	542 (104)	577 (56)	557 (65)	586 (66)	532 (133)
	Final	982 (744)	890 (873)	1019 (890)	895 (884)	942 (786)	769 (463)
Study 2	Initial	315 (44)	306 (58)	316 (48)	315 (56)	301 (43)	253 (66)
	Final	543 (260)	515 (225)	529 (210)	516 (225)	484 (331)	420 (181)
Study 3	Initial	580 (53)	545 (53)	576 (54)	547 (48)	592 (72)	542 (89)
	Final	1096 (2536)	764 (640)	1078 (2524)	765 (669)	1166 (2709)	831 (523)

C. Inclusion of all trials

Table S2.

Direction of change proportions (%) by Congruency status (congruent; incongruent) in Study 1 (Stroop task), Study 2 (Flanker task) and Study 3 (Stroop task) including all missed load and missed deadline trials. All missed deadline trials were coded as "0" (i.e., incorrect response).

		"00"	"01"	"10"	"11"
Study 1	Incongruent	10.5	47.2	1.1	41.2
	Congruent	2.8	35.3	1.2	60.7
Study 2	Incongruent	7.2	52.3	1.3	39.2
	Congruent	1.2	41.3	0.4	57.2
Study 3	Incongruent	8.7	46.2	1.9	43.2
	Congruent	2.5	32.2	1.7	63.7

Note. "00" = incorrect initial and incorrect final response; "01" = incorrect initial and correct final response; "10" = correct initial and incorrect final response; "11" = correct final and correct initial response.

D. Stroop task results of Study 3

Accuracy

As Figure S1A shows, we replicated the key pattern of results that we observed in Study 1. When participants were allowed to deliberate, they typically managed to solve incongruent trials correctly, but they still performed better on congruent compared to incongruent trials. The mean accuracy for the initial responses of the congruent trials was 79.8% (*SD* = 18.2%) and differed from 25% chance, $t(140) = 35.87$, $p < .001$. The mean accuracy for the initial responses of the critical incongruent trials was 63.6% (*SD* = 24.3%) and also differed from 25% chance, $t(138) = 18.67$, $p < .001$. This suggests that even when participants were forced to rely on intuitive, automatic processing, they were often able to produce correct responses. To see if there was an effect of the response stage (initial; final) and the congruency status (congruent; incongruent) on the accuracy of the Stroop responses, a two-way within-subjects ANOVA was conducted. As Figure S1A shows, the accuracy for congruent trials was higher than for incongruent trials, $F(1, 137) = 70.41$, $p < .001$, $\eta^2g = 0.103$, and the accuracy at the final stage was higher than at the initial stage, $F(1, 137) = 275.40$, $p < .001$, $\eta^2g = 0.287$, indicating that accuracy improved after deliberation. Finally, the difference between initial and final accuracy was higher for incongruent compared to congruent trials, as indicated by the response stage by congruency interaction, $F(1, 137) = 60.93$, $p < .001$, $\eta^2g = 0.287$.

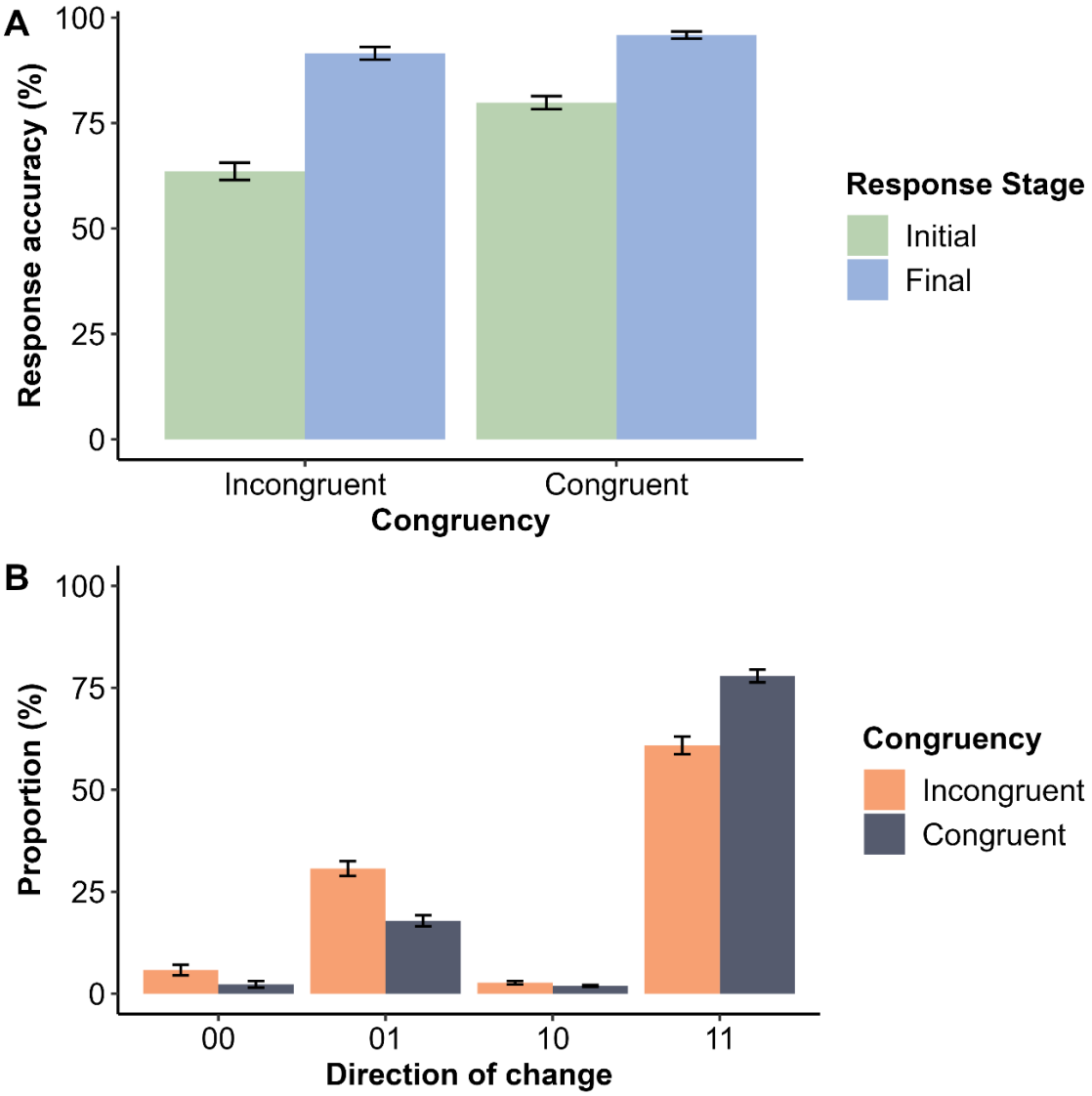

Figure S1. Accuracy and Direction of Change in the Stroop task of Study 3. **A)** Response accuracy at incongruent and congruent trials as a function of response stage. **B)** Proportion of each direction of change category in incongruent and congruent trials. The error bars represent the Standard Error of the Mean. "00" = incorrect initial and incorrect final response; "01" = incorrect initial and correct final response; "10" = correct initial and incorrect final response; "11" = correct final and correct initial response.

Stability index

The average stability index for the initial responses of the critical, incongruent trials was 74.2% (*SD* = 13.8%). If initial responding was prone to systematic guessing, we would expect more inconsistency in participants' initial responses across trials.

Direction of change

To get a more precise picture of how participants changed their responses after deliberation we conducted a direction of change analysis (Bago & De Neys, 2017, 2019a). The proportions of each direction of change were very similar to those of Study 1. As Figure S1B shows, the vast majority of the critical, incongruent trials had a "11" pattern (60.9%). The high "11" proportion was accompanied by a low "00" proportion (5.8%), and "10" proportion (2.7%). More importantly, the proportion of "11" trials was higher than that of the "01" trials (30.7%). The non-correction rate (i.e., proportion 11/11+01) reached 66.5%. This confirms the results of Study 1 and indicates that, in most correct final trials, the correct response was already generated when deliberate control was minimized.

As Figure S1B shows, and as it was expected, a similar pattern was observed for congruent trials. In most trials, correct responses were intuitively generated and the non-correction rate reached 81.3%.

Reaction Times

The average reaction time at the initial response stage was 545 ms (*SD* = 53 ms) for the congruent trials, and 580 ms (*SD* = 53 ms) for the incongruent trials. Participants spent longer on the final response stage, with an average of 764 ms (*SD* = 640 ms) at congruent trials and 1096 ms (*SD* = 2536 ms) at incongruent trials. Supplementary Material section B gives a full overview of reaction times according to response accuracy.

Exploratory analysis

To make maximally sure that participants did not deliberate during the initial response stage, we excluded a considerable amount of trials. As mentioned in Study 1, this could have artificially boosted the critical non-correction rate. To examine this possibility, we re-ran the direction of change analysis while including all missed load and missed deadline trials. As in Study 1, we opted for the strongest possible test and coded the accuracy of all missed deadline trials as "0" (i.e., incorrect). In the missed load trials both initial and final responses were recorded. The analysis, as reported in Supplementary Material section C, pointed to a higher proportion of "01" incongruent trials (46.2%), but the proportion of "11" (43.2%) responses and the non-correction rate remained high (48.3%). As

in Study 1, even in this extremely conservative analysis, correct incongruent responses were still generated intuitively about half of the time.

To summarize, regarding the Stroop task, the results of Study 3 replicated those of Study 1, with a much larger sample. This confirms the main finding of Study 1: even when deliberate control is minimized, participants can typically still provide correct Stroop responses. This suggest that more often than not, correct responding on the Stroop trial seems to be done intuitively in the absence of deliberate controlled correction.

E. Full cross tabulation table of correlation

Table S3

Pearson's product-moment correlation tests between the proportion of each direction of change (i.e., "00", "01", "10", "00") of each individual at the Stroop task, and the proportion of each direction of change of that individual at the Reasoning task of Study 3. Correlations are reported both at the composite level and separately for each type of reasoning problem.

Direction Reasoning	Task	Direction Stroop							
		00		01		10		11	
		r	p	r	p	r	p	r	p
00	BB	0.14	0.157	**−0.23**	0.016	0.11	0.251	0.07	0.439
	BR	**0.19**	0.029	−0.05	0.581	0.13	0.137	−0.10	0.252
	SYL	0.06	0.521	−0.04	0.676	0.03	0.743	−0.01	0.902
	Composite	**0.17**	0.040	−0.09	0.313	0.12	0.166	−0.06	0.503
01	BB	−0.08	0.432	**0.20**	0.033	−0.06	0.545	−0.11	0.276
	BR	−0.08	0.333	0.04	0.622	−0.11	0.206	0.04	0.675
	SYL	0.12	0.169	0.10	0.260	−0.07	0.404	−0.14	0.108
	Composite	−0.55	0.949	**0.17**	0.044	−0.12	0.147	−0.11	0.180
10	BB	−0.06	0.516	0.09	0.326	−0.03	0.769	−0.03	0.760
	BR	−0.07	0.433	−0.13	0.142	−0.05	0.598	0.15	0.073
	SYL	0.07	0.439	0.14	0.098	**0.19**	0.022	**−0.19**	0.023
	Composite	−0.02	0.781	0.07	0.427	0.12	0.161	−0.06	0.457
11	BB	−0.12	0.227	0.15	0.122	−0.10	0.291	−0.02	0.817
	BR	−0.5	0.136	0.06	0.475	−0.06	0.472	0.04	0.654
	SYL	−0.13	0.128	−0.06	0.464	−0.06	0.449	0.14	0.092
	Composite	**−0.18**	0.038	0.76	0.930	−0.12	0.172	0.12	0.149

Note. BB = Bat-and-ball; BR = Base-rates; SYL = Syllogisms; "01" = incorrect initial and correct final response; "00" = incorrect initial and incorrect final response; "10" = correct initial and incorrect final response; "11" = correct final and correct initial response. Significant correlations ($p < .05$) are in bold.

F. Reasoning confidence

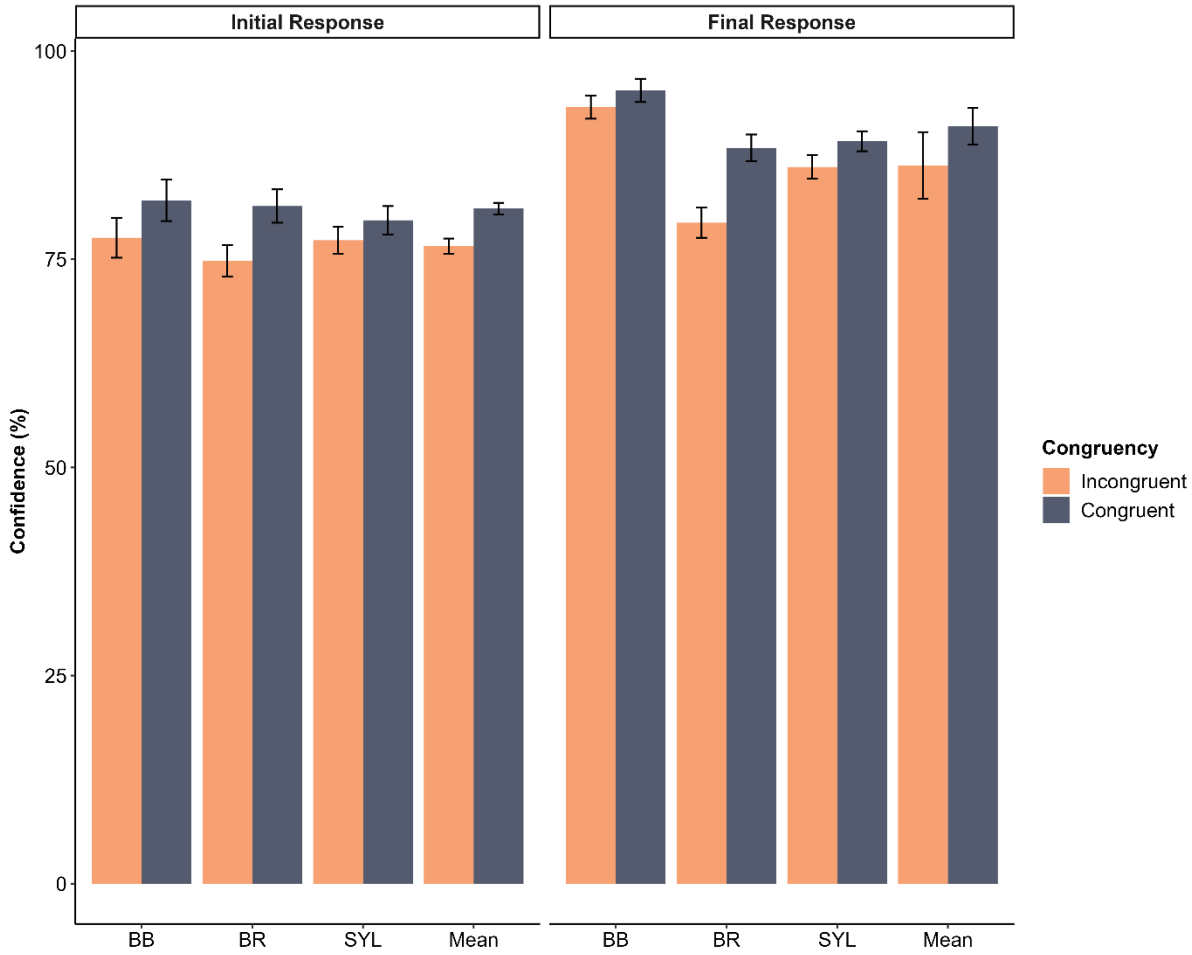

Figure S2. Confidence at initial and final responses, at congruent and incongruent trials in the Reasoning task of Study 3, separately for each problem type and for the mean across the three problem types. The error bars represent the Standard Error of the Mean. BB = Bat-and-ball; BR = Base-rates; SYL = Syllogisms; Mean = the mean across the four tasks.

Table S4.

Mean (*SD*) of the reported confidence at the initial responses of the Reasoning task of Study 3, as a function of congruency status (Congruent; Incongruent) and problem type (BB; BR; SYL; Mean).

	BB	BR	SYL	Mean
Incongruent	77.6 (28.9)	74.8 (26.4)	77.3 (25.4)	76.6 (1.5)
Congruent	82.1 (30.2)	81.4 (26.3)	79.7 (26.0)	81.1 (1.2)

Note. BB = Bat-and-ball; BR = Base-rates; SYL = Syllogisms; Mean = the mean across tasks.

G. Correlation results according to "11" conflict level

Conflict detection in the Reasoning task of Study 3 was calculated by subtracting the baseline confidence (i.e., the confidence at the correct congruent trials), from the confidence at the incongruent trials (e.g., De Neys et al., 2013; Mevel et al., 2015; Pennycook et al., 2015). The higher the difference between the two, the more conflict is thought to be experienced by the reasoner in the incongruent trials. In sum, high conflict detection is equivalent to low response confidence.

Table S5

Summary statistics of the initial conflict detection at the "11" trials of the Reasoning task (Study 3), separately for the half of the group that had a high conflict detection at "11" trials ("High half") and for the half of the group that had a low conflict detection at "11" trials ("Low half"). Negative values point to an overall successful conflict detection.

	N	Min	Max	Median	Q1	Q3	IQR	Mean	SD	SE	CI
High half	58	−100	-5	−16.86	−24.69	−12.5	12.19	−21.17	15.64	2.05	4.11
Low half	58	−3.87	33.33	0.44	0	9.67	9.67	5.31	8.84	1.16	2.32

Table S6

Correlation tests between the proportion of each direction of change of each individual at the Stroop task (Study 3), and the proportion of each direction of change of that individual at the Reasoning task (Study 3), for the half of the participants that had a high conflict detection at "11" trials.

	BB		BR		SYL		Composite	
	r	*p*	r	*p*	r	*p*	r	*p*
00	0.19	0.226	0.25	0.061	−0.30	0.841	0.22	0.109
01	**0.34**	0.029	***0.41***	0.002	0.003	0.982	**0.35**	0.009
11	0.10	0.524	**0.29**	0.029	0.23	0.088	**0.34**	0.010
10	−0.06	0.718	−0.07	0.617	**0.33**	0.014	0.18	0.185

Note. BB = Bat-and-ball; BR = Base-rates; SYL = Syllogisms; "00" = incorrect initial and incorrect final response; "01" = incorrect initial and correct final response; "10" = correct initial and incorrect final response; "11" = correct final and correct initial response. Significant correlations at the 0.05 level are in bold. Significant correlations at the 0.01 level are in bold and italics.

Table S7

Correlation tests between the proportion of each direction of change of each individual at the Stroop task (Study 3), and the proportion of each direction of change of that individual at the Reasoning task (Study 3), for the half of the participants that had a low conflict detection at "11" trials.

	BB		BR		SYL		Composite	
	r	*p*	r	*p*	r	p	r	*p*
00	0.15	0.332	0.25	0.053	0.16	0.230	0.25	0.056
01	0.12	0.450	−0.17	0.211	**0.34**	0.009	0.15	0.248
11	−0.07	0.664	−0.13	0.349	0.25	0.060	0.07	0.581
10	0.05	0.985	−0.12	0.357	0.07	0.604	0.002	0.985

Note. BB = Bat-and-ball; BR = Base-rates; SYL = Syllogisms; "00" = incorrect initial and incorrect final response; "01" = incorrect initial and correct final response; "10" = correct initial and incorrect final response; "11" = correct final and correct initial response. Significant correlations ($p < .05$) are in bold.

Supplementary material for Chapter 4

A. Accuracy Correlations

Table S1.

Pearson's product-moment correlation tests between the average accuracy of each individual at the conflict problems of session 1, and the accuracy of that individual at the conflict problems of session 2, separately for each reasoning task.

Response stage	Task	r	df	t
Initial response	BB	0.69	143	11.37*
	BR	0.67	140	10.76*
	SYL	0.65	145	10.44*
	CONJ	0.62	146	9.45*
Final response	BB	0.84	143	18.41*
	BR	0.65	140	10.14*
	SYL	0.71	145	12.09*
	CONJ	0.68	146	11.29*

Note. BB = Bat-and-ball; BR = Base-rates; SYL = Syllogisms; CONJ = Conjunction Fallacies.

***** $p < .001$.

B. Conflict Detection

As it can be seen in Figure S1 (note that negative values point to an overall successful conflict sensitivity) participants detected the conflict of their answers both at the initial and the final response stages, both at session 1 (initial: M = –7.0, SD = 8.6; final : M = –6.7, SD = 8.4) and session 2 (initial: M = –3.7, SD = 6.5; final: M = –3.6, SD = 6.2). The overall individual conflict detection at session 1 was significantly correlated with that of session 2 at the initial responses (r = 0.32, $t(149)$ = 4.08, p < .001), but not at the final responses (r = 0.27, $t(149)$ = 3.41, p < .001).

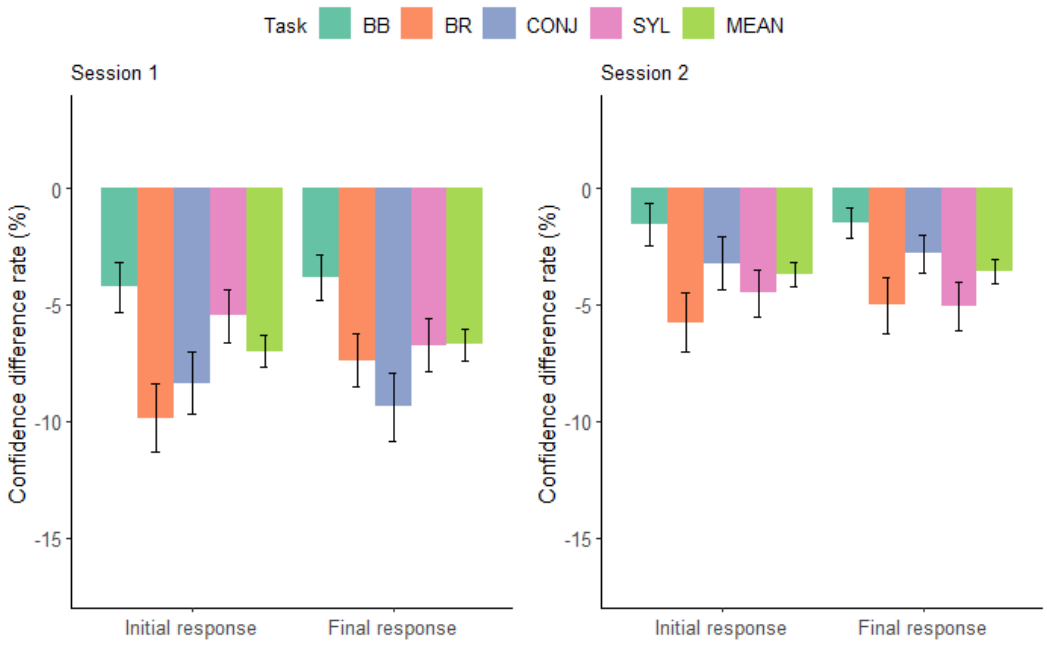

Figure S1. Confidence difference rates (%) between the conflict trials and the correct no-conflict trials (i.e., Confidence $_{conflict}$ – Confidence $_{no\text{-}conflict\_correct}$), separately for each session, each response stage, each reasoning task and for the composite mean across the four tasks. Negative values point to an overall successful conflict sensitivity. The error bars represent the Standard Error of the Mean. BB = Bat-and-ball; BR = Base-rates; CONJ = Conjunction Fallacies; SYL = Syllogisms; MEAN = the composite mean across the four tasks.

C. (Predictive) Conflict Detection on Incorrect Conflict trials

For completeness, in this section we re-ran the conflict detection and predictive conflict detection analyses by discarding the correct conflict trials when calculating conflict detection (i.e., conflict detection = Confidence $_{conflict\_incorrect}$ – Confidence $_{no\text{-}conflict\_correct}$). Due to the exclusion of incorrect conflict trials, we could only focus on the "00" and "01" directions.

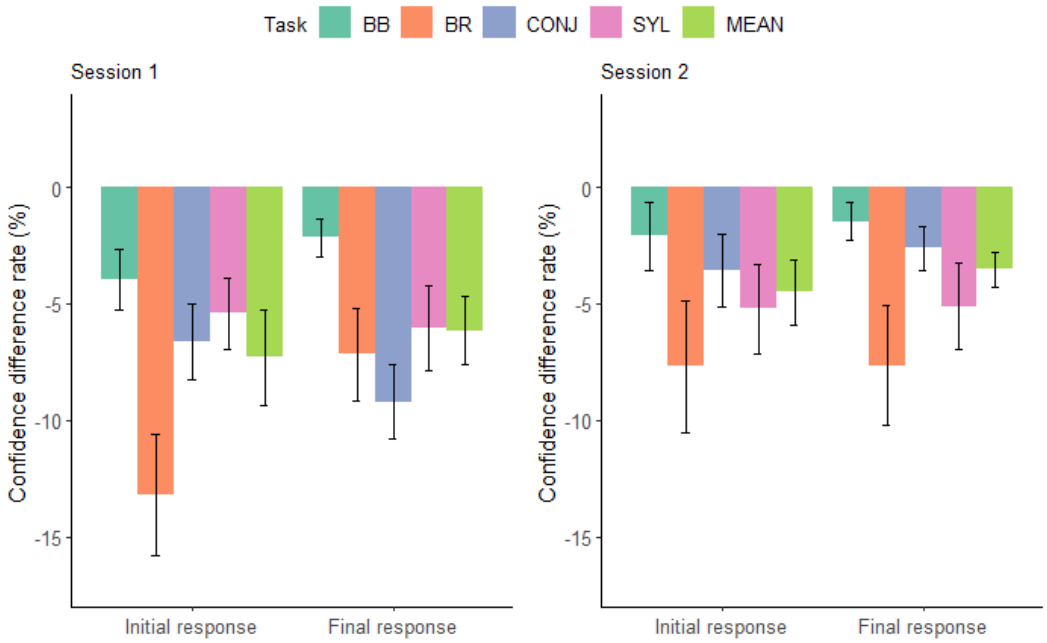

Figure S2. Confidence difference rates (%) between the incorrect conflict trials and the correct no-conflict trials, separately for each session, each response stage, each reasoning task and for the composite mean across the four tasks. Negative values point to an overall successful conflict sensitivity. The error bars represent the Standard Error of the Mean. BB = Bat-and-ball; BR = Base-rates; CONJ = Conjunction Fallacies; SYL = Syllogisms; MEAN = the composite mean across the four tasks.

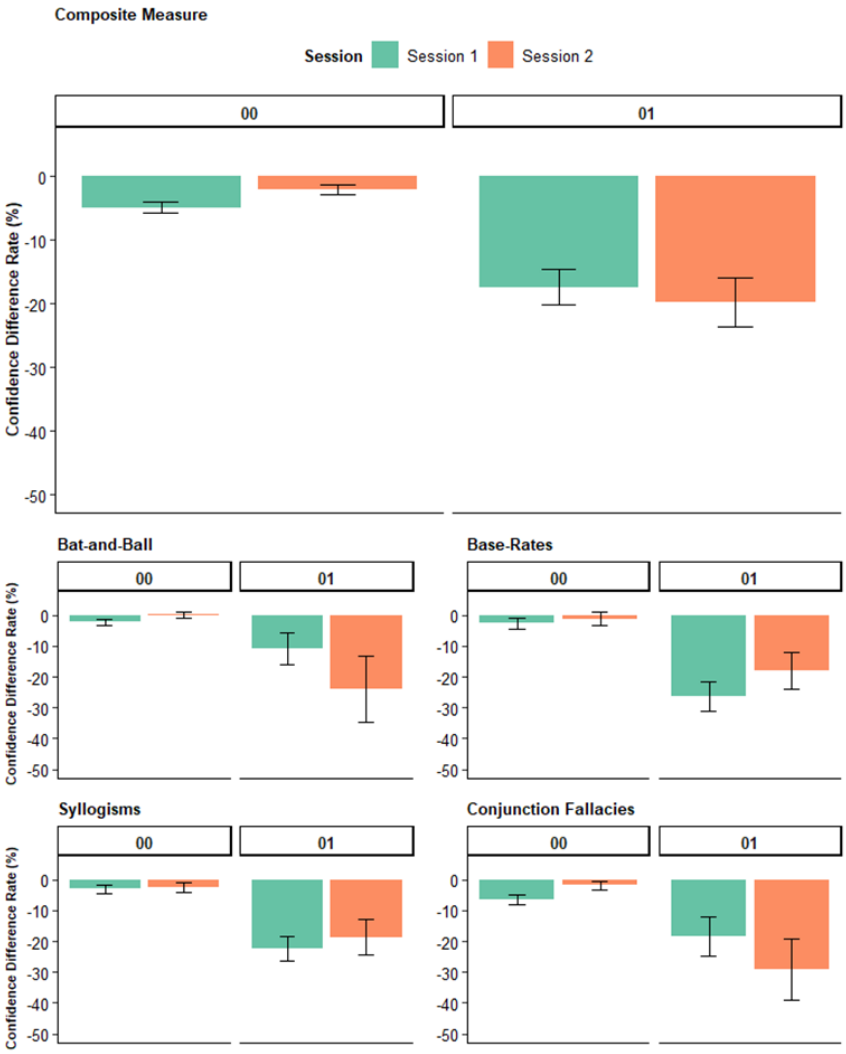

Figure S3. The mean confidence difference rate (%) according to the direction of change category (i.e., "01" trials represent the "change" category, "00" trials represent the "no change" category), separately for each session, each reasoning task and the composite measure across the four reasoning tasks. Negative values point to an overall successful conflict sensitivity. The error bars represent the Standard Error of the Mean. BB = Bat-and-ball; BR = Base-rates; CONJ = Conjunction Fallacies; SYL = Syllogisms; MEAN = the composite mean across the four tasks.

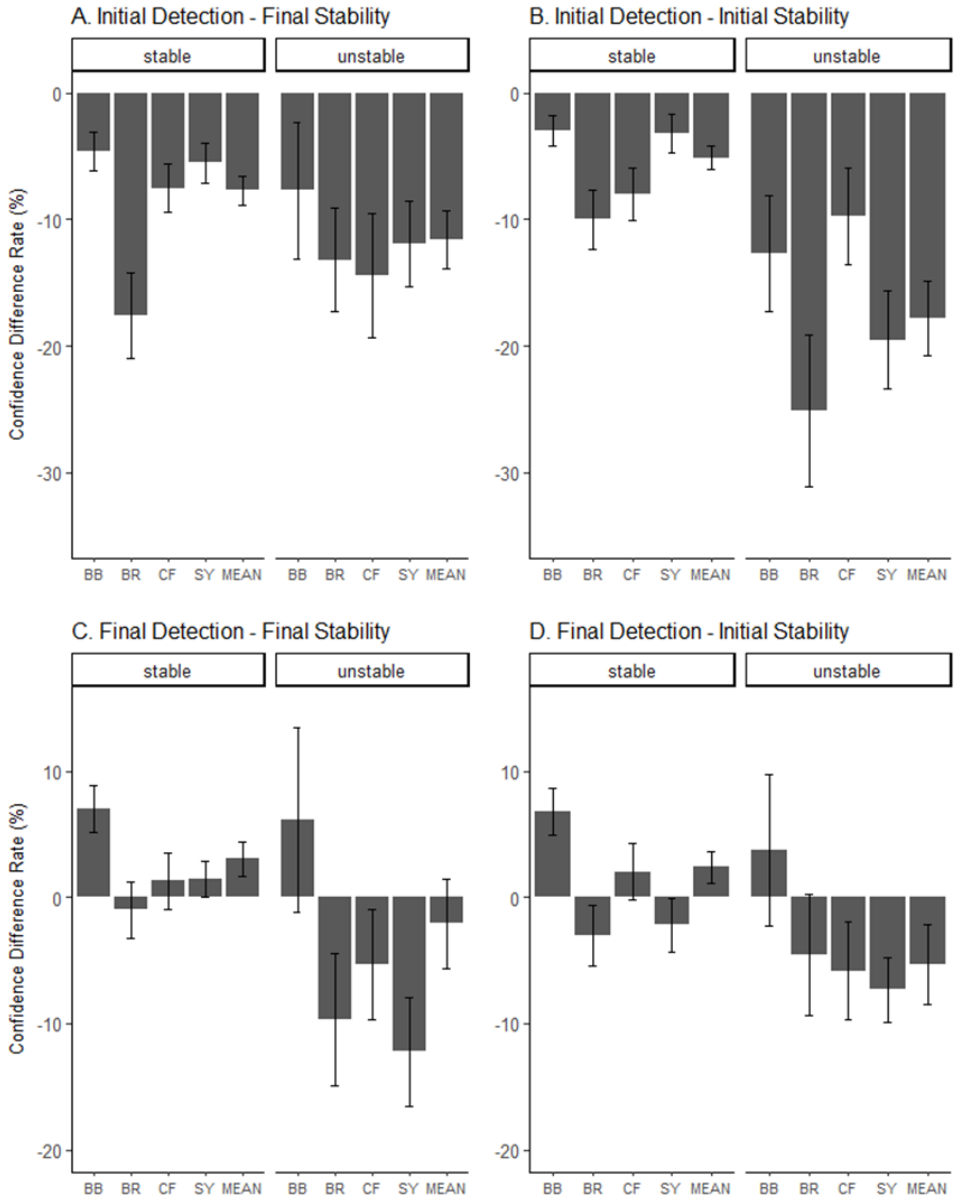

Figure S4. The (initial and final) conflict detection (i.e., Confidence $_{conflict\_incorrect}$ – Confidence $_{no\text{-}conflict\_correct}$) grand means according to stability (stable; unstable). Negative values point to an overall successful conflict sensitivity. Panel A shows the average initial conflict detection according to the stability of the final responses, Panel B shows the average initial conflict detection according to the initial responses' stability, Panel C shows the average final conflict detection according to the final responses' stability, and Panel D shows the average final conflict detection according to the initial responses' stability, separately for each reasoning task and for the composite mean across the four tasks. The error bars represent the Standard Error of the Mean. BB = Bat-and-ball; BR = Base-rates; CF = Conjunction Fallacies; SY = Syllogisms; MEAN = the composite mean across the four tasks.

Table S2.

Paired-samples t-tests between the mean conflict detection of the stable items and the mean conflict detection of the unstable items of each individual.

		Mean (SD) stable	Mean (SD) unstable	t	df
Initial detection	Final stability	–5.6 (13.8)	–11.6 (22.4)	2.36*	87
	Initial stability	–3.8 (10.3)	–14.6 (25.9)	3.95***	94
Final detection	Final stability	4.8 (14.8)	–6.5 (21.1)	3.88***	76
	Initial stability	3.7 (17.6)	–3.5 (19.5)	2.71**	89

***** $p < .05$.

****** $p < .01$.

******* $p < .001$.

D. (Predictive) Confidence Values

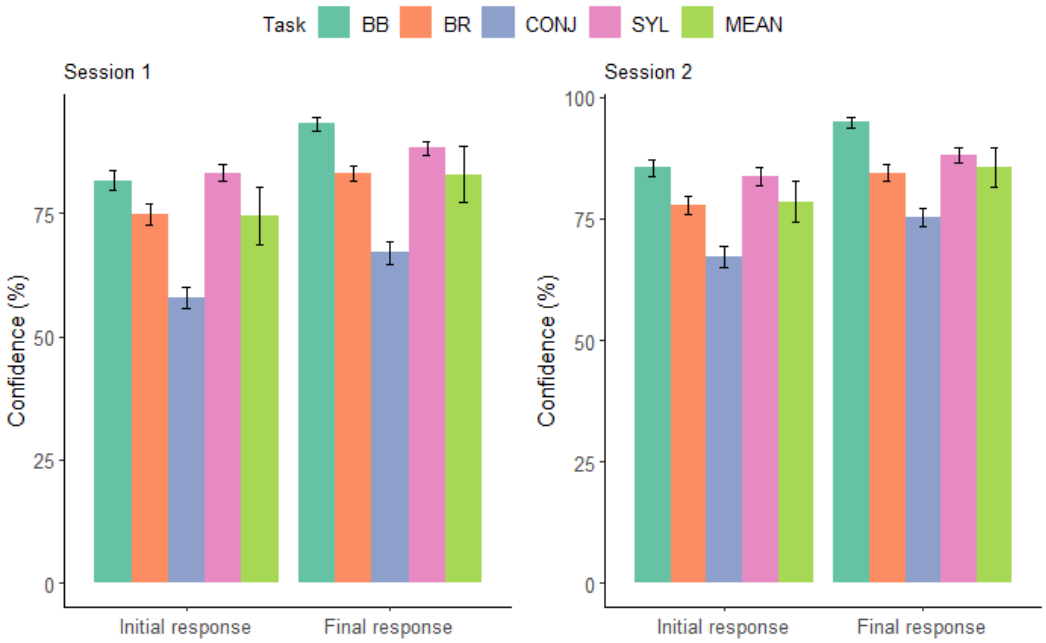

Figure S5. Confidence rates (%) at the conflict trials, separately for each session, each response stage, each reasoning task and for the composite mean across the four tasks. The error bars represent the Standard Error of the Mean. BB = Bat-and-ball; BR = Base-rates; CONJ = Conjunction Fallacies; SYL = Syllogisms; MEAN = the composite mean across the four tasks.

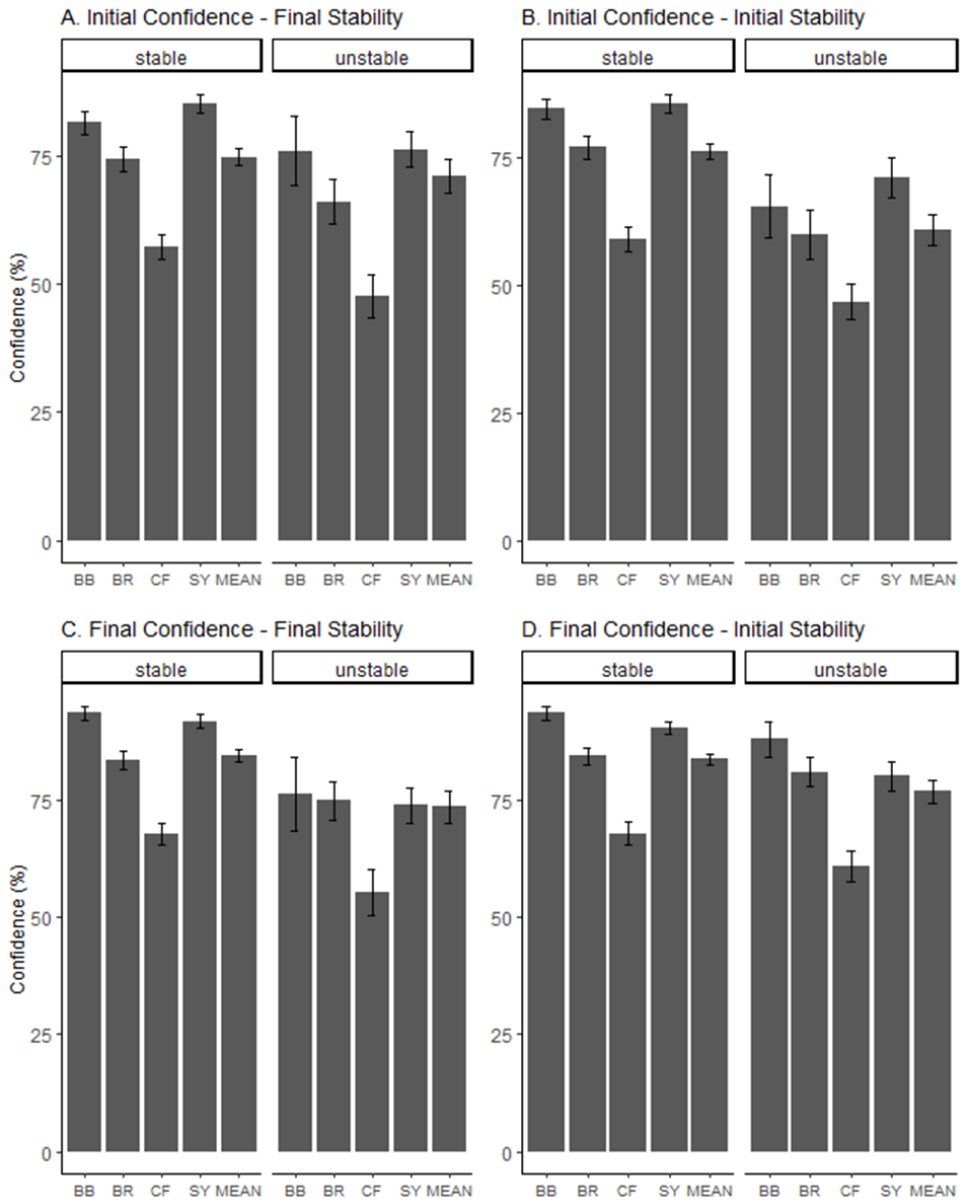

Figure S6. The (initial and final) confidence grand means according to stability (stable; unstable). Panel A shows the average initial confidence according to the stability of the final responses, Panel B shows the average initial confidence according to the initial responses' stability, Panel C shows the average final confidence according to the final responses' stability, and Panel D shows the average final confidence according to the initial responses' stability, separately for each reasoning task and for the composite mean across the four tasks. The error bars represent the Standard Error of the Mean. BB = Bat-and-ball; BR = Base-rates; CF = Conjunction Fallacies; SY = Syllogisms; MEAN = the composite mean across the four tasks.

Table S3.

Paired-samples t-tests between the mean confidence of the stable items and the mean confidence of the unstable items of each individual.

		Mean (SD) stable	Mean (SD) unstable	t	df
Initial confidence	Final stability	76.6 (24.6)	65.1 (29.4)	4.47*	115
	Initial stability	78.4 (22.3)	60.2 (31.1)	6.80*	123
Final confidence	Final stability	87.1 (18.9)	69.1 (30.1)	6.88*	115
	Initial stability	87.4 (18.1)	76.6 (24.8)	5.98*	123

***** $p < .001$.

E. Predictive Conflict Detection of Final Responses

Final Detection and Final Stability

By calculating the grand mean of conflict detection at the final responses, we found that there was a conflict detection effect for the items that had unstable final responses (M = –7.5, SD = 20.4), but a lack of conflict detection effect for the items with stable final responses (M = 2.3, SD = 12.8), as indicated by the positive confidence difference between conflict and no-conflict trials. As Figure S7A shows, this trend is observed in most individual reasoning tasks. To test the statistical significance of these results we compared participants' composite (final) conflict detection index at their stable and at their unstable items. Evidently, we only included the subjects that had both stable and unstable items (N = 114). Any participants with solely stable items were discarded from this analysis (there were no participants with only unstable items). A paired-samples t-test revealed a significant difference in the final conflict detection indices between stable (M = 3.3, SD = 14.1) and unstable (M = –9.1, SD = 24.5) items; $t(113) = 4.89$, p <.001. As expected, the unstable items had a higher conflict detection compared to the stable ones. It is worth noting that participants with only stable items (N = 37), did not show a conflict detection effect (M = 3.6, SD = 6.4).

Final Detection and Initial Stability

By calculating the grand mean of conflict detection at the final response, we found that there was a conflict detection effect for the items that had unstable initial responses (M = –2.3 , SD = 15.8), but no conflict detection effect for the items that had stable initial responses (M = 1.4, SD = 11.9). As Figure S7B shows, this trend is observed in most individual reasoning tasks. To test the statistical significance of these results we compared participants' composite conflict detection index at their stable and at their unstable items. Again, we only included the subjects that had both stable and unstable items (N = 122). Any participants with solely stable items were discarded from this analysis (there were no participants with only unstable items). A paired-samples t-test revealed a significant difference in the conflict detection indices between stable (M = 3.5, SD = 10.9) and unstable (M = –3.2, SD = 18.6) items; $t(121) = 4.09$, p <.001. As expected, the unstable items had a higher conflict detection compared to the

stable ones. Like in the above analysis, participants with only stable items (N = 29), did not show a conflict detection effect (*M* = 2.3, *SD* = 6.9).

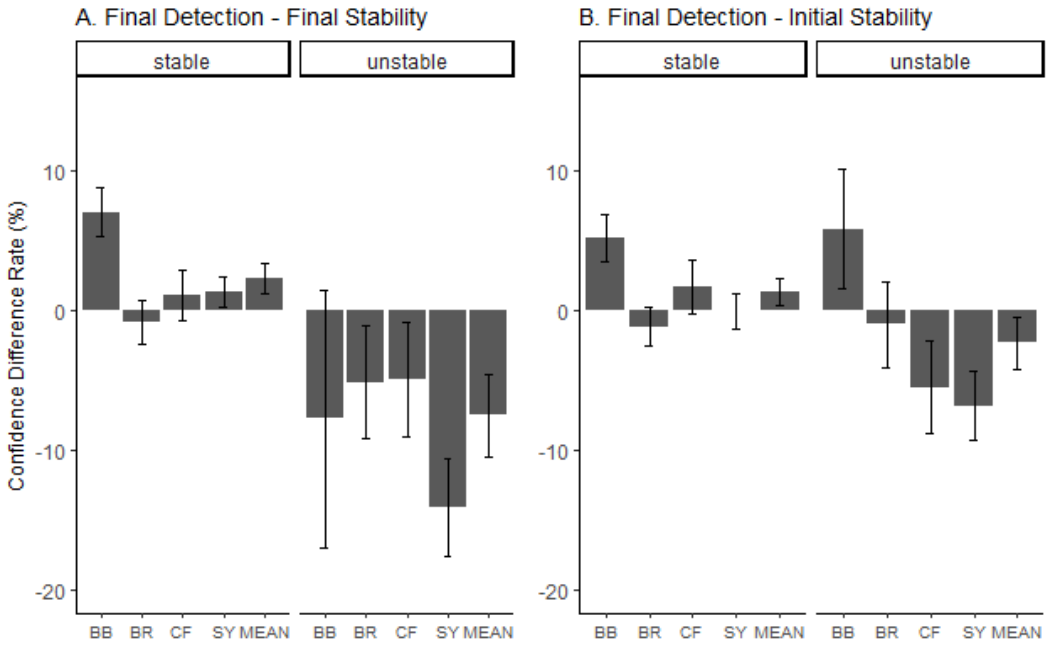

Figure S7. The grand means of the final conflict detection index (i.e., Confidence $_{conflict}$ – Confidence $_{no\text{-}conflict\_correct}$) according to stability (stable; unstable). Negative values point to an overall successful conflict sensitivity. Panel A shows the average final conflict detection according to the stability of the final responses and Panel B shows the average final conflict detection according to the stability of the initial responses, separately for each reasoning task and for the composite mean across the four tasks. The error bars represent the Standard Error of the Mean. BB = Bat-and-ball; BR = Base-rates; CF = Conjunction Fallacies; SY = Syllogisms; MEAN = the composite mean across the four tasks.

Résumé de la thèse

Contexte théorique

La réflexion caractérise l'expérience humaine, nous guidant constamment à travers une myriade de choix. Parfois, la réflexion nécessite du temps et des efforts pour parvenir à des solutions. Par exemple, avant de choisir un plan de crédit hypothécaire, il est probable que l'on étudie attentivement les options disponibles. À l'inverse, dans certaines situations, la réflexion peut se faire sans effort, comme lorsqu'il s'agit d'attraper ses clés avant de quitter la maison ou de résoudre des problèmes mathématiques de base tels que « 2 + 2 ». Cette dualité de la cognition humaine a conduit à l'idée qu'il existe deux modes de pensée distincts : l'un rapide, intuitif et sans effort, l'autre plus lent, réfléchi et exigeant davantage d'efforts (Frankish & Evans, 2009). La distinction entre un processus de pensée plus rapide et intuitif et un processus de pensée plus lent et délibéré a été au centre des théories duales de la cognition humaine, qui les ont popularisées en les appelant respectivement « Système 1 » et « Système 2 » (Epstein, 1994 ; Evans, 2008 ; Kahneman, 2011 ; Sloman, 1996).

L'influence des théories du double processus sur la pensée humaine a été considérable et appliquée à de nombreuses disciplines (Melnikoff & Bargh, 2018). Les théories du double processus ont été appliquées, entre autres, à la recherche sur les biais cognitifs et l'économie comportementale (Evans 2002 ; Kahneman, 2011), le jugement moral (Greene & Haidt, 2002), la coopération humaine (Rand et al., 2012), l'éducation (par exemple, Beaulac & Kenyon, 2018), la sensibilité aux fake news (Pennycook & Rand, 2019) et les algorithmes intelligents (Bonnefon & Rahwan, 2020). L'un des premiers domaines des sciences cognitives à populariser les modèles de double processus a été l'étude des biais dans le raisonnement logique (Kahneman 2000, 2011 ; Wason & Evans, 1975). Les études dans ce domaine ont montré que les individus enfreignent souvent les principes logico-mathématiques et probabilistes fondamentaux lorsqu'ils résolvent des tâches qui suscitent des réponses heuristiques en conflit avec ces principes. Un exemple célèbre de ces tâches d'heuristiques et de biais est le problème de Linda (Tversky & Kahneman, 1983) :

> Linda a 31 ans, est célibataire, extravertie et très brillante. Elle a obtenu une licence en philosophie. En tant qu'étudiante, elle était profondément préoccupée par les questions de discrimination et de justice sociale, et a également participé à des manifestations antinucléaires.
> Quelle affirmation est la plus probable ?
> a) Linda est banquière
> b) Linda est militante féministe et banquière

Alors que l'affirmation « Linda est banquière » ne correspond pas à la description stéréotypée de Linda, l'affirmation « Linda est militante féministe » s'aligne fortement sur la description. Ainsi, face à ce scénario, la majorité des raisonneurs utilisent le stéréotype comme un raccourci ou une « heuristique » et optent pour l'option b (Tversky & Kahneman, 1983). Cependant, selon la théorie des probabilités, la possibilité qu'un seul événement se produise est toujours plus élevée que la possibilité de la conjonction du même événement avec un autre, de sorte que l'énoncé unique (option a) est toujours le bon choix. Malgré la simplicité de cette règle, la majorité des individus l'ignorent et choisissent plutôt l'option qui correspond à l'heuristique stéréotypée (Tversky & Kahneman, 1983). Des résultats similaires ont été obtenus avec d'autres tâches d'heuristiques et de biais, qui suscitent des réponses heuristiques allant à l'encontre des principes logiques de base (Evans, 2008 ; Evans & Over, 1996 ; Kahneman & Frederick, 2002 ; Kahneman & Tversky, 1973). Ces résultats suggèrent que lorsque les individus sont confrontés à des problèmes qui appellent à la fois une réponse basée sur des règles logiques et une réponse heuristique convaincante, ils ont souvent tendance à négliger les principes logiques et à fournir des réponses biaisées basées sur l'heuristique (Kahneman, 2011).

Les théories du double processus offrent une explication élégante de ce phénomène de biais (Evans, 2008 ; Kahneman, 2011). Traditionnellement, ces théories postulent que pour surmonter les réponses heuristiques biaisées et prendre en compte les principes logiques, les individus doivent généralement s'engager dans une délibération qui demande un effort (Evans 2002, 2008 ; Evans & Over, 1996 ; Kahneman, 2011 ; Stanovich & West, 2000). Cependant, les raisonneurs humains sont des économes cognitifs qui préfèrent ne pas dépenser de temps et de ressources supplémentaires lorsqu'ils sont déjà parvenus à une réponse rapide et intuitive (Evans & Stanovich, 2013 ; Kahneman, 2011). Par conséquent, ils s'en tiennent souvent à leurs décisions intuitives, même si celles-

ci vont à l'encontre des principes logiques. Ils restent donc biaisés. Seuls les quelques raisonneurs qui disposent des ressources et de la motivation nécessaires pour s'engager dans une délibération et surmonter la réponse intuitive biaisée parviendront à fournir des réponses fondées sur la logique (Stanovich & West, 2000).

Il est important de souligner que les théories du double processus ne soutiennent pas que la pensée intuitive conduit toujours à des réponses biaisées ou qu'une délibération laborieuse garantit une réponse logique. Au contraire, les théoriciens du double processus se sont opposés à de telles simplifications (Evans, 2011 ; Evans & Stanovich, 2013). Par exemple, il est universellement reconnu que les adultes instruits peuvent résoudre avec précision le problème « 2 + 2 » sans s'engager dans une délibération ou que, même après une réflexion approfondie, le raisonneur moyen ne parviendra pas à résoudre un problème mathématique très complexe concernant, par exemple, les équations de la physique nucléaire. Les théories du double processus du raisonnement logique se concentrent plutôt sur des scénarios spécifiques dans lesquels le traitement intuitif et le traitement délibéré d'un problème sont supposés produire des réponses contradictoires (Frederick, 2005).

Des exemples de ces scénarios sont présentés dans le Cognitive Reflection Test (CRT), un groupe de problèmes dont la solution peut être facilement calculée, mais qui suscitent également des réponses heuristiques convaincantes qui contredisent les normes logiques (Frederick, 2005). L'exemple le plus célèbre du CRT est le problème de la batte et de la balle : « Une batte et une balle coûtent ensemble 1,10 $. La batte coûte 1 $ de plus que la balle. Combien coûte la balle ? ». Face à ce problème, la plupart des individus répondent rapidement que la balle coûte 10 centimes. Cependant, après réflexion et en résolvant l'équation « X+Y=1.10, Y=1+X, Résoudre pour X », il devient clair que la balle coûte en fait 5 centimes et que la batte, qui coûte 1 dollar de plus, coûte 1,05 dollars. Ces tâches d'heuristiques et de biais, telles que le problème de la batte et de la balle et le problème de Linda présentés ci-dessus, sont conçues pour créer systématiquement un conflit entre les réponses heuristiques et les réponses logiques. C'est dans ce contexte que l'on estime que le fait de surmonter les réponses heuristiques biaisées nécessite un effort de délibération.

Cependant, contrairement aux hypothèses des modèles traditionnels de double processus, des études récentes dans le domaine du raisonnement logique ont révélé que les réponses qui étaient auparavant considérées comme nécessitant une délibération peuvent également être traitées intuitivement (De Neys & Pennycook, 2019). Plus précisément, même lorsque la délibération est « éliminée » par des manipulations de contraintes expérimentales, la présumée réponse logique et délibérée est toujours observée. Les preuves directes à l'appui de cette affirmation proviennent d'études qui ont adopté de nouveaux paradigmes expérimentaux et, plus particulièrement, le paradigme à deux réponses (De Neys & Pennycook, 2019 ; Thompson et al., 2011). Dans ce paradigme, les participants sont généralement invités à donner deux réponses consécutives à des problèmes d'heuristiques et de biais. Dans un premier temps, il leur est demandé de donner la première réponse qui leur vient à l'esprit le plus rapidement possible. Immédiatement après, on leur montre à nouveau le même problème et on leur demande de prendre tout le temps qu'ils souhaitent pour y réfléchir avant de fournir leur réponse finale. Pour s'assurer que la réponse initiale est fournie de manière intuitive, sans délibération, certaines études imposent une pression temporelle et/ou des contraintes de charge cognitive au cours de la phase initiale du paradigme (Bago & De Neys, 2017 ; Newman et al., 2017). La raison derrière cela est la suivante : étant donné que la délibération nécessite du temps et des ressources cognitives pour fonctionner, en limitant ces deux éléments, les individus sont contraints au maximum de répondre de manière intuitive (Bago & De Neys, 2017). Par conséquent, ce paradigme permet aux chercheurs d'examiner séparément la nature des réponses plus intuitives et délibérées (Bago & De Neys, 2017, 2020 ; Raoelison et al., 2020 ; Thompson et al., 2011).

Selon les modèles traditionnels du double processus, les réponses intuitives initiales dans le paradigme à deux réponses devraient être biaisées, puisque les individus devraient être influencés par les indices heuristiques du problème. Cependant, contrairement aux hypothèses traditionnelles du double processus, les résultats du paradigme à deux réponses montrent que lorsque les individus parviennent à fournir une réponse logique à l'étape finale, après avoir délibéré, ils sont souvent parvenus à la même réponse dès l'étape initiale, intuitive (par exemple, Bago & De Neys, 2017, 2019a). Cela suggère que dans les tâches d'heuristiques et de biais, la délibération n'est pas toujours nécessaire pour passer outre les réponses intuitives et parvenir à une réponse logique, car les réponses

intuitives peuvent déjà être logiques. Ces résultats remettent en question les conceptions conventionnelles de l'intuition et de la délibération proposées par les théories du double processus, et montrent que la présumée réponse délibérée peut également être guidée par l'intuition. Il est intéressant de noter que des résultats similaires ont également été trouvés au-delà du raisonnement logique, dans les domaines de la prise de décision morale (Bago & De Neys, 2019b ; Vega et al., 2021) et prosociale (Bago et al., 2021 ; Kessler et al., 2017). Dans ces domaines, il est également prouvé que la présumée réponse délibérée (c'est-à-dire la décision morale utilitaire ou le choix prosocial égoïste) peut souvent être générée par un simple traitement intuitif.

Des études utilisant le paradigme de détection des conflits (De Neys & Pennycook, 2019) apportent un soutien supplémentaire au traitement intuitif des normes logiques. Ce paradigme se concentre sur les cas où les individus restent biaisés ; lorsqu'ils fournissent des réponses qui entrent en conflit avec les principes logiques du problème. Les études utilisant ce paradigme opposent généralement les problèmes standard (appelés « problèmes conflits »), dans lesquels le traitement intuitif et le traitement délibéré donnent des réponses conflictuelles, aux problèmes de contrôle (appelés « problèmes non-conflits »), dans lesquels le traitement intuitif et le traitement délibéré sont censés générer la même réponse et où aucun conflit n'est créé entre les deux. Par exemple, la version de contrôle, non-conflit, du problème Linda d'introduction, comporterait les options de réponse suivantes : « a. Linda est militante féministe; b. Linda est militante féministe et banquière ». Ici, l'option a s'aligne à la fois sur la description stéréotypée de Linda et sur les principes probabilistes, puisqu'elle se réfère à un seul événement et non à la conjonction. Par conséquent, même en cas de réponse intuitive, la majorité des individus choisiraient généralement l'option a comme réponse.

Des études utilisant le paradigme de détection des conflits ont montré que les individus biaisés se montrent généralement sensibles au fait que leurs réponses entrent en conflit avec des principes logiques concurrents. Par exemple, ils ont tendance à faire état d'une confiance moindre dans leurs réponses et de temps de réaction plus longs lorsqu'ils résolvent des problèmes conflit par rapport à leurs versions de contrôle, non-conflit (Bialek & De Neys, 2016 ; Frey et al., 2018 ; Gangemi et al., 2015 ; Mata, 2020 ; Srol & De Neys, 2019 ; Vartanian et al., 2018 ; voir De Neys, 2017 pour une revue, mais aussi Travers et al., 2016, ou Mata et al., 2017, pour les conclusions négatives). Étant donné que la seule

différence entre les deux versions est le conflit créé entre les indices logiques et heuristiques, ces résultats suggèrent que les individus traitent les indices logiques même lorsqu'ils fournissent des réponses heuristiques. Autrement dit, si les individus biaisés ignoraient complètement les indices logiques sous-jacents, leurs performances resteraient les mêmes dans les versions conflit et non-conflit du problème.

Cette incertitude concernant la réponse initiale, aussi appelée détection de conflit, persiste même lorsque les participants répondent intuitivement aux problèmes et que la délibération est minimisée grâce à des manipulations de charge cognitive et/ou de contrainte temporelle (Johnson et al., 2016 ; Pennycook et al., 2014 ; Thompson & Johnson, 2014). Plus précisément, même lorsque les individus fournissent une réponse biaisée au cours de la phase initiale et intuitive du paradigme à deux réponses, ils signalent généralement une diminution de la confiance dans la réponse par rapport à leur confiance de base (Bago & De Neys, 2017, 2019b ; Bialek & De Neys, 2017 ; Buric & Srol, 2020 ; Buric & Konradova, 2021). Cela indique que la sensibilité aux conflits fonctionne de manière plutôt automatique et fournit une preuve supplémentaire que les principes logiques peuvent être traités de manière intuitive.

En outre, la détection des conflits est également considérée comme un mécanisme qui influence le changement de réponse (De Neys, 2012 ; Pennycook et al., 2015 ; Purcell et al., 2023 ; Thompson et al., 2011). Les résultats issus du paradigme à deux réponses ont montré que les individus qui éprouvent plus de conflit dans leurs réponses intuitives initiales ont tendance à être plus enclins à réviser leurs réponses au cours de la phase délibérative finale (Bago & De Neys, 2017, 2020 ; Thompson & Johnson, 2014).

Sur la base des résultats ci-dessus, il apparaît clairement que, bien que les théories traditionnelles du double processus soutiennent que les réponses fondées sur des normes logiques ne peuvent généralement être déclenchées qu'après délibération, il n'existe pas de preuves empiriques solides pour étayer cette affirmation (De Neys, 2022). Pour résumer, deux résultats principaux contredisent cette hypothèse : premièrement, les réponses qui sont traditionnellement considérées comme survenant après la délibération sont également fournies intuitivement (Bago & De Neys, 2017, 2019a ; Newman et al., 2017 ; Raoelison & De Neys, 2019). Deuxièmement, même lorsque les raisonneurs fournissent des réponses biaisées, ils font toujours preuve de sensibilité aux indices logiques du

problème, et cette sensibilité opère souvent de manière intuitive (Bago & De Neys, 2017 ; Burič & Srol, 2020 ; Mata, 2020 ; Pennycook et al., 2014 ; Thompson & Johnson, 2014 ; mais voir aussi Mata et al., 2014, et Mata & Ferreira, 2018 pour les résultats négatifs). Il semble donc que le traitement intuitif n'entraîne pas seulement des réponses heuristiques, mais aussi des réponses logiques.

Cependant, cette reconceptualisation du traitement intuitif et délibéré n'implique pas que la délibération ne génère jamais de réponses logiques ou qu'elle n'est jamais nécessaire pour corriger nos intuitions. Au contraire, il est prouvé que lorsque les individus résolvent des tâches d'heuristiques et de biais, ils fournissent des réponses légèrement plus logiques après avoir délibéré et, dans les études utilisant le paradigme à deux réponses, la délibération est parfois nécessaire pour corriger les réponses intuitives biaisées (par exemple, Bago & De Neys, 2017). Le point essentiel ici est que le modèle délibératif correctif n'est pas aussi fréquent qu'on le supposait auparavant et que, le plus souvent, la présumée réponse délibérée est générée intuitivement.

Pour tenir compte de ces nouveaux résultats, les chercheurs ont introduit un modèle actualisé du double processus, parfois appelé théorie du double processus 2.0 (De Neys, 2017). Ce modèle affirme que la réponse qui a été traditionnellement considérée comme étant déclenchée par la délibération peut également être déclenchée intuitivement. Plus précisément, il propose que lorsqu'un individu traite intuitivement un problème de « biais », il génère plusieurs types d'intuitions qui entrent en concurrence les unes avec les autres. Deux intuitions principales entrent en jeu : l'une qui suscite une réponse heuristique (également appelée « intuition heuristique ») et l'autre qui suscite une réponse logique (également appelée « intuition logique »). Les « intuitions heuristiques » sont souvent basées sur des associations sémantiques stockées et contredisent les règles logiques, tandis que les « intuitions logiques » découlent d'une connaissance automatisée des principes mathématiques et probabilistes (De Neys, 2022 ; De Neys, 2012 ; Evans, 2019 ; Stanovich, 2018). L'intuition la plus forte et la plus activée finira par devenir la réponse intuitive sélectionnée. Lorsque les niveaux d'activation des intuitions concurrentes sont similaires, l'individu se sentira plus incertain, ou en conflit avec sa réponse. Cette incertitude peut susciter une délibération plus approfondie qui, à son tour, confirmera ou modifiera le choix intuitif (Pennycook et al., 2015). En revanche, si une intuition domine clairement

l'autre en termes de force, le raisonneur se sentira certain de son choix intuitif et l'intuition dominante conduira à une réponse sans délibération.

L'idée est que les « intuitions logiques » découlent d'un processus d'apprentissage et de pratique (Bago & De Neys, 2019a ; De Neys, 2012 ; Evans, 2019 ; Stanovich, 2018 ; Raoelison et al., 2021). Plus précisément, bon nombre des principes logico-mathématiques utilisés dans les tâches d'heuristiques et de biais sont enseignés pendant la scolarité. Par conséquent, les personnes qui sont exposées à ces principes au fil des années développent la capacité de les mettre en pratique de manière automatique (De Neys, 2012 ; Stanovich, 2018). Cela correspond à la façon dont les experts trouvent des problèmes complexes plus faciles à résoudre que les novices ; leur grande expérience leur permet de reconnaître immédiatement les schémas familiers. De la même manière, les personnes non spécialistes peuvent développer une familiarité similaire avec les concepts mathématiques et probabilistes fondamentaux.

Comme mentionné précédemment, les preuves de cette nouvelle caractérisation du raisonnement intuitif proviennent principalement de tâches classiques d'heuristiques et de biais, comme le problème de la batte et la balle (Raoelison & De Neys, 2019). Pourtant, il est important d'explorer si les résultats actuels s'étendent au-delà des tâches de raisonnement logique, une préoccupation récemment soulevée par divers chercheurs dans le domaine (e.g., March et al., 2023). En effet, bien que les modèles de double processus « rapide et lent » aient été principalement popularisés pour rendre compte des résultats dans le domaine d'heuristiques et de biais, leurs idées de base ont été appliquées dans de nombreux domaines (Melnikoff & Bargh, 2018). Si les nouvelles hypothèses centrales (révisées) doivent fournir quelque chose qui s'apparente à une théorie générale de la cognition (Reber & Allen, 2022), il est évident qu'il faut tester la généralisation des résultats centraux dans différents domaines. L'extension des résultats actuels au-delà du raisonnement logique a également des implications méthodologiques. Plus précisément, de nombreuses tâches cognitives ont été utilisées comme facteur prédictif des capacités de délibération (Frederick, 2005 ; Sirota et al., 2021). Toutefois, si l'on constate dans ces tâches que la présumée réponse délibérée peut également être déclenchée de manière intuitive, cela remettrait en cause leur utilisation en tant que facteur prédictif. Enfin, l'exploration de la logique intuitive au-delà des tâches d'heuristiques et de biais peut donner

des indications sur la conception d'interventions et de politiques visant à atténuer les biais dans chaque domaine.

La présente thèse aborde ces questions et cherche à mieux comprendre l'interaction entre l'intuition et la délibération, à travers deux dimensions clés. L'axe principal 1 vise à vérifier si les preuves d'une réponse intuitive correcte (chapitres 1, 2, 3) et d'une sensibilité aux conflits (chapitres 1 et 2) s'étendent au-delà des tâches classiques d'heuristiques et de biais. Pour ce faire, trois grands domaines sont explorés dans lesquels une réponse correcte est traditionnellement considérée comme le résultat d'un effort de délibération : la prise de décision en situation de risque, les tâches de traitement sémantique de haut niveau et les tâches de contrôle cognitif de bas niveau. Un axe 2 supplémentaire de cette thèse se concentre sur la stabilité des réponses aux tâches classiques d'heuristiques et de biais dans le temps, ainsi que sur l'impact à long terme de la sensibilité au conflit sur le changement de réponse (chapitre 4). Il vise ainsi à généraliser, sur une fenêtre temporelle plus étendue, les conclusions antérieures sur la détection des conflits dans le domaine du raisonnement.

Méthodes

Tout au long des chapitres de cette thèse, les participants ont été invités à résoudre différentes tâches. Dans chaque tâche, les participants devaient résoudre des problèmes conflits et non-conflits. Dans les problèmes conflits, le traitement intuitif et le traitement délibéré suscitent des réponses contradictoires. A l'inverse, dans les problèmes non-conflits, le traitement intuitif et le traitement délibéré génèrent tous deux la même réponse, et il n'y a donc pas de conflit entre les deux. Les problèmes non-conflits sont en fait des problèmes de contrôle ; si les participants s'abstiennent de deviner au hasard et accordent une attention minimale à la tâche, leurs performances devraient être maximales dans ces problèmes (Bago & De Neys, 2019a). Les types de tâches utilisées dans cette thèse sont brièvement décrits ci-dessous :

Tâches de choix risqué

Jeu de pari

Dans le chapitre 1, nous avons utilisé un jeu de pari basé sur la tâche d'aversion aux pertes de Keysar et al. (2012). Les participants se voyaient

proposer des paris qui pouvaient se traduire par un gain ou par une perte. Chaque pari indiquait la probabilité de gagner une certaine somme d'argent (par exemple, 5 % de probabilité de gagner 110 euros) et la probabilité de perdre une certaine somme d'argent (par exemple, 95 % de probabilité de perdre 5 euros), et les participants devaient dire s'ils voulaient prendre le pari ou non. Chaque pari comportait une version conflit et une version non-conflit, dont des exemples sont présentés ci-dessous :

Version conflit	Version non-conflit
Si vous prenez ce pari, vous avez :	*Si vous prenez ce pari, vous avez :*
5%** de probabilité de **GAGNER 110€	***95%** de probabilité de **GAGNER 110€***
95%** de probabilité de **PERDRE 5€	***5%** de probabilité de **PERDRE 5€***
Acceptez-vous le pari ?	*Acceptez-vous le pari ?*
○ *Oui*	○ *Oui*
○ *Non*	○ *Non*

Les paris conflits avaient une espérance mathématique positive et une forte probabilité de perdre de l'argent. Par conséquent, un conflit a été créé entre le fait d'éviter une perte potentielle (c'est-à-dire en ne prenant pas le pari) et le fait de prendre un risque afin d'obtenir un gain potentiel plus important (c'est-à-dire en prenant le pari). Il convient de noter que les instructions insistent sur le fait que l'objectif est de réaliser le plus de bénéfices possible. Ainsi, selon un calcul objectif des résultats et en l'absence d'un biais d'aversion aux pertes, les participants devraient toujours prendre le pari (à une espérance positive).

Les paris non-conflits avaient une espérance positive et une faible probabilité de perdre de l'argent. Ces paris ont été construits en inversant le P_{win} et le P_{lose} des items conflits, tout en gardant les valeurs identiques. Ainsi, dans les paris non-conflits, les participants avaient toujours une très forte probabilité de gagner une somme importante et une très faible probabilité de perdre une petite somme. Par conséquent, les items ne devraient pas (ou très peu) susciter d'aversion aux pertes et ne devraient pas (ou très peu) créer de conflit avec les considérations relatives à l'espérance. Si les individus s'abstenaient de deviner au hasard, nous nous attendions à une forte préférence pour le choix maximisant l'espérance dans ces items.

Jeu de loterie

Au chapitre 1, nous avons également utilisé un jeu de loterie qui était une variante de la tâche de choix de loterie de Holt-Laury (Holt & Laury, 2002). Les participants se voyaient présenter deux paires de loteries et devaient choisir une loterie dans chaque paire (loterie A ou loterie B). Les deux loteries (A et B) comportaient une forte probabilité de gagner une grosse somme d'argent (par exemple, 70 % de probabilité de gagner 350 euros) et une faible probabilité de gagner une petite somme d'argent (par exemple, 30 % de probabilité de gagner 10 euros). Dans chaque paire de loteries, la loterie A et la loterie B avaient les mêmes probabilités de gagner une grande et une petite somme d'argent. Un exemple de problème est présenté ci-dessous :

Version conflit

Loterie A	*Loterie B*
70% *de probabilité de* ***gagner 350€***	***70%*** *de probabilité de* ***gagner 230€***
30% *de probabilité de* ***gagner 10€***	***30%*** *de probabilité de* ***gagner 160€***

Quelle loterie choisissez-vous ?

- *A*
- *B*

Version non-conflit

Loterie A	*Loterie B*
70% *de probabilité de* ***gagner 10€***	***70%*** *de probabilité de* ***gagner 160€***
30% *de probabilité de* ***gagner 35€***	***30%*** *de probabilité de* ***gagner 230€***

Quelle loterie choisissez-vous ?

- *A*
- *B*

Dans les paires de loteries conflits, l'une des loteries avait toujours l'espérance mathématique la plus élevée de l'ensemble, tandis que l'autre loterie avait une espérance plus faible mais le gain minimal garanti le plus élevé. Par conséquent, un conflit s'est créé entre le choix d'une loterie avec un gain potentiellement important mais incertain, et une loterie avec un gain plus faible mais plus certain. Dans l'exemple ci-dessus, la loterie A a une espérance plus

élevée, mais ne garantit qu'un gain de 10 euros, tandis que la loterie B a une espérance plus faible, mais garantit un gain de 160 euros. Il est à noter que les participants ont reçu pour instruction d'essayer de faire le plus de profit possible. Ainsi, selon un calcul objectif des résultats et en l'absence d'un biais d'aversion aux pertes, ils devraient toujours choisir la loterie dont l'espérance est la plus élevée.

Dans les paires de loteries non-conflits, l'une des loteries présentait toujours à la fois l'espérance mathématique la plus élevée de l'ensemble et le gain garanti le plus élevé. Par conséquent, aucun conflit n'a été créé ; les participants étaient toujours censés préférer l'une des deux loteries, à la fois en termes de certitude et de gain potentiel. Les paires non-conflits ont été construites en inversant le P_{large} et le P_{small} dans chaque loterie, tout en gardant les valeurs identiques. Dans l'exemple ci-dessus, nous nous attendons à ce que les participants choisissent la loterie B.

Tâche d'illusion sémantique

Au chapitre 2, nous avons utilisé une tâche d'illusion sémantique comprenant des problèmes de culture générale adaptés de Speckmann et Unkelbach (2021 ; expérience 2). Pour chacune des questions sélectionnées, nous avons créé une version conflit (également appelée « anomalie ») et une version non-conflit (également appelée « sans anomalie »), qui sont présentées ci-dessous :

Version conflit « anomalie »	Version non-conflit « sans anomalie »
Dans le récit biblique, combien d'animaux de chaque espèce Moïse a-t-il emportés dans l'arche ? ○ *Deux* ○ *Trois* ○ *Cette question ne trouve pas de réponse dans ce formulaire* ○ *Ne sait pas*	*Dans le récit biblique, combien d'animaux de chaque espèce Noé a-t-il emportés dans l'arche ?* ○ *Deux* ○ *Trois* ○ *Cette question ne trouve pas de réponse dans ce formulaire* ○ *Ne sait pas*

La version non-conflit utilisait le mot original, non déformé, de la question de culture générale (par exemple, « Noé »), tandis que la version conflit utilisait

un mot « imposteur » sémantiquement lié (par exemple, « Moïse »). De la même manière que Speckmann et Unkelbach (2021), chaque question comportait quatre options de réponse différentes. La première option était la réponse « non déformée » (par exemple, « deux » pour la question sur Moïse) et pouvait être correcte ou incorrecte selon la version de la question (conflit ou non-conflit). La deuxième option (par exemple, « trois ») était toujours incorrecte. La troisième option de réponse était «Cette question ne trouve pas de réponse dans ce formulaire », et pouvait être correcte ou incorrecte selon la version de la question (conflit ou non-conflit). La quatrième option était « Ne sait pas », qui était toujours codée comme incorrecte.

Tâches de contrôle cognitif

Tâche de Stroop

Au chapitre 3, nous avons utilisé une tâche de Stroop. Sur la base d'Aïte et al. (2016), nous avons créé seize stimuli mot-couleur en combinant quatre noms de couleur différents (« rouge », « vert », « bleu » et « jaune ») avec quatre couleurs d'encre correspondantes (codes de couleur RVB 255;0;0, 0;255;0, 0;0;255 et 255;255;0). La réponse considérée comme correcte était toujours celle qui correspondait à la couleur d'encre du mot. Avec ces stimuli, nous avons créé des essais conflits (également appelés « incongruents ») et non-conflits (également appelés « congruents »), comme ceux présentés ci-dessous :

Version conflit « incongruente »	Version non-conflit « congruente »
vert	**bleu**

Tâche de flanker

Au chapitre 3, nous avons également utilisé une tâche de Flanker. Les stimuli consistaient en une rangée de cinq flèches. Cette rangée comprenait une flèche centrale flanquée de deux flèches de chaque côté, toutes avec des pointes de flèche pointant vers la gauche ou vers la droite. Dans les stimuli non-conflits (également appelés « congruents »), les flèches environnantes pointaient dans la même direction que la flèche centrale, tandis que dans les stimuli conflits

(également appelés « incongruents »), les flèches environnantes pointaient dans la direction opposée à la flèche centrale. La réponse considérée comme correcte était celle qui identifiait la direction de la flèche centrale. Les stimuli sont présentés ci-dessous :

Version conflit « incongruente »	Version non-conflit « congruente »
←←→←← ou →→←→→	←←←←← ou →→→→→

Tâches d'heuristiques et de biais

Tâche de la balle et de la batte

Dans le chapitre 3, nous avons utilisé la tâche de la batte et de la balle avec des items tirés de Bago et De Neys (2019a). Dans le chapitre 4, nous avons utilisé une variante de cette tâche, qui incluait des montants au lieu de prix. Les éléments de cette tâche ont été tirés de Raoelison et De Neys (2019). Chaque item avait une version conflit et une version non-conflit. Des exemples de problèmes sont présentés ci-dessous :

	Version conflit	Version non-conflit
Prix	*Un crayon et une gomme coûtent au total 1,10 $.* *Le crayon coûte 1 $ de plus que la gomme.* *Combien coûte la gomme ?* ○ *5 cents* ○ *1 cent* ○ *10 cents* ○ *15 cents*	*Un crayon et une gomme coûtent au total 1,10 $.* *Le crayon coûte 1 $.* *Combien coûte la gomme ?* ○ *5 cents* ○ *1 cent* ○ *10 cents* ○ *15 cents*

Montants	*Un parc national compte 650 roses et fleurs de lotus au total.* *Il y a 600 roses de plus que de fleurs de lotus.* *Combien y a-t-il de fleurs de lotus ?* ○ *25* ○ *50*	*Un parc national compte 650 roses et fleurs de lotus au total.* *Il y a 600 roses.* *Combien y a-t-il de fleurs de lotus ?* ○ *25* ○ *50*

Les participants se voyaient toujours proposer quatre options de réponse : l'option logique (« 5 cents » dans la tâche de la batte et la balle originale), qui est la réponse correcte, l'option heuristique (« 10 cents » dans la tâche de la batte et la balle originale), et deux options de remplissage. Les deux options étaient toujours la somme de la réponse correcte et de la réponse heuristique (par exemple, « 15 cents » dans les unités originales de la batte et de la balle) et de leur deuxième plus grand diviseur commun (par exemple, « 1 cent » dans l'original). Les versions non-conflits ont été construites en supprimant l'énoncé « plus que » des versions non-conflits (« Un crayon et une gomme coûtent 1,10 $ au total. Le crayon coûte 1 dollar. Combien coûte la gomme ? »).

Tâche des taux de base

Dans les chapitres 3 et 4, nous avons utilisé la tâche de taux de base, et les items ont été tirés de Pennycook et al. (2014). Chaque problème se composait d'une phrase décrivant la composition d'un échantillon (par exemple, « Cette étude contient des hommes d'affaires et des pompiers »), d'une phrase contenant une description stéréotypée d'une personne aléatoire de l'échantillon (par exemple, « La personne 'K' est courageuse ») et d'une phrase contenant l'information sur le taux de base (par exemple, « Il y a 996 hommes d'affaires et 4 pompiers »). Les participants devaient ensuite choisir le groupe auquel la personne aléatoire appartenait le plus probablement. Les versions non-conflits ont été construites en inversant les taux de base des versions conflits. Par exemple, dans sa version non-conflit, la deuxième phrase du problème ci-dessus se lirait comme suit : « Il y a 4 hommes d'affaires et 996 pompiers ». L'option de réponse considérée comme correcte était toujours celle qui correspondait à la grande majorité des individus de l'échantillon. Un exemple de problème est présenté ci-dessous :

Version conflit	Version non-conflit
Cette étude comprend des hommes d'affaires et des pompiers. *La personne « K » est courageuse.* *Il y a 996 hommes d'affaires et 4 pompiers.* *La personne « K » a-t-elle plus de chances d'être :* ○ *Un homme d'affaires* ○ *Un pompier*	*Cette étude comprend des hommes d'affaires et des pompiers.* *La personne « K » est courageuse.* *Il y a 4 hommes d'affaires et 996 pompiers.* *La personne « K » a-t-elle plus de chances d'être :* ○ *Un homme d'affaires* ○ *o Un pompier*

Tâche de raisonnement syllogistique

Dans les chapitres 3 et 4, nous avons utilisé la tâche de raisonnement syllogistique et les items ont été tirés de Bago et De Neys (2017). Chaque problème se composait d'une prémisse majeure (par exemple, « Tous les fruits peuvent être mangés »), d'une prémisse mineure (par exemple, « Les fraises sont des fruits ») et d'une conclusion (par exemple, « Les fraises peuvent être mangées »). Les participants devaient toujours considérer les prémisses comme vraies et devaient dire si la conclusion découlait logiquement des prémisses ou non. Une conclusion n'était considérée comme logique que si elle était valide. Dans les problèmes conflits, la crédibilité et la validité des problèmes étaient en conflit, ce qui signifie qu'un syllogisme était soit valide et non crédible, soit invalide et crédible. Au contraire, dans les problèmes non-conflits, les syllogismes étaient soit valides et crédibles, soit non valides et non crédibles. Un exemple de ces problèmes est présenté ci-dessous :

Version conflit	Version non-conflit
Tous les fruits peuvent être mangés. *Les fraises peuvent être mangées.* *Les fraises sont des fruits.* *La conclusion est-elle logique ?* ○ *Oui* ○ *Non*	*Tous les fruits peuvent être mangés.* *Les fraises sont des fruits.* *Les fraises peuvent être mangées.* *La conclusion est-elle logique ?* ○ *Oui* ○ *Non*

Tâche de l'erreur de conjonction

Dans le chapitre 4, nous avons utilisé la tâche de l'erreur de conjonction, et les items ont été tirés de Frey et al. (2018), à l'exception d'un item, le problème de Linda, qui a été adapté du matériel de Tversky et Kahneman (1983). Chaque problème consistait en une description stéréotypée d'un individu, suivie de deux affirmations sur cet individu, et les participants devaient choisir l'affirmation la plus susceptible d'être vraie. La première option de réponse consistait en une seule affirmation liée à la personne (par exemple, « Jon joue dans un groupe de rock »), tandis que la deuxième option de réponse était une conjonction de la première affirmation avec une deuxième affirmation (par exemple, « Jon joue dans un groupe de rock et est comptable »). L'une des deux affirmations correspondait fortement à la description stéréotypée, tandis que la seconde y correspondait moins. Étant donné que la possibilité qu'un seul événement se produise est toujours plus élevée que la possibilité de la conjonction, l'affirmation unique a toujours été considérée comme le bon choix. Les versions non-conflits ont été créées en remplaçant l'option singulière par l'affirmation qui correspondait le mieux à la description stéréotypée. Un exemple de problème est présenté ci-dessous :

Version conflit	Version non-conflit
John a 32 ans. *Il est intelligent et ponctuel, mais sans imagination et quelque peu inerte.* *A l'école, il était fort en mathématiques mais faible en langues et en art.* *Quelle est l'affirmation la plus probable ?* ○ *John joue dans un groupe de rock* ○ *John joue dans un groupe de rock et est comptable*	*John a 32 ans.* *Il est intelligent et ponctuel, mais sans imagination et quelque peu inerte.* *A l'école, il était fort en mathématiques mais faible en langues et en art.* *Quelle est l'affirmation la plus probable ?* ○ *John est comptable* ○ *John joue dans un groupe de rock et est comptable*

Paradigme à deux réponses

Nous avons utilisé le paradigme à deux réponses (Thompson et al., 2011) pour la présentation de tous les items. Dans ce paradigme, les participants sont invités à fournir deux réponses consécutives à chaque essai. Le format du

paradigme est basé sur des études récentes dans lesquelles, pendant la réponse initiale, les participants sont invités à effectuer une tâche de mémorisation de la charge ainsi qu'à répondre dans un délai strict, qui est pré-testé pour être exigeant pour la tâche respective (par exemple, Bago & De Neys, 2017, 2019a ; Raoelison et al., 2020). Pendant la réponse finale, il n'y a pas de charge ni de délai. Ainsi, en limitant le temps de traitement et en ajoutant une charge de mémorisation au cours de la première étape, l'implication du système 2 est minimisée. Par conséquent, on peut être sûr que la réponse initiale est fournie intuitivement (c'est-à-dire sans délibération), tandis que dans la phase de réponse finale, les raisonneurs sont autorisés à délibérer.

Pour éviter toute confusion, il est important de souligner que le délai et la charge cognitive concurrente diffèrent en fonction de la tâche. L'objectif de ces deux contraintes est de minimiser l'implication de la délibération et de favoriser la pensée intuitive. Cependant, il n'existe pas de procédure standard qui puisse garantir que les individus répondront de manière intuitive, et la définition de « ressources cognitives limitées » dépend toujours de la tâche à accomplir. Par exemple, les tâches d'heuristiques et de biais sont longues (quelques phrases de préambule et la lecture des options de réponse), de sorte que les délais sont basés sur les temps de lecture moyens pré-testés qui sont généralement de quelques secondes (les participants doivent disposer du temps minimum pour lire le problème avant de répondre). Au contraire, la réponse aux tâches de Stroop et de Flanker est considérablement plus rapide puisque les participants ne voient qu'un seul stimulus (c'est-à-dire un mot ou une rangée de flèches), de sorte qu'un délai strict ne peut pas être plus long qu'une seconde. Il en va de même pour la charge cognitive, dont l'objectif est de solliciter les ressources cognitives des participants. La sollicitation des ressources peut dépendre de la nature spécifique de la tâche. C'est pourquoi, pour chacune de nos tâches, nous avons opté pour une charge dont il a été indépendamment démontré dans la littérature qu'elle sollicitait les ressources cognitives et diminuait les performances dans ce type de tâche spécifique.

Chapitres expérimentaux

Au chapitre 1, je me concentre sur la prise de décision en situation de risque. Bien que ce domaine soit au cœur de la théorie des perspectives

(Kahneman & Tversky, 1979), qui est centrale de la recherche d'heuristiques et de biais, il n'existe pas de preuves empiriques systématiques démontrant que la délibération est nécessaire pour prendre en compte les risques maximisant l'espérance mathématique (par exemple, Mechera-Ostrovsky et al., 2022). Lorsque les individus prennent des risques, ils sont souvent sensibles au biais d'aversion aux pertes, qui les amène à surestimer l'impact négatif des pertes par rapport à la perspective de gains potentiels comparables (Kahneman & Tversky, 1979). Selon les théories du double processus, la délibération est nécessaire pour surmonter ce biais et prendre en compte l'espérance mathématique d'un pari (Slovic et al., 2005). Cependant, il existe peu de preuves empiriques à l'appui de cette idée. Pour tester directement cette idée, j'ai présenté aux participants des paris à espérance positive en utilisant le paradigme à deux réponses (Thompson et al., 2011), et ils devaient choisir entre une option sûre d'aversion aux pertes et une option risquée de maximisation de l'espérance. Les résultats montrent que, dans la plupart de leurs choix, les individus restent réticents aux pertes, tant après un simple traitement intuitif qu'après une délibération. Cependant, lorsqu'ils optaient pour le choix maximisant l'espérance après délibération, ils étaient souvent parvenus à cette réponse dès le stade initial, intuitif. En outre, même lorsque les individus avaient une aversion aux pertes, ils s'apercevaient souvent que leur réponse était en contradiction avec les principes de l'espérance du problème (comme le montre la diminution de la confiance). Ces résultats montrent que la délibération n'est pas la voie principale pour une réponse basée sur l'espérance mathématique dans la prise de décision risquée. La plupart du temps, lorsque les individus parviennent à prendre des risques maximisant l'espérance, ils le font en utilisant un simple traitement intuitif.

Après avoir démontré l'existence d'une réponse intuitive saine et d'une sensibilité au conflit dans les décisions risquées, j'étudie, au chapitre 2, la nature de la réponse correcte dans les tâches sémantiques de traitement du langage de haut niveau. Je me concentre plus particulièrement sur les illusions sémantiques, qui sont des tâches de rappel de mémoire ayant une solution correcte, mais qui en même temps déclenchent une réponse heuristique incorrecte (Erickson & Mattson, 1981). Prenons l'exemple de la question suivante : « Quel est le nom des courtisanes vêtues de kimonos qui divertissent les Chinois ? ». Face à cette question, la plupart des individus répondraient « Geisha », sans remarquer que les Geishas font partie de la culture japonaise et non chinoise. Les théories du

double processus attribuent ce biais à l'absence de traitement délibéré (Koriat, 2017). Selon ce point de vue, un traitement délibéré lent est nécessaire pour détecter les anomalies dans les phrases déformées et corriger les réponses intuitives superficielles. Néanmoins, les données disponibles ne nous permettent pas de savoir si la délibération est toujours nécessaire pour détecter les anomalies dans les phrases. Pour tester cette hypothèse, j'ai présenté à des participants des illusions sémantiques en utilisant le paradigme à deux réponses (Thompson et al., 2011). Les résultats indiquent que, le plus souvent, les individus doivent s'engager dans un traitement lent et délibéré pour surmonter l'illusion. Cependant, ils parviennent à fournir des réponses intuitives correctes dans un nombre non négligeable de cas. En outre, même lorsque les individus tombent dans l'illusion, ils sont sensibles au fait que leur réponse n'est pas entièrement justifiée, comme le montre la baisse de confiance. Les résultats du chapitre 2 révèlent que les réponses intuitives correctes ne se limitent pas aux tâches qui requièrent une compréhension de base des concepts mathématiques, mais qu'elles sont également présentes dans les tâches de compréhension linguistique de haut niveau.

Sur la base des résultats précédents, qui ont démontré l'existence d'une réponse intuitive correcte dans les tâches de haut niveau, j'explore au chapitre 3 la nature de la réponse correcte dans les tâches de contrôle cognitif de bas niveau. Ces tâches ont été utilisées pour exploiter directement les processus de contrôle de bas niveau, plutôt que les fonctions d'ordre supérieur telles que le raisonnement (Botvinick et al., 2001 ; Diamond, 2013). Elles demandent généralement aux individus d'inhiber une réponse puissante mais non pertinente pour la tâche et de choisir une réponse moins dominante. On estime que le fait de résister aux réponses automatiques et tentantes exige un traitement contrôlé et laborieux (Botvinick et al., 2001). En d'autres termes, le contrôle cognitif est supposé avoir un rôle correctif ; des réponses rapides et incorrectes sont automatiquement générées et sont ensuite corrigées par des processus plus lents et contrôlés (Botvinick et al., 2001). Ce schéma est similaire à celui proposé par les théories du double processus de raisonnement. Pour tester empiriquement le modèle correctif du contrôle délibéré dans les tâches de bas niveau, j'ai présenté aux participants deux des tâches de contrôle cognitif les plus couramment utilisées dans un format à deux réponses : la tâche de Stroop (Stroop, 1935) et la tâche de Flanker (Eriksen & Eriksen, 1974). L'objectif était de déterminer si les individus

pouvaient fournir des réponses précises lorsque le temps et les ressources cognitives étaient limités. Les résultats ont montré que les bonnes performances dans ces tâches sont dues à un traitement intuitif précis plutôt qu'à une correction lente et contrôlée de réponses automatiques erronées. Ainsi, tant dans la tâche de Stroop que dans la tâche de Flanker, des réponses correctes sont produites lorsque le contrôle délibéré est limité. Dans un deuxième temps, j'ai exploré le lien entre les performances intuitives dans la tâche de Stroop et dans les tâches d'heuristiques et de biais (Abreu-Mendoza et al., 2020 ; De Neys et al., 2011 ; Handley et al., 2004). Les résultats indiquent une corrélation faible, pour laquelle j'explore plusieurs explications théoriques.

En résumé, dans l'axe 1 de cette thèse (chapitres 1 à 3), j'ai trouvé des preuves que la présumée réponse délibérée peut souvent résulter d'un simple traitement intuitif, dans les domaines de la prise de décision en situation de risque, des tâches de non-raisonnement de haut niveau et des tâches de contrôle cognitif de bas niveau. En outre, dans les chapitres 1 et 2, j'ai montré que même lorsque les participants fournissaient une réponse biaisée et heuristique, ils pouvaient détecter que leur réponse était en conflit avec les éléments « logiques » corrects sous-jacents du problème. Il est important de noter que cette sensibilité à la bonne réponse n'a pas seulement été constatée pour les réponses délibérées, mais aussi pour les réponses intuitives, ce qui suggère que le mécanisme de détection des conflits peut fonctionner automatiquement dans ces domaines.

Dans le chapitre 4, je me concentre plus précisément sur la sensibilité aux conflits et j'examine son impact à long terme sur la stabilité de la performance dans les tâches d'heuristiques et de biais. Comme indiqué précédemment, des études antérieures sur deux réponses ont montré que la détection des conflits prédit le changement de réponse au niveau intra-essai ; plus un participant est en conflit avec sa réponse à l'étape initiale, plus il est susceptible de la changer à l'étape finale, lorsqu'il est autorisé à délibérer (Bago & De Neys 2017, 2019a ; Thompson & Johnson, 2014). Dans le chapitre 4, je vérifie si la sensibilité aux conflits peut également avoir un impact à long terme sur le changement de réponse. Pour ce faire, j'ai demandé aux participants de résoudre les mêmes tâches d'heuristiques et de biais deux fois lors de deux sessions de test, à deux semaines d'intervalle. J'ai utilisé le paradigme à deux réponses pour tester la stabilité des réponses intuitives initiales et des réponses délibérées finales. Tout d'abord, les résultats ont montré que les réponses des participants aux tâches

d'heuristiques et de biais sont très stables dans le temps (Stango & Zinman, 2020). Autrement dit, les participants ont rarement modifié leurs réponses intuitives et délibérées après le premier test. Toutefois, malgré cette grande stabilité, les réponses intuitives et délibérées présentaient encore une certaine variabilité au bout de deux semaines. Il est important de noter que cette variabilité à long terme n'était pas entièrement aléatoire, mais qu'elle pouvait être prédite par la détection des conflits. Plus les individus étaient en conflit avec leur réponse intuitive à un problème au cours de la première session de test, plus ils étaient susceptibles de modifier leur réponse (intuitive et délibérée) au même problème deux semaines plus tard.

En somme, l'axe 2 démontre, dans l'un des rares tests directs de la stabilité des tâches d'heuristiques et de biais (Białek & Pennycook, 2018 ; Stango & Zinman, 2020), que les biais individuels restent très stables au fil du temps. Surtout, il montre que la détection intuitive des conflits peut prédire la variabilité des réponses intuitives et délibérées au fil du temps. De cette manière, il confirme que le couplage entre la détection des conflits et le changement de réponse peut être généralisé sur une fenêtre temporelle plus longue, ce qui ouvre la voie à une nouvelle application intéressante du paradigme à deux réponses.

Dans l'ensemble, les résultats de cette thèse permettent de mieux comprendre la nature des réponses intuitives-automatiques au-delà des tâches d'heuristiques et de biais et suggèrent que, dans divers domaines, la prise de décision peut être mieux comprise comme une interaction entre diverses intuitions « rapides », plutôt que comme une dichotomie entre une pensée « rapide » et une pensée « lente ».

Conclusion

Cette thèse a démontré que le traitement intuitif constitue une voie fiable vers une réponse correcte dans un large éventail de domaines. En outre, les individus font preuve d'une sensibilité intuitive à leurs erreurs, et cette sensibilité sert de prédiction pour le changement de réponse à long terme. Bien que l'ampleur de la réponse intuitive correcte et de la sensibilité aux conflits varie en fonction des caractéristiques propres à la tâche, un ensemble cohérent de preuves valide leur présence. Ces résultats s'alignent sur des études antérieures dans le domaine du raisonnement logique et montrent que dans des domaines impliquant des choix

risqués, des associations sémantiques et un contrôle cognitif, les réponses traditionnellement considérées comme résultant d'une délibération peuvent également découler d'un simple traitement intuitif. Cela suggère que le rôle principal de la délibération n'est pas nécessairement de nature corrective et que l'intuition joue un rôle plus important qu'on ne le pensait pour arriver à une pensée saine.

References

Abreu-Mendoza, R. A., Coulanges, L., Ali, K., Powell, A. B., & Rosenberg-Lee, M. (2020). Children's discrete proportional reasoning is related to inhibitory control and enhanced by priming continuous representations. *Journal of Experimental Child Psychology*, *199*, 104931. https://doi.org/10.1016/j.jecp.2020.104931

Aïte, A., Cassotti, M., Linzarini, A., Osmont, A., Houdé, O., & Borst, G. (2016). Adolescents' inhibitory control: Keep it cool or lose control. *Developmental Science*, *21*(1), e12491. https://doi.org/10.1111/desc.12491

Bago, B., Bonnefon, J.-F., & De Neys, W. (2021). Intuition rather than deliberation determines selfish and prosocial choices. *Journal of Experimental Psychology: General*, *150*(6), 1081–1094. https://doi.org/10.1037/xge0000968

Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition*, *158*, 90–109. https://doi.org/10.1016/j.cognition.2016.10.014

Bago, B., & De Neys, W. (2019a). The Smart System 1: Evidence for the intuitive nature of correct responding on the bat-and-ball problem. *Thinking & Reasoning*, *25*(3), 257–299. https://doi.org/10.1080/13546783.2018.1507949

Bago, B., & De Neys, W. (2019b). The intuitive greater good: Testing the corrective dual process model of moral cognition. *Journal of Experimental Psychology: General*, *148*(10), 1782–1801. https://doi.org/10.1037/xge0000533

Bago, B., & De Neys, W. (2020). Advancing the specification of dual process models of higher cognition: A critical test of the hybrid model view. *Thinking & Reasoning*, *26*(1), 1–30. https://doi.org/10.1080/13546783.2018.1552194

Beaulac, G., & Kenyon, T. (2018). The Scope of Debiasing in the Classroom. *Topoi*, *37*(1), 93–102. https://doi.org/10.1007/s11245-016-9398-8

Bialek, M., & De Neys, W. (2016). Conflict detection during moral decision-making: Evidence for deontic reasoners' utilitarian sensitivity. *Journal of Cognitive Psychology*, *28*(5), 631–639. https://doi.org/10.1080/20445911.2016.1156118

Bialek, M., & Neys, W. D. (2017). Dual processes and moral conflict: Evidence for deontological reasoners' intuitive utilitarian sensitivity. *Judgment and Decision Making*, *12*(2), 148–167. https://doi.org/10.1017/S1930297500005696

Bialek, M., & Pennycook, G. (2018). The cognitive reflection test is robust to multiple exposures. *Behavior Research Methods*, *50*(5), 1953–1959. https://doi.org/10.3758/s13428-017-0963-x

Bonnefon, J.-F., & Rahwan, I. (2020). Machine Thinking, Fast and Slow. *Trends in Cognitive Sciences*, *24*(12), 1019-1027. https://doi.org/10.1016/j.tics.2020.09.007

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624–652. https://doi.org/10.1037/0033-295X.108.3.624

Burič, R., & Konrádová, ľubica. (2021). Mindware Instantiation as a Predictor of Logical Intuitions in Cognitive Reflection Test. *Studia Psychologica*, *63*(2), 114–128. https://doi.org/10.31577/sp.2021.02.822

Burič, R., & Šrol, J. (2020). Individual differences in logical intuitions on reasoning problems presented under two-response paradigm. *Journal of Cognitive Psychology*, *32*(4), 460–477. https://doi.org/10.1080/20445911.2020.1766472

De Neys, W. (2012). Bias and Conflict: A Case for Logical Intuitions. *Perspectives on Psychological Science*, *7*(1), 28–38. https://doi.org/10.1177/1745691611429354

De Neys, W. (Ed.). (2017). *Dual process theory 2.0*. Routledge.

De Neys, W. (2022). Advancing theorizing about fast-and-slow thinking. *Behavioral and Brain Sciences*, 1–68. https://doi.org/10.1017/S0140525X2200142X

De Neys, W., & Pennycook, G. (2019). Logic, Fast and Slow: Advances in Dual-Process Theorizing. *Current Directions in Psychological Science*, *28*(5), 503–509. https://doi.org/10.1177/0963721419855658

Diamond, A. (2013). Executive functions. *Annual review of psychology, 64*, 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Epstein, S. (1994). Integration of the Cognitive and the Psychodynamic Unconscious. *American Psychologist*, 16.

Erickson, T. D., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior*, *20*(5), 540–551. https://doi.org/10.1016/S0022-5371(81)90165-1

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143–149. https://doi.org/10.3758/BF03203267

Evans, J. St. B. T. (2002). Logic and human reasoning: An assessment of the deduction paradigm. *Psychological Bulletin*, *128*(6), 978–996. https://doi.org/10.1037/0033-2909.128.6.978

Evans, J. St. B. T. (2008). Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition. *Annual Review of Psychology*, *59*(1), 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Evans, J. St. B. T. (2011). Dual-process theories of reasoning: Contemporary issues and developmental applications. *Developmental Review*, *31*(2), 86–102. https://doi.org/10.1016/j.dr.2011.07.007

Evans, J. St. B. T. (2019). Reflections on reflection: The nature and function of type 2 processes in dual-process theories of reasoning. *Thinking & Reasoning*, *25*(4), 383–415. https://doi.org/10.1080/13546783.2019.1623071

Evans, J. St. B. T., & Over, D. E. (1996). Rationality in the selection task: Epistemic utility versus uncertainty reduction. *Psychological Review*, *103*(2), 356–363. https://doi.org/10.1037/0033-295X.103.2.356

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Frankish, K., & Evans, J. St. B. T. (2009). The duality of mind: An historical perspective. In J. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 1–30). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199230167.003.0001

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Frey, D., Johnson, E. D., & De Neys, W. (2018). Individual differences in conflict detection during reasoning. *Quarterly Journal of Experimental Psychology*, *71*(5), 1188–1208. https://doi.org/10.1080/17470218.2017.1313283

Gangemi, A., Bourgeois-Gironde, S., & Mancini, F. (2015). Feelings of error in reasoning—In search of a phenomenon. *Thinking & Reasoning*, *21*(4), 383–396. https://doi.org/10.1080/13546783.2014.980755

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work?. *Trends in cognitive sciences*, *6*(12), 517-523. https://doi.org/10.1016/S1364-6613(02)02011-9

Handley, S. J., Capon, A., Beveridge, M., Dennis, I., & Evans, J. S. B. (2004). : Working memory, inhibitory control and the development of children's reasoning. *Thinking & Reasoning*, *10*(2), 175–195. https://doi.org/10.1080/13546780442000051

Holt, C. A., & Laury, S. K. (2002). Risk Aversion and Incentive Effects. *The American Economic Review*, *92*(5), 1644–1655. https://doi.org/10.1257/000282802762024700

Johnson, E. D., Tubau, E., & De Neys, W. (2016). The Doubting System 1: Evidence for automatic substitution sensitivity. *Acta Psychologica*, *164*, 56–64. https://doi.org/10.1016/j.actpsy.2015.12.008

Kahneman, D. (2000). A psychological point of view: Violations of rational rules as a diagnostic of mental processes. *Behavioral and Brain Sciences*, *23*(5), 681–683. https://doi.org/10.1017/S0140525X00403432

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.

Kahneman, D., & Frederick, S. (2002). Representativeness Revisited: Attribute Substitution in Intuitive Judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and Biases* (1st ed., pp. 49–81). Cambridge University Press. https://doi.org/10.1017/CBO9780511808098.004

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263–291. https://doi.org/10.2307/1914185

Kessler, J. B., Kivimaki, H., & Niederle, M. (2017). Thinking fast and slow: generosity over time. *Unpublished manuscript. https://users.nber.org/~kesslerj/papers/KesslerKivimakiNiederle_GenerosityOverTime.pdf*

Keysar, B., Hayakawa, S. L., & An, S. G. (2012). The Foreign-Language Effect: Thinking in a Foreign Tongue Reduces Decision Biases. *Psychological Science*, *23*(6), 661–668. https://doi.org/10.1177/0956797611432178

Koriat, A. (2017). Can people identify "deceptive" or "misleading" items that tend to produce mostly wrong answers? *Journal of Behavioral Decision Making*, *30*(5), 1066–1077. https://doi.org/10.1002/bdm.2024

March, D., Olson, M., & Gaertner, L. (2023). Automatic threat processing shows evidence of exclusivity. Behavioral and Brain Sciences, 46, E131. https://doi.org/10.1017/S0140525X22002928

Mata, A. (2020). Conflict detection and social perception: Bringing meta-reasoning and social cognition together. *Thinking & Reasoning*, *26*(1), 140–149. https://doi.org/10.1080/13546783.2019.1611664

Mata, A., & Ferreira, M. B. (2018). Response: Commentary: Seeing the conflict: an attentional account of reasoning errors. *Frontiers in Psychology*, *9*. https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00024

Mata, A., Ferreira, M. B., & Sherman, S. J. (2013). The metacognitive advantage of deliberative thinkers: A dual-process perspective on overconfidence. *Journal of*

Personality and Social Psychology, *105*(3), 353–373. https://doi.org/10.1037/a0033640

Mata, A., Ferreira, M. B., Voss, A., & Kollei, T. (2017). Seeing the conflict: An attentional account of reasoning errors. *Psychonomic Bulletin & Review*, *24*(6), 1980–1986. https://doi.org/10.3758/s13423-017-1234-7

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. https://doi.org/10.1016/j.cognition.2014.07.011

Mechera-Ostrovsky, T., Heinke, S., Andraszewicz, S., & Rieskamp, J. (2022). Cognitive abilities affect decision errors but not risk preferences: A meta-analysis. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-021-02053-1

Melnikoff, D. E., & Bargh, J. A. (2018). The mythical number two. *Trends in cognitive sciences*, *22*(4), 280-293. https://doi.org/10.1016/j.tics.2018.02.001

Newman, I. R., Gibb, M., & Thompson, V. A. (2017). Rule-based reasoning is fast and belief-based reasoning can be slow: Challenging current explanations of belief-bias and base-rate neglect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(7), 1154–1170. https://doi.org/10.1037/xlm0000372

Neys, W. D., Novitskiy, N., Geeraerts, L., Ramautar, J., & Wagemans, J. (2011). Cognitive Control and Individual Differences in Economic Ultimatum Decision-Making. *PLOS ONE*, *6*(11), e27107. https://doi.org/10.1371/journal.pone.0027107

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2014). Cognitive style and religiosity: The role of conflict detection. *Memory & Cognition*, *42*(1), 1–10. https://doi.org/10.3758/s13421-013-0340-7

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. https://doi.org/10.1016/j.cogpsych.2015.05.001

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39-50. https://doi.org/10.1016/j.cognition.2018.06.011

Purcell, Z. A., Wastell, C. A., & Sweller, N. (2023). Eye movements reveal that low confidence precedes deliberation. *Quarterly Journal of Experimental Psychology*, *76*(7), 1539–1546. https://doi.org/10.1177/17470218221126505

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, *489*(7416), Article 7416. https://doi.org/10.1038/nature11467

Raoelison, M., Boissin, E., Borst, G., & De Neys, W. (2021). From slow to fast logic: The development of logical intuitions. *Thinking & Reasoning*, *27*(4), 599–622. https://doi.org/10.1080/13546783.2021.1885488

Raoelison, M., & Neys, W. D. (n.d.). Do we de-bias ourselves?: The impact of repeated presentation on the bat-and-ball problem. *Judgment and Decision Making*, 9.

Raoelison, M., Thompson, V. A., & De Neys, W. (2020). The smart intuitor: Cognitive capacity predicts intuitive rather than deliberate thinking. *Cognition*, *204*, 104381. https://doi.org/10.1016/j.cognition.2020.104381

Reber, A. S., & Allen, R. (2022). *The Cognitive Unconscious: The First Half Century*. Oxford University Press.

Sirota, M., Dewberry, C., Juanchich, M., Valuš, L., & Marshall, A. C. (2021). Measuring cognitive reflection without maths: Development and validation of the verbal cognitive reflection test. *Journal of Behavioral Decision Making*, *34*(3), 322–343. https://doi.org/10.1002/bdm.2213

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3–22. https://doi.org/10.10 37/0033-2909.119.1.3

Slovic, P., Peters, E., Finucane, M. L., & MacGregor, D. G. (2005). Affect, risk, and decision making. *Health psychology*, *24*(4S), S35. https://doi.org/10.1037/0278-6133.24.4.S35

Speckmann, F., & Unkelbach, C. (2021). Moses, money, and multiple-choice: The Moses illusion in a multiple-choice format with high incentives. *Memory & Cognition*, *49*(4), 843–862. https://doi.org/10.3758/s13421-020-01128-z

Šrol, J., & De Neys, W. (2021). Predicting individual differences in conflict detection and bias susceptibility during reasoning. *Thinking & Reasoning*, *27*(1), 38-68. https://doi.org/10.1080/13546783.2019.1708793

Stango, V., & Zinman, J. (2020). *Behavioral Biases are Temporally Stable* (w27860; p. w27860). National Bureau of Economic Research. https://doi.org/10.3386/w27860

Stanovich, K. E. (2018). Miserliness in human cognition: The interaction of detection, override and mindware. *Thinking & Reasoning*, *24*(4), 423–444. https://doi.org/10.1080/13546783.2018.1459314

Stanovich, K. E., & West, R. F. (2000). 24. Individual Differences in Reasoning: Implications for the Rationality Debate?. *Behavioural and Brain Science*, *23*(5), 665-726.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662. https://doi.org/10.1037/h0054651

Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning*, *20*(2), 215–244. https://doi.org/10.1080/13546783.2013.869763

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*(3), 107–140. https://doi.org/10.1016/j.cogpsych.2011.06.001

Travers, E., Rolison, J. J., & Feeney, A. (2016). The time course of conflict on the Cognitive Reflection Test. *Cognition*, *150*, 109–118. https://doi.org/10.1016/j.cognition.2016.01.015

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315. https://doi.org/10.1037/0033-295X.90.4.293

Vartanian, O., Beatty, E. L., Smith, I., Blackler, K., Lam, Q., Forbes, S., & De Neys, W. (2018). The Reflective Mind: Examining Individual Differences in Susceptibility to Base Rate Neglect with fMRI. *Journal of Cognitive Neuroscience*, *30*(7), 1011–1022. https://doi.org/10.1162/jocn_a_01264

Vega, S., Mata, A., Ferreira, M. B., & Vaz, A. R. (2021). Metacognition in moral decisions: Judgment extremity and feeling of rightness in moral intuitions. *Thinking & Reasoning*, *27*(1), 124–141. https://doi.org/10.1080/13546783.2020.1741448

Wason, P. C., & Evans, J. ST. B. T. (1975). Dual processes in reasoning? *Cognition*, *3*(2), 141–154. https://doi.org/10.1016/0010-0277(74)90017-1