LUND UNIVERSITY

The neurocognitive basis of confabulatory introspection

Choice blindness and the brain

Vogel, Gabriel

2024

Document Version:
Publisher's PDF, also known as Version of record

Link to publication

Citation for published version (APA):
Vogel, G. (2024). *The neurocognitive basis of confabulatory introspection: Choice blindness and the brain* (Lund University Cognitive Studies ed.). [Doctoral Thesis (compilation), Cognitive Science]. Lund University (Media-Tryck).

Total number of authors:
1

Creative Commons License:
CC BY-ND

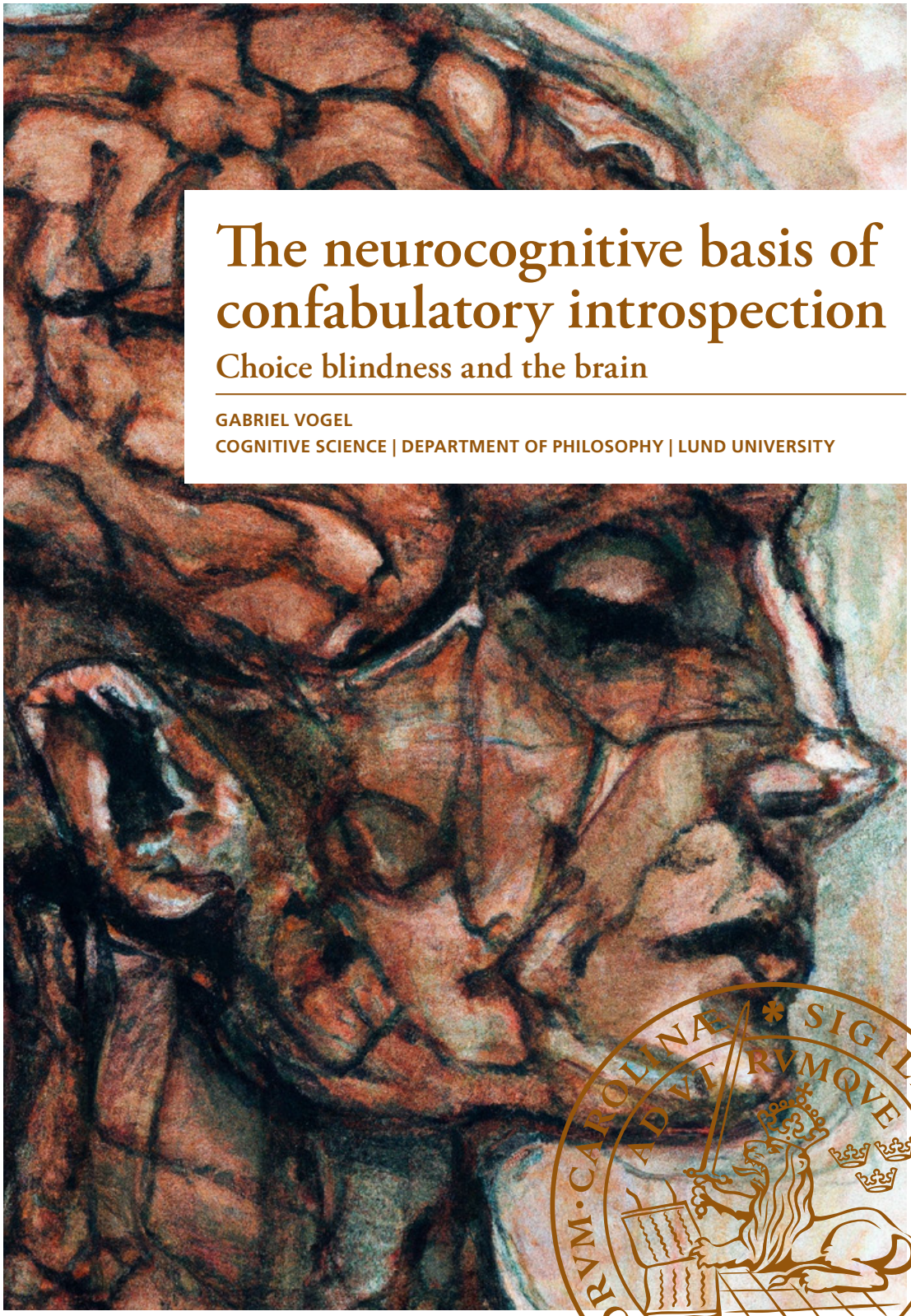
General rights
Unless other specific re-use rights are stated the following general rights apply:
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.
• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: https://creativecommons.org/licenses/

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00



The neurocognitive basis of confabulatory introspection

Choice blindness and the brain

GABRIEL VOGEL
COGNITIVE SCIENCE | DEPARTMENT OF PHILOSOPHY | LUND UNIVERSITY

Printed by Media-Tryck, Lund 2024 NORDIC SWAN ECOLABEL 3041 0903

The neurocognitive basis of confabulatory introspection

This thesis investigates the neurocognitive mechanisms of introspection using choice blindness and neuroimaging (fMRI). Choice blindness is a phenomenon in which people fail to notice that the outcome of their decision is not what they had originally chosen, and end up confabulating explanations for why they chose it. This thesis provides the first description of the brain networks involved in the failure to detect manipulations and in the production of confabulated explanations in choice blindness. It also investigates more deeply the condition under which illusions of choices arise in choice blindness, showing that this phenomenon persists even when people are instructed to detect manipulation, and how people can reject their genuine choice when reached through a wrong action. These empirical results are integrated in a broader review of the research in cognitive science, which suggests that introspection relies on the same interpretative neurocognitive mechanisms we use to understand other people's behaviour.

LUND UNIVERSITY

Joint Faculties of Humanities and Theology
Department of Philosophy
Cognitive Science

Lund University Cognitive Studies 187
ISBN 978-91-89874-26-8
ISSN 1101-8453

The neurocognitive basis of confabulatory introspection

The neurocognitive basis of confabulatory introspection

Choice blindness and the brain

by Gabriel Vogel

LUND
UNIVERSITY

DOCTORAL DISSERTATION

Thesis advisors: Petter Johansson, Lars Hall, Philip Pärnamets
Faculty opponent: Mark Schram Christensen

Doctoral dissertation for the degree of Doctor of Philosophy (PhD) at the Faculty of Humanities and Theology at Lund University to be publicly defended on the 12th of April 2024 in LUX room C121.

Organization: Lund University Cognitive Science

Department of Philosophy

LUND UNIVERSITY

Document name: DOCTORAL DISSERTATION **Date of issue: April 12, 2024**

Author(s): Gabriel Vogel

Title and subtitle: The neurocognitive basis of confabulatory introspection: Choice blindness and the brain

Abstract: The goal of this thesis is to advance our understanding of introspection by studying when it fails, without us being aware of it. To do so, I have used the choice blindness paradigm. Choice blindness is a surprising phenomenon in which people fail to detect mismatches between their intention and outcome in a decision task, and then spontaneously confabulate reasons why they preferred an alternative they did not choose. Very little is known about the mechanisms of choice blindness, both when people detect or not, and how this leads to confabulation. My contribution consists in making this phenomenon less puzzling, by dissecting its neurocognitive basis. This involves building a first framework of false feedback detection in choice blindness, and in so doing investigating the monitoring mechanisms and reasoning processes that allows us to keep track of our intentions and their consequences in the world. In addition, by studying how our brain uses confabulation and post hoc rationalization to integrate false information about one's' choices, I want to highlight the deeply interpretative nature of our self-knowledge, a facet that has often escaped our intuitive understanding of ourselves. In the introduction of this thesis, I also review the state of the art on interpretative models of introspection, highlight gaps in the literature, and formulate new research tracks in light of my findings.

In paper 1, I show that CB can arise without deception, as failures to detect false feedback persist even when participants are instructed to detect them. The study also shows the limits of our monitoring mechanism as well as how prior beliefs modulate false feedback acceptance. Building on these results, I outline a framework to understand choice blindness as the result of an interplay between automatic monitoring and reasoning systems. In paper 2, I show that the neural correlates of false feedback detection are consistent with the monitoring and inference framework described in paper 1. In the study, I find that detection is associated with reward monitoring (midbrain, basal ganglia, insula, ACC), sensory predictions (superior temporal sulcus, angular gyrus), as well as dorsal frontoparietal networks associated with executive control and reasoning. Paper 3 goes one step further, studying how outcome and motor levels of monitoring interact with each other. In the study, I show that motor monitoring can override outcome monitoring, leading people to reject outcomes they want when these are obtained through an action they perceive to be wrong. In paper 4, I investigate what happens at a neural level once monitoring failes, and people start constructing reasons for choices they never made. Here the finding is that confabulation involves the same theory of mind network that we use to make sense of others (mPFC, TPJ, STS) as well as areas related to reality monitoring (rPFC BA10) and executive function (dlPFC). This strengthens the interpretative perspective on introspection, suggesting that we use the same cognitive mechanisms to understand ourselves as those we use to understand others.

Key words: confabulation, introspection, fMRI, choice blidness, sense of agency, decision, self-knowledge, metacognition, cognitive neuroscience

Language: English **ISSN and key title:** 1101-8453 Lund University Cognitive Studies 187

ISBN:978-91-89874-26-8 (print), 978-91-89874-27-5 (digital).

Number of pages: 180

I, the undersigned, being the copyright owner of the abstract of the above-mentioned dissertation, hereby grant to all reference sources permission to publish and disseminate the abstract of the above-mentioned dissertation.

Signature Date 2024-02-13

The neurocognitive basis of confabulatory introspection

Choice blindness and the brain

by Gabriel Vogel

LUND
UNIVERSITY

Coverphoto by DALL·E 2

Copyright pp. 1-82 Gabriel Vogel 2024
Paper 1 © by the Authors (Manuscript unpublished)
Paper 2 © by the Authors (Manuscript unpublished)
Paper 3 © Elsevier under Creative Commons CC-BY
Paper 4 © by the Authors (Manuscript unpublished)

Faculty of Humanities and Theology
Department of Philosophy
Cognitive Science

ISBN 978-91-89874-26-8 (print)
978-91-89874-27-5 (digital).
ISSN 1101-8453 Lund University Cognitive Studies 187

Printed in Sweden by Media-Tryck, Lund University
Lund 2024

Media-Tryck is a Nordic Swan Ecolabel certified provider of printed material. Read more about our environmental work at www.mediatryck.lu.se

MADE IN SWEDEN

Acknowledgements

I would first like to thank my supervisors without whom I could not have reached the end of this adventure. Among many things, I am particularly grateful for Petter Johansson's unwavering support and kindness, which were invaluable during this long and challenging process; Lars Hall's suggestions of new tracks when I was uncertain where I was headed; and Philip Pärnamets' help in pushing my statistical knowledge to the next level.

My research journey involved an exciting yet testing transition to neuroimaging. I want to thank Johan Mårtensson and Peter Mannfolk for believing in my - slightly crazy - project to learn fMRI from scratch in the middle of my PhD and supporting me throughout. I would also like to thank the whole MRI team at Lund University Bioimaging Centre, especially all the radiographers who ran the scanner, helped with data collection and brought a great energy all along! Thanks, Theodor Rumetshofer, for the help with the analyses.

I am very grateful to all the people at the department for the great atmosphere their friendliness and openness created! The administrative team deserves a special thank for being so helpful and making everything so smooth. I am especially grateful for sharing so many stimulating discussions, good times - and great music - with Trond Arild Tjøstheim and Andrey Anikin. Thanks to my office mates, Anton Wrisberg, Alexander Tagesson and Matthew Tompkins, for bringing fun and laughter in the process!

I also want to thank all the Master students and research assistants who participated in my project. Their enthusiasm and fresh eyes brought me a lot. Special thanks to Daniel Zander, for being the best conference buddy!

I would also like to thank the Swedish Research Council, for funding my PhD (Grant No. 2014-1371).

Finally, I am deeply grateful for my family and friends back in France, for having been there all along, despite the distance.

Sammanfattning på svenska

Målet med den här avhandlingen är att fördjupa vår förståelse av introspektion genom att studera när vi inte är medvetna om att den misslyckas. För att göra detta har jag använt mig av fenomenet beslutsblindhet. Beslutsblindhet är den överraskande upptäckten att försöksdeltagare ofta inte upptäcker förändringar som sker mellan att de väljer någonting och de får se konsekvenserna av sitt val, och där de sedan spontant konfabulerar skäl till varför de föredrog alternativet de egentligen inte valde. Mycket lite är känt om mekanismerna bakom beslutsblindhet, både vad som avgör huruvida en förändring upptäcks eller inte, och hur det kan leda till konfabulatoriska förklaringar. Mitt bidrag består i ett försök att göra detta fenomen mindre svårbegripligt, genom att dissekera dess neurokognitiva grund. Detta innefattar att konstruera ett första ramverk för hur falsk återkoppling upptäckts i beslutsblindhetssituationer och därigenom undersöka de monitoreringsmekanismer och resonemangsprocesser som tillåter oss hålla reda på våra intentioner och deras konsekvenser i världen. Dessutom, genom att studera hur vår hjärna använder konfabulation för att integrera falsk återkoppling om våra egna beslut, vill jag belysa hur vår självkunskap i grunden är en form av tolkning eller slutledning, snarare än bara en intuitiv förståelse, så som det traditionellt har ansetts. I inledningen till denna avhandling granskar jag också den senaste forskningen om olika modeller av introspektion, lyfter fram luckor i litteraturen, och formulerar nya forskningsspår i ljuset av mina fynd.

I artikel 1 visar jag att beslutsblindhet kan uppstå utan dolda manipulationer, eftersom misslyckanden med att upptäcka falsk återkoppling kvarstår även när deltagarna instrueras att upptäcka dem. Studien visar också gränserna för våra monitoreringsmekanismer, samt hur tidigare föreställningar som försöksdeltagarna har påverkar acceptansen av falsk återkoppling. Baserat på dessa resultat skisserar jag ett ramverk för att förstå beslutsblindhet som resultatet av ett samspel mellan automatisk monitorering och inferenssystem. I artikel 2 visar jag att de neurala korrelaten för upptäckt av falsk återkoppling är i överensstämmelse med monitorerings- och inferensramverket som beskrivs i artikel 1. I studien finner jag att upptäckt är associerat med aktivitet i belöningssystem (midbrain, basal ganglia, insula, ACC), prediktion av sinnesintryck (superior temporal sulcus, angular gyrus), samt dorsala frontoparietala nätverk som är associerade med exekutiv kontroll i hjärnan. Artikel 3 går sedan ett steg längre och studerar hur övergripande monitorering av konsekvenserna av beslut, och övervakning på lägre

motoriska nivåer av vilken konkret handling som gjorts, interagerar med varandra. I studien visar jag att motorisk monitorering kan vägas tyngre än konsekvensmonitorering, vilket leder till att försöksdeltagare avvisar ett alternativ som de faktiskt vill ha när de har fått det här alternativet genom en handling som känns felaktig. I artikel 4 undersöker jag vad som händer på en neural nivå när monitoreringen misslyckas och människor börjar konstruera skäl för val de aldrig gjorde. Här är upptäckten att konfabulation involverar samma "theory of mind" nätverk, som vi använder för att förstå avsikter och åsikter hos andra människor (mPFC, TPJ, STS), samt områden relaterade till monitorering av huruvida en handling verkligen har utförts eller bara föreställts (rPFC BA10) och exekutiv funktioner (dlPFC). Detta stärker perspektivet att introspektion är en form av tolkning eller slutledning, och antyder att vi använder samma kognitiva mekanismer för att förstå oss själva som de vi använder för att förstå andra.

Table of Contents

List of original papers

Paper I

Vogel, G., Pärnamets, P., Hall, L., & Johansson, P. (2023). Choice blindness without deception: failures to notice false-feedback persist in explicit detection tasks. https://doi.org/10.31234/osf.io/amqwy

Paper II

Vogel, G., Mårtensson, J., Mannfolk, P., Hall, L., van Westen, D., & Johansson, P. (2024). The neural correlates of outcome monitoring and false feedback detection in choice blindness: a fMRI study. https://doi.org/10.31234/osf.io/smecn

Paper III

Vogel, G., Hall, L., Moore, J., & Johansson, P. (2024). The right face at the wrong place: How motor intentions can override outcome monitoring. *Iscience*, 27(1).

Paper IV

Vogel, G., Mårtensson, J., Mannfolk, P., Hall, L., van Westen, D., & Johansson, P. (2024). Catching the brain in the act of confabulation: a fMRI study. https://doi.org/10.31234/osf.io/e9vg7

Abbreviations

ACC	Anterior cingulate cortex
AG	Angular gyrus
BA	Brodmann area
BG	Basal ganglia
dlPFC	dorsolateral prefrontal cortex
mPFC	Medial prefrontal cortex
rPFC	Rostral prefrontal cortex
TPJ	Temporoparietal junction
STS	Superior temporal sulcus

Conceptual note

Throughout this thesis, I refer to introspection as the ability to know our mental states and processes. However, it is important to note that the way I use the term "introspection" does not presuppose the existence of a dedicated mechanism that directly measures or monitors our mental states and processes. Introspection is merely the ability - subserved by whichever mechanism - that allows to fulfil such function. One of the main points of my thesis is actually to suggest that most of introspection does not rely on a dedicated direct access mechanism, but on more general and indirect inferential processes. Other terms have been used to refer to very similar cognitive functions, such as "self-knowledge" or "metacognition". I chose the term introspection for several reasons. Self-knowledge tends to have a broader meaning, including our knowledge of our personality, or our ability to predict our future mental states (e.g. our satisfaction with a future experience, Wilson, 2009). The term metacognition refers to "thinking about thinking" (Norman et al., 2019), but has in recent research been more narrowly associated with the study of for example confidence judgments or how performance judgements track our actual performance (Fleming, 2024). Hence, I prefer using the term introspection to specifically refer to the ability that allows us to know our mental states and processes. Still, introspection, self-knowledge and metacognition will at times be used interchangeably, as their meaning often overlap.

Scope and summary

Explaining the reasons behind our choices and behaviour is an integral part of our daily interactions. Why did you choose this job? Why did you say this to me? Why do you like this city? Why did you vote for this party? We tend to come up so readily with an answer that we may think that the ability to introspect our motives is just an elementary part of our cognitive equipment. What caused your action is you. The reason lies within you, so you just need to look inside, find it, and report it back. This intuition of a *direct introspective access* to one's own mind has several roots. One relates to a longstanding western philosophical tradition that we could trace back to Descartes, according to which the only certainty we can have is about our subjective experience. Our everyday life can also make us feel that we are transparent to ourselves. When we find explanations for our own behaviour with such ease, why would we think that they may be wrong, or that we may be inventing stories rather than reaching inside to find the truth?

However, when experimental psychology started to scientifically probe the validity of our introspective reports, a very different picture started to emerge. We come up with believable reasons for our consumer choices, even when the two products we chose from were actually exactly the same (Nisbett & Wilson, 1977). When people get their hemispheres disconnected, they spontaneously come up with stories to explain decisions based on information that their "language" hemisphere does not have access to (Gazzaniga, 2014). At times, people spontaneously and confidently justify choices they never made, even in political or moral domains, and change their preferences accordingly, without being aware of it (Hall et al., 2012; Johansson et al., 2005; Strandberg et al., 2018). In contrast with the intuition of a privileged introspective mechanism, most cognitive models paint introspection as being self-interpretative in nature, i.e. we use the same interpretative resources to make sense of ourselves and others.

The ongoing endeavour of understanding how we construct our representation of the self is not devoid of methodological challenges. A delicate matter has long

been: how to find an objective ground to pit subjective reports against? How can we manage to legitimately say: "you may very well be sincere and fully believe in your explanation, but you are wrong about yourself"? In more philosophical terms, how can we bypass the first-person authority generally granted to introspective statements? Recently, the choice blindness paradigm provided a fundamental methodological advance to solve this issue. In a choice blindness experiment, we ask people to choose between two alternatives and ask them to explain why they did so. However, sometimes, instead of showing back the selected alternative, we show back the other as if it was the chosen one. When people fail to detect this manipulation (which they often do), we can be confident that people's justifications are confabulatory, because they did not choose the alternative they explain having chosen.

Using choice blindness as a model, the purpose of this thesis is to go one step further in our understanding of the interpretative nature of introspection. Indeed, despite having been extensively replicated in many different contexts, choice blindness and its neurocognitive underpinnings remain poorly understood. Getting a better understanding of this phenomenon however bears promising insights into how we construct our representation of the self.

First, choice blindness allows us to study how the brain monitors (and fails to do so) its own intentions and whether it manages to carry them out in the real world. This would be the foundation of our sense of agency, i.e. our sense of control of our bodily movements and their outcomes (Haggard & Chambon, 2012). Secondly, choice blindness as an experimental paradigm allows us to empirically study the production of confabulated reports in the normal population. Other paradigms based on misinformation can also produce confabulation by implanting false memories, but they tend to focus on factual questions, while choice blindness typically targets subjective preferences and attitudes (e.g. Loftus, 2005). Thirdly, choice blindness sheds light on how we unknowingly shape and change our preferences, based on the choices we *think* we have made.

In Paper 1, 2 and 3, I focused on the monitoring component of choice blindness. Paper 1 highlighted the limits of our monitoring mechanism, showing how people can fail to notice outcome manipulations even when specifically instructed to detect them. I proposed a first tentative framework of choice blindness, suggesting that self-attributions of unintended outcomes result from the interplay of automatic monitoring mechanisms and reasoning processes. In Paper 2, I made

the first attempt to uncover the neural correlates of false feedback detection in choice blindness using fMRI. In line with the monitoring and inference framework of Paper 1, I found that false feedback detection was associated with reward monitoring (midbrain, basal ganglia, insula, ACC), sensory predictions (superior temporal sulcus, lateral parietal cortex), memory monitoring (rPFC) as well as dorsal frontoparietal networks associated with executive control and reasoning. In Paper 3, I went one step further, investigating how different levels of monitoring, motor and outcome related, interact with each other during self-attributions. I showed that motor monitoring can override outcome monitoring, leading people to reject outcomes they wanted when they are obtained through erroneous actions. This allowed to investigate the inferential component of detection, as here people integrated different cues and actively rejected outcomes that they actually desired based on action cues.

But monitoring and the sense of agency is only one part of the choice blindness phenomenon. Another fascinating aspect is confabulation: how we invent plausible but inaccurate explanations for our behaviour, without being aware of their constructed nature. Confabulation in everyday life is an important prediction of self-interpretative models of introspection. However, the study of the brain basis of confabulation had so far been limited to clinical population. In paper 4, I provided the first fMRI study of everyday confabulation in a normal population using choice blindness. I showed that confabulation involves the same theory of mind network that we use to make sense of others (mPFC, TPJ, precuneus) as well as areas related to reality/memory monitoring (rPFC, BA10) and executive control (dlPFC). One interpretation of this intriguing finding is that choice blindness induced confabulation relies on the same brain basis of mentalizing, but requires more rationalization effort in order to process mismatches between true memories and wrong beliefs about one's own choices.

The third component of choice blindness, i.e. preference change, was not at the heart of this dissertation. However, the studies included here generally involved a measure of preference change, and other projects of mine not reported here included a more in-depth exploration of preference change. Hence, I will also briefly discuss this component in the final part of the introduction.

Models are an essential part of the scientific process. We need to use simplified approximations of the phenomenon we are interested in to break it down and get a finer grained understanding of its constituents. That is why my empirical work

went from introspection to choice blindness. However, this dissertation is also an opportunity to explore bigger theoretical questions about introspection, following the opposite path, from choice blindness to introspection in general. In this introduction, I will also give more space for fundamental questions that were only alluded to in our empirical work. Why is it the case that our brain is not endowed with a direct introspective mechanism? What could a general model of introspection look like, or do we need an array of specialized models for specific introspective processes? What are the next steps for the study of introspection? I hope that taking a step back from the nitty gritty of empirical research to take a more speculative outlook on introspection will be fruitful to motivate further research.

Hence, in the theoretical introduction to our papers, I will start by reviewing the literature on interpretative models of introspection, as well as alternative, competing "direct access" models of introspection.

Summary of papers

Paper I - Choice blindness without deception: failures to notice false-feedback persist in explicit detection tasks.

Research question

In all choice blindness studies so far, the participants receive false feedback on their choice without being informed that such manipulations will occur. In this paper, we investigated whether people would still fail to recognize manipulations even when being explicitly instructed to detect them. If people would still fail to detect manipulations in this context, that could shed light on the limited precision of the monitoring mechanism we use to assess the outcome of our decisions. I also present a monitoring and inference framework of detection in choice blindness and use it to interpret the results.

Procedure

Two hundred people were recruited on prolific to participate in an online choice blindness task. They had to choose which of two faces they found the most attractive, and were then presented again with the purported chosen option and had to report which facial feature mattered the most in their decision. However, in 8 trials, a choice blindness manipulation occurred, meaning that participants were presented with the option they did not choose as feedback on their choice. They could reject the manipulated outcome by pressing a button "I actually preferred the other face". Half of the participants were assigned to a standard implicit choice blindness condition, where they were not informed that feedback on their choice might be altered. The other half participated in an explicit detection condition, and were told that manipulations would occur, and were instructed to try to detect them. Finally, to measure possible preference change,

the participants were represented with the same pairs of faces and were asked to again choose the face preferred.

Results and conclusion

In the standard implicit choice blindness condition, the participants failed to detect 58% of the manipulated trials. However, even in the explicit detection condition, a large proportion of the manipulations were not detected (24%). Failures to detect a manipulation led to a preference change in both the explicit and the implicit choice blindness condition, meaning that in the second round of choices, the participants were more likely to choose a face they had previously been led to believe they liked. In addition, in the implicit choice blindness task, the later the first manipulation occurred, the less likely the participants were to detect it. Together, these results suggest that the mechanisms of outcome monitoring might have a limited precision. In addition, prior beliefs about feedback reliability may influence the likelihood to detect manipulation. This is in line with the monitoring and inference framework I outline in the paper. Not only monitoring matters for detection, but also other information such as feedback reliability can be integrated through an inference to self-attribute choices.

Paper II - The neural correlates of outcome monitoring and false feedback detection in choice blindness: a fMRI study.

Research question

Monitoring the outcome of our decisions is deemed to be integral to learning and our sense of agency. Here, we investigated which brain-based mechanisms are involved in outcome monitoring and the detection of manipulation in choice blindness. This was the first study to do so with fMRI. Based on the monitoring and inference framework of Paper 1, we expected to observe activations related to outcome monitoring, but also to reasoning processes.

Procedure

57 people participated in a choice blindness task while being scanned in a 7T fMRI scanner at Lund University Hospital. The choice blindness task was very similar to the one described in Paper 1 about facial attractiveness.

Results and conclusion

Detection was associated with activity in areas related to reward monitoring (midbrain, basal ganglia, insula, anterior cingulate cortex [ACC]), sensory predictions (superior temporal sulcus, lateral parietal cortex), memory monitoring (rPFC), as well as dorsal frontoparietal networks associated with executive control and reasoning. Failure to detect false feedback was not characterized by any activations, but a host of deactivations in more posterior/occipital regions (lingual, fusiform, parahippocampal gyrus, posterior cingulate cortex). These findings suggest that feedback attribution in choice blindness relies on automatic monitoring mechanisms generating error signals, which are then integrated and interpreted by reasoning and executive systems.

Paper III - The right face at the wrong place: How motor intentions can override outcome monitoring.

Research question

Monitoring is deemed to occur in a hierarchical manner. However, less is known about how different levels of monitoring interact. In this experiment, we investigated how motor and outcome monitoring are integrated in our judgments of agency. We tested whether people sometime rely more on their action than the outcome of a choice, and if they might reject the outcome they actually wanted as a consequence of errors at the motor level.

Procedure

80 participants took part in an adapted choice blindness task about facial attractiveness. The participants had to select the face they found the most attractive by dragging a mouse cursor to it. We induced motor errors by forcefully

deviating the cursor during selection, or we created outcome errors by switching the position of the chosen face, or we did both at the same time. In this last and theoretically most interesting condition, a motor error was experienced, despite the outcome being correct.

Results and conclusion

In the last condition, the participants rejected the outcome they wanted when their action to reach it was wrong in a majority of trials (59%). This rejection was made with very high confidence and had downstream effect on the participants' preferences, such that after having rejected the initially preferred alternative they were much less likely to choose this alternative a second time. This suggests that monitoring may be less straightforward than a process of matching intention and outcome, contrary to what is typically assumed.

Paper IV - Catching the brain in the act of confabulation: a fMRI study.

Research question

Choice blindness have shown how we sometime invent plausible but inaccurate explanations for our own behaviour without being aware of their constructed nature. In line with this result, interpretative models of introspection have postulated that confabulation and introspection rely on very similar neurocognitive mechanisms generally used for social cognition. Especially, the left-brain interpreter model supposes that this story-making process is mostly performed by the left hemisphere. However, these assumptions have not been tested with neuroimaging in the normal population. To investigate the relationship between introspection and confabulation at the neural level, we conducted a fMRI study using the choice blindness paradigm.

Procedure

The experiment was the same as the one used in paper II. This paper is based on the specific analysis of when people explained their decision silently in the scanner

when seeing the outcome of their choice. In the analysis, we compared brain activity related to confabulation (non-detected manipulations) and non-confabulation (non-manipulated trial).

Results and conclusion

The study showed that confabulation in the normal population is associated with right-sided activations. Confabulation was associated with the mentalizing network typically involved in social cognition (right temporoparietal junction, medial prefrontal cortex, precuneus). Confabulation also recruited areas related to reality and memory monitoring and executive functions (right rostral and dorsolateral prefrontal cortex). However, non-confabulation did not appear to recruit any other areas. This suggested that confabulation and introspection share the same interpretative basis related to social cognition, although confabulation may require a more effortful rationalization activity in the context of choice blindness. This would explain increased activity in the mentalizing network with no other specific activations in non-confabulation. rPFC and dlPFC activity may reflect executive and memory mechanism related to the processing of a mismatch between true memories of one's choice and false beliefs about one's choice induced by choice blindness. This result also challenges the left-brain interpreter theory of confabulation, as confabulation-related activations were exclusively right-sided.

The interpretative nature of introspection: a theoretical and empirical review

1. Summary

Contrary to common assumptions, we don't appear to have a direct access to our decision making and judgment processes. Reviewing the theoretical and empirical literature on introspection in cognitive science, I show that interpretative models are numerous and that they are supported by a wide range of evidence. In contrast, "direct access" models are underdeveloped and rest on a relatively limited empirical basis. After reviewing how modularity, motivated cognition, mentalistic categorization and strategic self-deception may explain why introspection is interpretative, I outline new research avenues to further our understanding of introspection. For example, I highlight the need for more precise and specialized computational models, and that the strategic self-deception hypothesis of introspection should be more thoroughly explored. Despite the multitude of evidence for interpretative models of introspection, I also emphasize the value of building more precise "direct access" models as a contrast, allowing us to draw more definitive conclusions on the nature of introspection.

2. Introduction

Our ability to introspect is at the heart of fundamental questions, spanning philosophical, epistemological, interpersonal, and institutional questions. At an intuitive level, one may sense how odd it would be for an agent to strive to reach goals, plan, act and interact with others without being aware of her own goals and reasons. But however odd this picture of the human condition may appear to be, cognitive science models suggest that it has an element of truth. We may of course not be fully blind to who we are. Even if our motives were entirely hidden from us, we could infer them from our actions, as an external observer would. This is actually the point made by what we call "interpretative models of introspection". According to these models, even if we know ourselves to some degree and have a private access to subjective feelings and thoughts, the cognitive machinery allowing us to know ourselves is the same one that we use to interpret others. And introspective blind spots may lie around the corner, without us being aware of how we fill them with plausible but sometime inaccurate narratives about ourselves.

There are many consequences of this positions. Scientifically, important controversies have surrounded the status of introspection, shaping the paradigm of current experimental psychology. Philosophically, the idea that we would have to surrender our first-person authority about our motives and mental states may seem puzzling. What would such a world be like when we are no longer authorities about what goes on in our own minds? Where to your sincere statement: "I want this job" other could justifiably say: "no you don't". But this view of the mind may also influence our understanding of how we live our everyday life. It has for example been argued that our lack of access to our motives may be at the root of multiple societal dysfunctions, in the educational, political or health domain (Simler & Hanson, 2017). It may also affect us at a very basic interpersonal level, as for example relationship success has been shown to be influenced by our capacity for self-knowledge (Tenney et al., 2013).

Answering the big philosophical questions is beyond the scope of the present work. But to be properly equipped to tackle them in the future, a fundamental first step is to build an understanding of how introspection is working in our mind/brain. What neurocognitive mechanisms allow us to report the (purported) reasons behind our choices, and how much we prefer or believe in something?

3. A short historical detour

3.1. The intuition of direct introspective access

Relying on arguments, intuitions, and daily life observations, many philosophers have noted that we appear to be in privileged position to know our own mind (Bar-On, 2004; Gertler, 2021, see Carruthers, 2011 for an overview). We so effortlessly reach a plausible explanation for our behaviour that we may very well be endowed with a *dedicated introspective neurocognitive mechanism,* allowing us direct access to our own mental states. This has sometimes been described as an "inner sense", allowing or brain to "sense", perceive, measure, monitor our own mental states, making them available to consciousness and verbal report (Carruthers, 2011).

Early attempts to frame this model in cognitive terms postulated that we have a monitoring system which can access our mental states, or more precisely our "propositional attitudes", such as our beliefs, preferences and desires (Nichols & Stich, 2003). We would be able to know that we prefer left wing parties to right wing parties because the information "I prefer left wing parties" is stored somewhere in a "belief module" in our brain. This information is accessed/retrieved by the monitoring mechanism and made available for verbal report (see section 4. for more details).

Even if this may sound plausible and intuitive a priori, the big question is: are empirical data consistent with such a model?

3.2. Early doubts about introspective access

Before going to the state of the art of models and data related to interpretative theories of introspection, it is valuable to take a short historic detour to the paper "Telling more than we can know" by Nisbett and Wilson, which has shaped the field of introspection and self-knowledge. In early psychology, Freud's notion of the unconscious already instilled scepticism as to whether we can access the content of our own mind. Similarly, the status of introspection was at the heart of foundational debates at the early age of experimental psychology, opposing introspectionist schools to behaviorists, whose radical empiricism excluded mental and subjective phenomena from scientific inquiry (Brock, 2013). After the advent of cognitive science as a new way to put to put the mind back in the study of

psychology, a landmark paper was published by Nisbett and Wilson in 1977. Using experimental data to question our ability to introspect our own mental states, they famously argued that that we have "little to no access to our *higher order* [emphasis added] cognitive processes" p.231.

Interestingly, this oft-cited paper was born in a scientific context in which the lack of access to *lower order* neurocognitive process such as perceptual processing, memory, motor control, was already taken as a well-documented -and almost self-evident- scientific truth. Self-evident because, as Nisbett and Wilson put it: "It would be absurd, for example, to ask a subject about the extent to which he relied on parallel line convergence when making a judgment of depth or whether he stored the meanings of animal names in a hierarchical tree fashion or in some other manner" (Nisbett & Wilson, 1977, p.232). If we had such an absolute introspective access to our mental processes, experimental psychology would be utterly superfluous. To the question: do perception rely on unconscious Bayesian inference, we would just look inside, and answer yes or no. As an attempt to empirically investigate this question, Miller (1962) showed that when people were asked factual question such as "what is your mother's name" and then asked how they came up with an answer, they generally avowed their ignorance, along the lines of "I don't know, it just came to me".

The real question that Nisbett and Wilson wanted to tackle was: do we have any introspective access to our *higher order cognitive processes* such as judgment, choice, inference, and problem solving? They recognized a surprising gap between our tendency to come up with explanation for our lower and higher cognitive process. In contrast with lower cognitive processes, for which we willingly acknowledge our ignorance, when it comes to explaining how we came up with a choice or judgment, we generally have a ready answer based on multiple reasons. Nisbett and Wilson's point was to show that these explanations often don't reflect a real knowledge but are fabricated on the spot.

Summarizing literature on subliminal perception, incubation and preference change, as well as reporting no less than 7 new experiments showing people's failures to assess the role of various factors in their decision-making, they concluded that we have "little to no access to our *higher order* cognitive processes" (Nisbett & Wilson, 1977, p.231). For example, in their famous "stocking" experiment, they asked passersby in a mall to choose between different pairs of stockings. After this, people spontaneously explained the reason for the decision.

However, the pairs of stocking were all exactly the same. Nisbett and Wilson hence suggested that rather than coming from a dedicated introspective mechanism, introspective reports mostly reflected the use of causal theories of how behaviour comes about. People would unknowingly try to infer the most likely reasons behind their choices. Or in their own words: "When reporting on the effects of stimuli, people may not interrogate a memory of the cognitive processes that operated on the stimuli; instead, they may base their reports on implicit, a priori theories about the causal connection between stimulus and response." (Nisbett & Wilson, 1977, p.233).

4. Introspection as self-interpretation

4.1. Review of interpretative models of introspection

From this point on, a large array of evidence has accumulated suggesting that our introspection relies on unconscious inferences, integrating our past memories, situational cues, current sensations as well our implicit causal theories about behaviour. Many cognitive models suppose that introspection is not different in kind to interpreting other's behaviour. According to this framework, we use the same interpretative mentalizing apparatus (sometimes called mindreading, theory of mind or social cognition) for self- and other- knowledge. To account for this, researchers often say that there is a "symmetry" or a "parallel" between self and other's knowledge (Carruthers, 2011; Gopnik, 1993). In table 1, I review 5 models originating from different research fields, based on different evidence, and focusing on different mental states, all converging on the same conclusion about introspection. Our self-attribution of mental states (e.g. emotion, beliefs, desires, preferences, attitudes, intentions…) is the result of unconscious inference based on various cues and memories as well as our causal theories about behaviour.

Table 1. Interpretative models of introspection

Source	Model	Area	Introspective target
(Bem, 1972)	Self-perception	Social psychology	Attitudes, emotions, internal states
(Gazzaniga, 1985)	Left-brain interpreter	Clinical cognitive neuroscience	NA
(Gopnik, 1993)	Theory theory of mentalizing	Developmental psychology	Psychological states, intentional states, belief, desires
(Wegner & Wheatley, 1999)	Apparent mental causation	Social psychology	Will, volition
(Carruthers, 2011)	Interpretative sensory access	Philosophy of mind	Mental states, attitudes

For example, the self-perception theory first claimed that the same processes are at play when we attribute mental states to ourselves and others. We rely on internal and external cues, and, when internal cues are weak or ambiguous, we are virtually in the same position as external observers when "introspecting" (Bem, 1972). Starting from clinical research on split brain patients whose brain hemispheres have been surgically disconnected and who unknowingly confabulated reasons for their actions when the crucial information was not available to their left "language" hemisphere, Gazzaniga (1985) formulated the "left-brain interpreter" theory, attributing to the left hemisphere the function of building plausible stories of how our behaviour originated (but see our results suggestive of a "right-brain" interpreter in Paper 4). Starting from developmental psychology, and observing the concomitant appearance of the ability to attribute specific mental states to oneself and others, Gopnik (1993) argued that introspection was always mediated by the interpretative and conceptual resources of our theory of mind. Pushing the philosophical implication of this view one step further, Wegner and Wheatley (1999) claimed that our experience of conscious will was illusory as consciousness would not be the cause of our actions (see figure 1). In other words, we think that our conscious thoughts cause our actions, but our thoughts are only epiphenomenal (they have no real causal power). The real physical cause of behaviour is unconscious neurocognitive processes. We illusorily grant a causal role to our conscious thoughts merely because of their close temporal contiguity and congruency with our actions, and because of the lack of other salient causal candidates. Building on the neuronal workspace theory of consciousness and the modularity of the mind, Carruthers (2011) suggested that the only representation available to introspection are sensory motor in nature. The jump from these cues to self-attributions of mental states necessitates an inferential step.

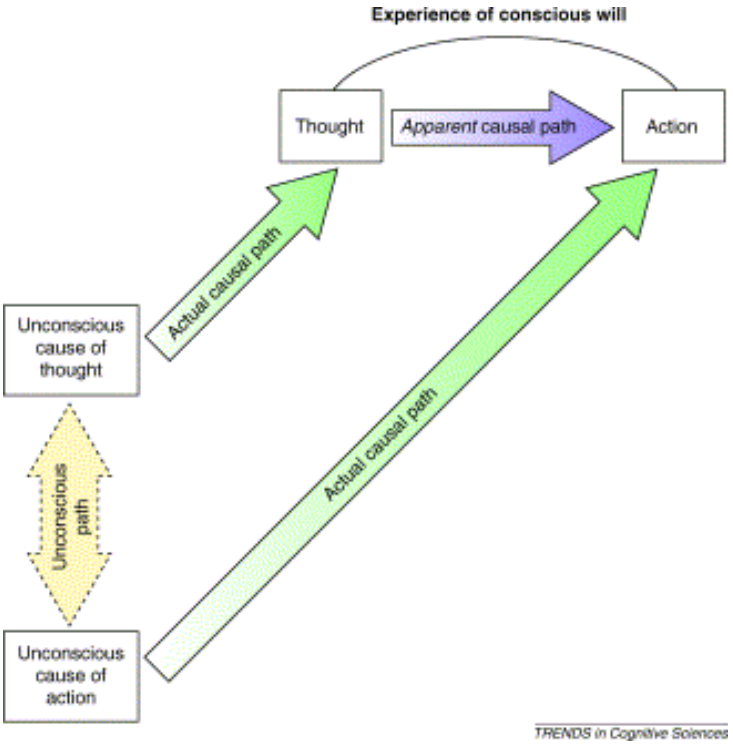


Figure 1. Illustrations of the apparent mental causation theory (Wegner, 2002). Reproduced with permission of Elsevier.

4.2. Empirical evidence for interpretative models

Abundant evidence has been supporting these interpretative models (see Carruthers, 2011 and table 2). First, strong evidence supports the hypothesis of a parallel between mentalizing and introspection. For example, mentalizing and self-processing rely on the same network in the brain (Denny et al., 2012; Kestemont et al., 2015; Legrand & Ruby, 2009). The ability to attribute various kinds of mental states to oneself and others appear at the same time during children's development (Gopnik, 1993, but see Nichols & Stich, 2003 for objection and Carruthers, 2011 for a response). Impairments of mentalizing are tied to impairments of introspection, and vice versa (Carruthers, 2011; Frith & Happé, 1999; Happé, 2003).

Another prediction of interpretative theories is that manipulating the cues available to people would influence their introspective judgements, even if they

did not influence their cognitive processes or decision. In accordance with this, it has been shown that sensory motor cues that are causally irrelevant in one's behaviour but congruent with a specific mental state influence people's introspective reports. For example, people reported stronger confidence in their judgment about arguments' persuasiveness when they were instructed to nod for irrelevant reasons (Briñol & Petty, 2003, 2022; Wells & Petty, 1980), and manipulating emotional prosodic characteristics of vocal feedbacks influences people's judgments about their emotional state (Goupil et al., 2021). Similar biases have been shown to influence our judgments of intentionality or agency. Wegner and Wheatley (1999) showed that people tend to self-attribute intentionality for uncontrolled cursor stops over items that were auditorily primed in close precedence to the stop. Additionally, when primed to entertain thoughts consistent with the observed outcome, people rate their causal influence on the outcomes as higher, despite having no control over it (Pronin et al., 2006). This influence also happens with non-specific sensory motor cues. For example, priming people to remember past pro- or anti- ecology behaviours biased their self-reported attitudes about environmentalism (Chaiken & Baldwin, 1981), or just letting them believe that they made a decision they never did changed their attitude about the topic (Albarracín & Wyer, 2000).

Table 2. Evidence for interpretative models of introspection

Type	Evidence	Selected references
Developmental	The ability to attribute mental states to oneself and others develop at the same age in children	(Gopnik, 1993)
Clinical	Association between impairment of introspection and mentalizing	(Frith & Happé, 1999; Happé, 2003)
Neuronal	Same brain basis for mentalizing and self-processing	(Denny et al., 2012; Kestemont et al., 2015; Legrand & Ruby, 2009)
Behavioural/ Neuronal	Confabulatory reports about the reasons for owns choices and judgments	Johansson et al., 2005, Vogel et al (paper 4)
Behavioural	Influence of sensory motor cues on introspective reports	Briñol & Petty, 2003; Wegner & Wheatley, 1999; Goupil et al., 2021
Behavioural	Failures to assess the role of information in one's decisions	(Nisbett & Wilson, 1977)
Behavioural	Gap between self-reported importance of attributes and their actual contribution in judgments	(Slovic & Lichtenstein, 1971)

Interestingly, another prediction of interpretative models of introspection is that, keeping the cues constant, people may infer different mental state depending on their causal beliefs or theories. This aspect was pointed out early on by Nisbett

and Wilson in 1977. However, to my knowledge, it has not been explicitly investigated. Results of paper 1 and 3 can be interpreted as showing the role of causal beliefs. For example, in the classic choice blindness condition in Paper 1, I showed that the more filler trials there were before the 1st manipulation, the less likely participants were to detect it. This can be interpreted as participant building a stronger prior that feedback provided on their choices is reliable, making them more likely to accept it as their own choice. For example, experimental failure may appear as a less likely cause of false feedback the more it is experienced as being reliable, making participants more likely to judge that the outcome is correct when it was not. Similarly, in paper 3, we showed that people tended to reject outcome they wanted when the action leading to it was wrong (i.e. we manipulated a mouse cursor to reach the undesired target, but then feedbacked the option that the subject wanted). This result can be interpreted as the product of a causal reasoning of the form: "if my action was wrong, it can't be the outcome I wanted". Indeed, in everyday life, the physical action (the cursor reaching the wrong target) should lead to the wrong outcome. Participants may struggle considering an alternative causal scenario where the wrong action leads to the right outcome. However, more research is needed to specifically and explicitly manipulate people's causal beliefs, for example by priming them with different theories about behaviour and seeing how their introspective judgments are affected (Nisbett & Wilson, 1977), or making them learn new causal relation (e.g. in virtual reality) and assess how it impacts their self-attributions.

Another more mixed line of evidence can be found in the self-insight literature in judgment and decision-making, where researchers have been investigating how well people assess the influence or importance of specific attributes in their judgments. One can evaluate people's "self-insight" by comparing their self-estimated importance of attributes and the objective attribute weights derived from a regression model. In line with interpretative models, an extensive review of early research on self-insight suggested that introspection of attribute weights was rather unreliable (Slovic & Lichtenstein, 1971). For instance, Slovic and Linchtenstein (1971) found that the correlation between implicit and explicit weights of 13 professional stockbroker was as low as 0.34 in a stock selection task. However, in their criticism of research on unconscious influence on decision making, Newell and Shanks (2014) highlighted that sometimes, people can also be very accurate in these tasks (Lagnado et al., 2006), and that several methodological issues may be at play. For example, the tendency of some

experiments to ask for attribute weight evaluation after a full block of judgments rather after each individual judgment. Newell and Shanks also formulate an interesting yet double-edged criticism. It could be the case that people don't arrive at a judgment through a weighted sum of attribute value, as typically assumed, but using different decision-making processes, such as heuristic or rule-based processes. However, the very fact that participant don't protest the cue weight evaluation as being incongruent with their true judgment process may already be a clue that they have little insight about it. Given these mixed results, there appears to be a need for more clear-cut experiments on self-insight, respecting the various criteria proposed by Newell and Shanks (2014), i.e. reliability, relevance, immediacy and sensitivity).

Finally, a crucial and fascinating line of evidence for interpretative models of introspection is the one at the heart of this dissertation, namely: choice blindness and confabulatory introspection. As we mentioned, when scientists started to pit introspective reports against objective grounds, they realized that we were prone to invent plausible but inaccurate explanations on the spot, without being aware of their constructed nature (Johansson et al., 2006; Nisbett & Wilson, 1977; Scaife, 2014). For now, we will just point at the fact that confabulation in daily life is an important prediction of interpretative models of introspection. We come up with wrong explanations about the cause of our behaviour, without knowing subjectively that they are false. Actually, we can even tend to be very confident in them (see Paper 3). This fits with the idea that we use the same interpretative mentalizing inference to explain our behaviour and the one of others. Paper 4 supported this view with fMRI analyses, showing the involvement of the mentalizing network in choice blindness induced confabulations.

4.3. Zooming in on choice blindness

As choice blindness plays a central role in my own thesis, I will also at some length review the prior findings using this paradigm. The choice blindness paradigm was developed as a new tool to investigate introspection and self-knowledge in decision making (Johansson et al., 2006). In a typical experiment, participants choose one of two pictures of faces they find the most attractive, and are asked for their reasons why this was the face preferred. However, some trials are followed by a surreptitious switch of the alternatives, where participants are shown the face they did not chose as if they had chosen it. A surprising finding is that participants

often fail to detect the manipulation and offer spontaneous, confident, introspectively derived reasons for why they chose the way they did. One of the values of this paradigm lies in clearly dissociating the process of making a choice and later explaining why this choice was made. After the discovery of choice blindness in Johansson et al (2005), an abundant literature has accumulated. As of now, no less than 60 studies have replicated the choice blindness phenomenon across a wide range of domains and stimuli (see table 1, 2 and annex 1 for a full list of current choice blindness studies).

Table 1. Domains of study of choice blindness. The frequency table reports the number of studies in each domain.

DOMAIN	Frequency
Facial attractiveness	15
Eyewitness/forensic psychology	7
Political judgment	5
Aesthetic judgment	3
Clinical	3
Reasoning	3
Consumer decision	2
Health	2
Memory	2
Moral judgment	2
Neuroimaging	2
Norm violation	2
Risk preference	2
Sympathy judgment	2
Behavioural norms	1
Experiential avoidance	1
Financial decision	1
Flatmate choice	1
Food preference	1
Own Personality	1
Preference for objects	1
Religious attitude	1
Selective attention	1
Tactile preference	1

After an initial focus on judgments of facial and aesthetic attractiveness, a first research effort has been devoted to investigating how choice blindness generalizes to other judgment and decision problems. Nowadays, choice blindness has been studies in a wide array of domains. For example, choice blindness studies have shown how people can justify political, moral, religious, and health-related

attitudes they did not originally hold (Hall et al., 2012, 2013; Law et al., 2017; Merckelbach et al., 2011a; Rieznik et al., 2017; Strandberg et al., 2018; Vranka & Bahnik, 2016). The responses to various questionnaires, such as personality, history of norm violation, experiential avoidance have also been successfully manipulated with choice blindness (Ambrus et al., submitted; Artenie et al., 2023; Sauerland, Schell, et al., 2013). In relation to behavioural economics, choice blindness has also been shown to occur in the domain of risk preference and financial and consumer decisions (Cheung et al., 2016; Hall et al., 2010; Kusev et al., 2022; McLaughlin & Somerville, 2013; Muda et al., 2020).

Table 2. Stimuli used in Choice Blindness experiments. The frequency table reports the number of studies using each type of stimulus.

Stimuli	Frequency
Pictures of faces	27
Political survey	5
Abstract Patterns	2
Lineups	2
Monetary gambles	2
Norm violating behaviour questionnaire	2
Reasoning problems	2
Scenario of accident	2
Can of soup	1
Descriptions of morally ambiguous behaviours	1
Experiential avoidance questionnaire	1
General knowledge	1
Health state scenarios	1
Investment portfolios	1
Jam and tea	1
Moral principles and issues	1
Objects	1
Personality traits	1
Pictures of chocolates	1
Pictures of natural sceneries	1
Pictures of objects	1
Prescriptive view of aging questionnaire	1
Psychiatric symptoms	1
Religious statements	1
Videos of events	1
Voice recordings	1

Choice blindness has shown to be robust across various modes of presentation, from real life slight of hands with picture cards or surveys (Johansson et al., 2005; Strandberg et al., 2018), to computerised task in the lab or online (Johansson & Hall, 2008; Vogel et al., 2023 [paper 1]), and even in virtual reality (Johansson et al., 2007). Interestingly, choice blindness arises in ecological situations such as tea and jam testing in a store (Hall et al., 2010), product evaluation (Cheung et al., 2016) or lineup identification (Sagana et al., 2013). Choice blindness was also tested in the clinical domain. It was shown to be associated with obsessive compulsive disorder (Aardema et al., 2014; Wong et al., 2020). A recent study including people with autism suggests that, although detection rate is not modulated by autism, this group experience a lower level of choice blindness induced preference change (Remington et al., 2023).

In addition, a substantial effort has been made to prove that choice blindness as such is not a mere artifact of demand effects, i.e. to prove that people genuinely fail to notice false feedback and then self-attribute choices they never made. Several studies have shown that choice blindness is not associated with social desirability (Aardema et al., 2014; Sauerland et al., 2014; Sauerland, Sagana, et al., 2013; Sauerland, Schell, et al., 2013), compliance (Sauerland et al., 2016; Sauerland, Sagana, et al., 2013) or suggestibility (Sauerland, Schell, et al., 2013). In addition, some studies have shown a form of "choice blindness blindness". That is, after having finished an experiment, participants are asked whether they think they would be able to detect false feedback if they would be exposed to it. A high rate of participants think they would detect them, although they failed to do so in the experiment they just completed (Johansson et al., 2005; Sauerland, Schell, et al., 2013). When participants are explicitly told that their responses will sometimes be manipulated and tasked to detect these mismatches, they still fail to detect a significant proportion of the false feedback (Vogel et al., 2023 [paper 1 in this thesis]). An eye tracking study showed that there is very little difference in pupil dilation between non-detected manipulations and non-manipulated trials, while a significant difference exists between detected manipulations and non-manipulated trials (Pärnamets et al., 2023). This clearly goes against the hypothesis that participants would detect the manipulation but refrain from reporting it. Interestingly, one study showed that choice blindness was resistant to some task incentives (i.e. getting the chocolate, jam or tea they selected in the experiment) (Hall et al., 2010; Somerville & McGowan, 2016).

In addition, several studies suggest that choice blindness cannot be explained by a mere memory deficiency account. Recognition memory performances after a choice blindness task have been shown to be very high (84% accuracy) (Pärnamets et al., 2015), and memory performances don't differ between manipulated and non-manipulated trials when people are informed that some trials were manipulated after a choice blindness task (Sagana et al., 2014a). In addition, several studies show that people remain very consistent with their original choice when not manipulated (Johansson et al., 2013; Vogel et al., 2023 [paper 1]). Finally, it has also been shown that undetected manipulations tend to change participants preferences and attitudes, suggesting that the wrong beliefs about one's choices are internalized (Hall et al., 2013; Johansson et al., 2013; Muda et al., 2020; Pärnamets et al., 2020; Strandberg et al., 2018; Taya et al., 2014; Vogel et al., 2023 [paper 1]).

Another line of inquiry of why choice blindness arises is to investigate various factors possibly associated with the failure to detect manipulations. For example, it has been shown that detection is influenced by the similarity between choice alternatives (Hall et al., 2010; McLaughlin & Somerville, 2013; Sauerland, Sagana, et al., 2013; Steenfeldt-Kristensen & Thornton, 2013; Vogel et al., 2023 [paper 1]), initial preference strength (Hall et al., 2010, 2012; Somerville & McGowan, 2016; Strandberg et al., 2018; Vogel et al., 2023 [paper 1]), confidence in one's decision (Strandberg et al., 2018), choice complexity (McLaughlin & Somerville, 2013) and familiarity with choice alternatives (Hall et al., 2012; McLaughlin & Somerville, 2013)

Several studies assessed the association between choice blindness and individual differences, most without finding any clear results. Working memory capacity did not appear correlated with choice blindness (Poorun et al., 2018). However, if the overall working memory capacity does not seem involved in choice blindness, ongoing working memory activity may play a role. Indeed, longer retention intervals between choice and false feedback tend to decrease detection rate (Johansson et al., 2005; Sauerland, Sagana, et al., 2013). Preference for consistency and need for cognition seem not to be associated with choice blindness (Strandberg et al., 2019). Within the big five personality traits, no correlations have generally been found with choice blindness (Law et al., 2017; Poorun et al., 2018; Sauerland, Schell, et al., 2013), except in one study where openness was related to lower detection rate (Sauerland, Schell, et al., 2013). Trait mindfulness may not influence false feedback

detection (Artenie et al., 2023), but actual meditative practice fosters higher detection rate (Lachaud et al., 2022). Depression symptoms may decrease detection rate (Aardema et al., 2014 but see Wong et al., 2020 for a contradictory result). Schizotypy was sometime found to be associated with choice blindness (Aardema et al., 2014) and sometime not (Aardema et al., 2014; Wong et al., 2020). However, it has been shown that critical thinking is associated with a higher detection rate (Strandberg et al., 2018, 2019).

Despite this abundant literature, an overarching framework to understand why people fail to detect manipulations in choice blindness is still lacking. However, monitoring the outcome of our decision is deemed important for learning and our sense of agency (Daw & Tobler, 2013; Moore, 2016; Ridderinkhof et al., 2004). Hence, it is important to understand why outcome monitoring typically fails in choice blindness. To address this, in paper 1, I propose a general framework of choice blindness, focused on the failure or success to detect false feedback. According to this monitoring and inference framework, there exist monitoring systems generating error signal when something in our environment or actions goes wrong. These error signals are then aggregated and weighted by inferential mechanisms incorporating prior beliefs and reasoning processes to decide to self-attribute or reject the false feedback. This is consistent with several models of the sense of agency suggesting that similar component are involved in self-attributing our actions and their results in the world (Moore & Fletcher, 2012; Synofzik et al., 2008).

According to this monitoring and inference framework, several scenarios can lead to the failure to detect manipulations. Error signals produced by monitoring mechanisms may be too small to reach detection. This can happen because of limited resources allocated to monitoring or limited precision of monitoring mechanisms. Another scenario involves the inferential component of choice blindness. For example, inferences can underweight monitoring-derived error signals based on a judgment that manipulations are unlikely to occur. In line with this, in Paper 1, the later the first manipulation happened in the implicit experiment, the less likely participants were to detect it. This may reflect the development of a stronger trust in the choice feedback in the experiment. In Paper 3, other ways in which reasoning can influence outcome attribution are suggested. In this study, participants mostly rejected the outcome they wanted when they reached it through an erroneous action. This might be explained by the

participants using a heuristic judgment such as: "if my cursor went to the wrong position, this cannot be my choice".

Another way in which I tried to improve our understanding of the mechanisms of choice blindness is by using neuroimaging. I decided to use fMRI for several reasons. Although its temporal resolution is lower than say EEG or MEG, it provides a much better spatial resolution and can capture signals from subcortical areas such as the basal ganglia, cerebellum, or midbrain. The basal ganglia and the midbrain, as part of the dopaminergic system, have been shown to be involved in outcome monitoring in the decision-making literature (Bartra et al., 2013; Clithero & Rangel, 2014; Jauhar et al., 2021). This was hence an important candidate monitoring mechanism to investigate in the context of choice blindness. In addition, the available equipment we had at Lund University (7T scanner) promised to have a better signal to noise ratio than EEG systems, allowing to study choice blindness without having to exceedingly increase the number of manipulated trials.

fMRI measures brain activity indirectly by detecting changes in blood flow. Typically, when neurons are active, a physiological response increase the blood flow to provide them with more oxygen. This increase in oxygen level in blood is detectable as an increased MRI signal This response is named the blood-oxygen-level-dependent (BOLD) response. More precisely, oxygen level decrease slightly right after neuronal activity and then increase sharply until reaching a peak after 4-6 seconds (Huettel et al., 2014). With this knowledge, we can analyse which brain areas respond to stimuli presented during the experiment in a way compatible with the BOLD response, suggesting their association with cognitive processes related to the task at hand.

In my fMRI study, I showed that false-feedback detection was associated with a wide range of brain activations. This included areas associated with reward monitoring (midbrain, basal ganglia, insula, medial prefrontal cortex), sensory prediction and the sense of agency (superior temporal sulcus, supplementary motor area, angular gyrus, precuneus). Aside from these activations that can be related to monitoring, I also observed activations in a frontoparietal network compatible with the involvement of executive functions and reasoning (rPFC, dlPFC, AG). This is consistent with the monitoring and inference framework I outlined above and in paper 1, as both monitoring and reasoning related activations appeared.

4.4. Reasons why introspection is interpretative

In sum, there is an impressive array of evidence supporting the idea that introspection is interpretative in nature, and there are many cognitive models trying to account for it. Hence, it is worth pondering: why do we need to self-interpret in the first place? Wouldn't it be "easier" to have a dedicated monitoring mechanism allowing us to directly access the content of our mind? Why would we not have it?

To explore the reasons why introspection would be interpretative, it is worth to return to the useful yet controversial distinction between content and process drawn by Nisbett and Wilson (1977). They suggested that only the results of our cognitive operations reaches our consciousness, but not the various processing steps that produced them. A consequence is that the introspectable contents reaching our consciousness are not the cognitive processes or operations themselves, but their products. This is an influential intuition that we find in many cognitive models. For example, in the global neuronal workspace model of consciousness, most basic cognitive operations are performed by specialized, encapsulated modules whose operation cannot be accessed by other modules (Dehaene et al., 1998; Mashour et al., 2020). Only their output can be sent in a global neuronal workspace akin to working memory, and then widely broadcast to other modules, thus becoming "conscious".

Another way to put it is: *consciousness does not perform computations*. I would call this the "passivity of consciousness" hypothesis. Consciousness would only maintain the representation of a content active, so that it can be further processed. But these contents are produced and computed by unconscious specialized modules. In other words, there is no conscious operations or actions; no "conscious processing" module. Consequently, there is nothing like consciously "judging" or "deciding". Various unconscious modules contribute to the computation of a judgment or a decision. What reaches consciousness is solely (some of) their outputs and inputs. More precisely, that is to the extent that a conscious content (e.g. mental image, or an inner speech sentence) is actually used as an input by a decision or judgment module that it may reflect part of a cognitive process. But which use is actually made of a conscious content (if it is used at all) lies beyond our introspective reach. It can at best be indirectly inferred from the following conscious events and the resulting decision/judgment. This *passivity of*

consciousness is at the heart of the apparent causation model (Wegner & Wheatley, 1999) and the illusion of conscious thought hypothesis (Carruthers, 2017).

This approach explains the need for self-interpretation by the modular architecture of the mind (see figure 2). Given how the brain is built, there is no other method available for introspection than interpretation. But other complementary explanations have been put forth. For example, self-perception and the "theory theory" of mentalizing claim that our introspective talk are grounded in mentalistic concepts and theories (e.g. attitudes, beliefs, desires…). Hence, introspection relies on applying learned categories and concepts, and on making use of all the categorization and reasoning apparatus that it requires. Interestingly, that means that the presence of mentalistic terms in an introspective report is a good indicator that self-inference has taken place. But it also opens the possibility of a less theory laden introspection, for example by training people to only describe their conscious experience, as objectively as possible, as micro phenomenology or early introspectionist psychological schools suggest (Brock, 2013; Petitmengin et al., 2019).

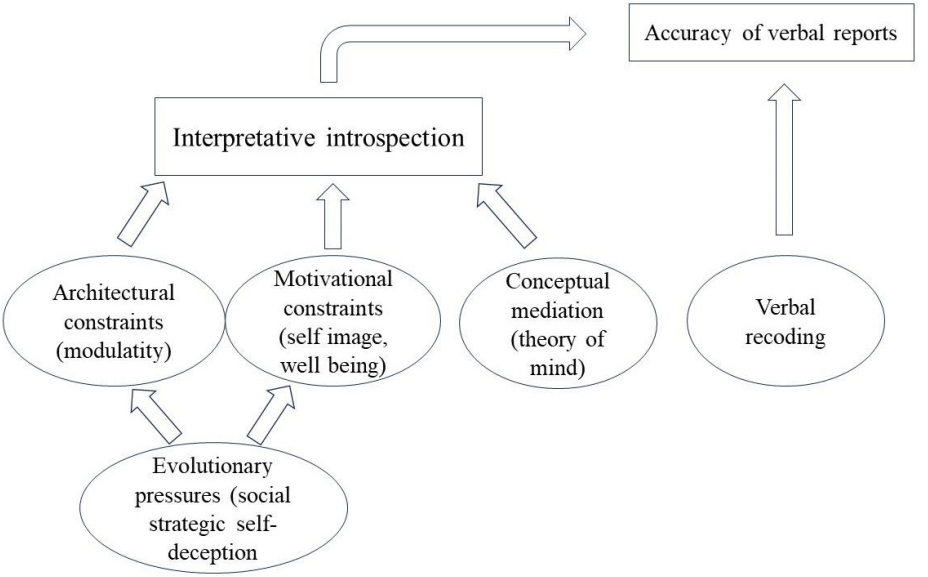


Figure 2. Possible causes of the interpretative nature of introspection, and their relationship with verbal reports accuracy.

A closely related reason is the need for linguistic recoding of nonverbal cognitive processes for social communication (Ericsson & Simon, 1980; Schultheiss & Strasser, 2012). This explanation does not really deny the possibility of a dedicated introspection/monitoring mechanism. Even if such a mechanism existed, one would need to translate the first order cognitive information to a linguistic format, a step that may induce errors or information loss. However, it is important to note that, although this perspective can explain inaccuracies in introspective reports, linguistic recoding does not necessarily entail inferential introspection (Ericsson & Simon, 1980, see figure 2).

Additionally, as suggested early on by Freud and contemporary experimental research, there can exist motivational limits to self-knowledge (Erdelyi, 1985). That is, one may be motivated not to be aware of specific mental contents or intentions, as it would prove detrimental, for example, for one's self- image, social standing, or psychological comfort. Wilson and Dunn (2004) reviewed evidence suggesting the existence of various mechanisms such as repression, suppression, intentional or complete forgetting. Hence, it is also possible that some mental contents could, in principle, be accessible to introspection, but fail to be accessed for self-serving reasons. This would suggest that some introspective inference may be biased in a self-serving way, especially when the stakes are high in for example social interaction.

This fits with a recently developed evolutionary strategic account regarding the limits of introspection (Simler & Hanson, 2017; Trivers, 2011). According to this perspective, the lack of direct access to one's motives is not a flaw but a cognitive adaptation that evolved to make us better at deceiving others. Simler and Hanson (2017) noted that as social individuals, we have a drive to fulfil our self-interest, while at the same time displaying prosocial and altruistic intent. An important part of the "social game" would consist in getting our selfish goals satisfied as much as possible, while avoiding being recognised as a selfish person unworthy of cooperation, which could lead to severe consequences such as ostracism. We would sometimes need to navigate social interaction by deceiving others about our true intent, hiding them or lying. Trivers (2011) suggested that a profitable strategy to deceive others is deceiving oneself. Lying comes at a cognitive cost and is often associated with tale tell signs, such as nervousness or difficulties in keeping tracks of stories that contradict each other. To discreetly pursue our deceitful intent, an efficient strategy is not being aware of them at all. Simler and Hanson

(2017) reviewed a wide array of evidence supporting this view, although clear cut, controlled experimental evidence would be desirable (see section 6.). It is worth noting that the evolutionary strategic account is compatible with other accounts previously mentioned. The brain may have evolved or exploited a modular architecture to keep selfish motives secret to other modules and consciousness. Similarly, our introspective inferences may be strongly biased by self-serving factors, taking into account social impression and strategic goals into account whenever we make an introspective report.

So far, we have seen that interpretative models of introspection have a lot of empirical support and are well grounded in current theories of cognition. One may wonder: what is the state of the alternative models supporting a direct access to our cognitive processes?

5. The alternative: direct access models of introspection

Maybe unsurprisingly, given the abundance of evidence for interpretative models, there are few plausible non-interpretative models of introspection. For example, even the landmark paper of Erikson and Simons (Ericsson & Simon, 1980), seen as the main critics of Nisbett and Wilson about the reliability of verbal reports, embraces the same consensus about introspection. They clearly stated that: "we will not assume that the verbalized description accurately reflects the internal structure of processes or of heeded information, or that it has any privileged status as a direct observation" p.217. Their aim was to model how introspective reports are produced and describe conditions in which they could accurately reflect underlying cognitive processes.

Ericsson and Simon's (1980) purpose was not to postulate a dedicated monitoring mechanism, but to assess how and under which condition verbal reports can be reliable indicators of cognitive processes. They drew a more optimistic yet humble conclusion than Nisbett and Wilson (1977), that is: "verbal reports, elicited with care and interpreted with full understanding of the circumstances under which they were obtained, are a valuable and thoroughly reliable source of information about cognitive processes" (Ericsson & Simon, 1980, p.247). To solicit the most accurate reports possible, they should be produced i) concurrently to the task at hand or very soon after the process occurred (<5s) while the information may still be in working memory, ii) we should be wary that the verbalization process does not add excessive cognitive load, iii) avoid asking questions that require filtering of information or inference (such as why questions) and rather focus on description of processes that are already verbal or only require linguistic recoding, iv) ask participant to think aloud, or avoid probes about hypothetical or general states, v) use undirected probes that don't hint at how the experimenter expect the cognitive process to operate.

Interestingly, their position is in line with many interpretative models of introspection. They claim that we can access the content of our working memory, which includes inner speech, visual or other sensory contents. This fits neatly with interpretative models and the neuronal global workspace of consciousness, which share the same assumption (Dehaene et al., 1998). In this model, the fine-grained steps of our cognitive processes tend not to be directly accessible either, and knowing what role if any a conscious content plays in an underlying cognitive

process is not a given, but something that can only be inferred, as the "passivity of consciousness" hypothesis and other models assume.

Ericsson and Simons (1980) reviewed extensive evidence showing that asking for verbal report tend not to add substantial cognitive load nor significantly impacts the underlying cognitive process. In addition, they reported some experiments demonstrating consistency between verbalizations and behaviour. For example, in a rule-based card sorting task, a vast majority of participants made decisions that were consistent with the rule they claimed to use (Dulany Jr. & O'Connell, 1963; S. H. Schwartz, 1966). As we mentioned, interpretative models are also compatible with the hypothesis that conscious content may reflect part of the underlying cognitive processes, although indirectly.

Other approached the debate on introspection by criticizing the extent to which our cognitive processes are unconscious (Newell & Shanks, 2014). However, Newell and Shanks (2014) did not provide an alternative model of how introspection works. They expressed interesting concerns about the extent to which our cognitive processes are unconscious, but they did not go as far as giving quantitative claims about how much of our lower order or higher order cognitive processes are accessible to consciousness. Nor do they account for the basic observation that we don't introspectively know anything about the details of the algorithms and information processing routines our brain uses.

A clear criterion to know which processing steps are available or not to consciousness still needs to be defined. It is uncertain whether there exists a general answer to this question, as it may depend on the specific cognitive process at play. However, I may speculatively suggest one tentative approach. According to the global workspace theory, awareness of information is determined by the need to share it with other modules for additional processing (Dehaene et al., 1998). One may approach this question abstractly or, rather, computationally. For example, given a specific algorithm, what is the cost and benefit of making an intermediate step available to other modules? For example, in judgment and decision making, what part of evidence sampling and value evaluation would benefit enough of language processing to be shared with a language module? What would be the cost of sharing a specific intermediate result in terms of axonal connexions and energy use? One can speculate that the degree of "encapsulation" that exists in the brain can be traced to optimal cost-benefit trade-offs refined over the long-time of nervous systems' biological evolution. This conjecture remains to be investigated.

Another non-interpretive empirically informed model of introspection is the dual-method model of Nichols and Stich (Nichols & Stich, 2003, see figure 3). It suggests that two methods can operate for introspecting one's mental states. One uses the interpretative, mentalizing resource generally used for social cognition, while the other one would use a specific monitoring mechanism which has a direct access to the content of a so called "belief box", that one can see as a sort of short- or long-term memory of propositional attitudes, generally seen as being in a linguistic or quasi linguistic format (Nichols & Stich, 2003). However, this model may be hard to fit with the current state of knowledge in cognitive science. For example, much of the content of our brain is not stored in a linguistic format and neural network/connexionist models suggest that non-symbolic encoding is very common (Harris, 2006; Perlovsky & Sakai, 2014).

Even conceptually, an attitude may not be appropriately seen as a representation stored in one's own mind. An attitude is a conceptual construct integrating evaluations, beliefs, emotions and action tendencies (VandenBos, 2007). An attitude about, for example, supporting the green party can only be an inference, as it is a generalization of our past and present evaluations and behaviours towards the green party (and it appears that this very introspection can be biased by priming recollection of past specific past behaviours, see Chaiken & Baldwin, 1981).

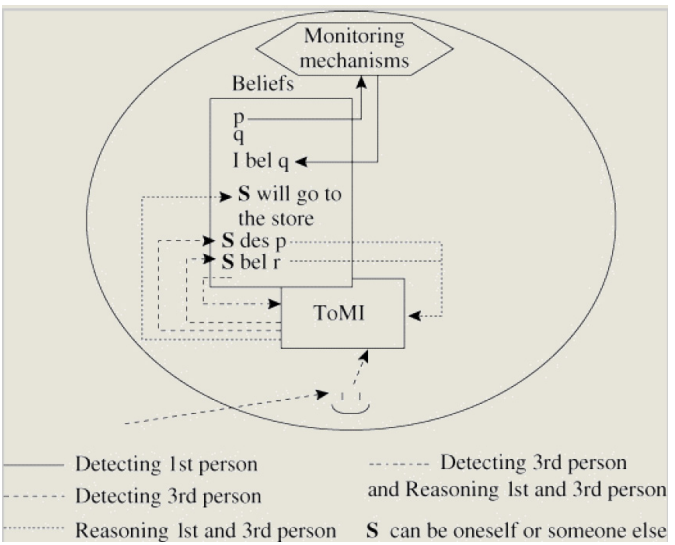


Figure 3. Dual method model of Nichols and Stich (2003). Reproduced with permission of Oxford University Press (© OUP).

One may see this "propositional monitoring" mechanism suggested by Nichols and Stich as something more akin to a model of introspection of overt or inner speech. In which case it might be reduced to interpretative models or Ericsson & Simon's (1980) model of verbal reports. Indeed, it is assumed by most models that we mentioned that we have access to our conscious inner speech. We can also retrieve our past linguistic utterances from episodic memory. However, these linguistic productions are mere cues of intermediary product of ongoing cognitive processes. For example, the fact that I think "that's an interesting proposition" while listening to a political speech is a good clue that i) I approve the proposition, and ii) that this proposition may influence my voting decision. However, in which way (if any) this proposition played a role in my actual voting decision and the evidence accumulation process that led to it does not seem to be introspectable. I don't know by how much hearing this proposition increased my political preference. And why exactly did I find this very proposition appealing? Which of my goals (my own benefit, fitting in my peer group, signalling my altruism or intellectual autonomy) gave value to this proposition? Even further, I may actually think "that's an interesting proposition" in a sarcastic way, or it may just express respectful disagreement. Again, the epiphenomenal and passivity of consciousness hypotheses are relevant here as well.

A possible suggestion would be that propositional attitudes in Nichols and Stich's model would not be stored in a linguistic format. However, we would still need to assume the existence of conceptual content about one's own attitudes (e.g. believing, liking, trusting, etc...) and a role for categorization which would intervene as an intermediary step in the introspection process. And, it is unclear how this intermediary step would stand as a case of direct access. In my opinion, this possible aspect of their model is not spelled out in detail and is difficult to interpret.

Despite being unsatisfactory in its current form, the dual method theory contains an element which has been rather influential in several domains of cognitive science: the notion of a monitoring mechanism which directly gets information from the inner processing of a specific cognitive module. It is worth noting that the operations of this monitoring mechanism don't have to be conscious, as long as the result of this processing is. You don't need to know how you know that you like something, as long as the information of how much you like something is represented somewhere and measured by some monitoring mechanism. This view

had quite a success in the study of metacognition, with the proposition that we derive our confidence from monitoring specific variable of our decision-making process. Interestingly, the two main alternatives mostly considered in the literature (decisional and post decisional locus theories of confidence) both postulate a direct monitoring mechanism, while the inferential approach is sometimes surprisingly neglected (Baranski & Petrusic, 1998; Fleming, 2024; Yeung & Summerfield, 2014; Zylberberg et al., 2012). Decisional locus theories of confidence judgments try to map confidence judgments to the monitoring of specific variables during the decision process within a diffusion drift framework. For example, one may base one's confidence judgment on monitoring the time to reach a decision, the balance of evidence, or the quantity of evidence discounted by the time used to reach it (Yeung & Summerfield, 2014). By opposition, post decisional locus theories suppose that confidence derives from the monitoring of decision variables after the decision is reached, such as how much evidence continues to be accumulated after a decision threshold is attained.

However, according to interpretative models of introspection, it could very well be the case that confidence also stems from inferences based on indirect cues. For example, Kornell (2014) and Schwartz et al. (1997) argued that direct access models tend to be disconfirmed in the field of memory metacognition. In these studies, it appears that cues such as fluency, ease of processing and familiarity strongly influence confidence judgments about memory, when actual memory quality is held constant (Alter & Oppenheimer, 2009; Kornell, 2014a; Undorf & Erdfelder, 2011). There is also experimental evidence that memory and confidence can be negatively correlated, with the final judgment relying more on ease of processing than memory strength (Benjamin et al., 1998; Besken & Mulligan, 2013). Kornell (2014b) went even one step further expressing a "healthy scepticism about what it is, exactly, that makes metacognition different from other situations in which animals respond based on complex cues" (p.160). In this view, metacognition would not be much different from regular inferences that we draw while reasoning and trying to answer various question.

To me, this is a very important point to stress. Indeed, various evidence points at the hypothesis that higher order cognitive abilities categorised as separate cognitive functions actually rely on very similar - if not the same - neurocognitive mechanism. For example, metacognition, the sense of agency, reasoning as well as mentalizing have been shown to rely on highly overlapping brain regions (Vaccaro

& Fleming, 2018; Valk et al., 2016; Van Overwalle, 2011, see table 3 about sense of agency and reasoning). It may very well be the case that all these cognitive abilities are just special cases of general reasoning and judgment processes.

Table 3. Overlap between neural correlates of reasoning and the sense of agency

	Insula	**TPJ**		**SMA/ preSMA**	**Precuneus**
	BA13, BA47	Inferior parietal BA39, BA40	Superior posterior temporal BA41, BA42, BA22	BA6	Medial BA7
Sense of agency	✓1	✓1	✓2	✓1	✓1
Reasoning	✓3	✓4	✓5	✓6	✓7

1 (Charalampaki et al., 2022; Haggard, 2017; Seghezzi et al., 2019; Sperduti et al., 2011). **2** (Charalampaki et al., 2022; Haggard, 2017; Sperduti et al., 2011). **3** (Fugelsang & Dunbar, 2005; Hobeika et al., 2016; Luo et al., 2003; L. Wang et al., 2020). **4**. (Blos et al., 2012; Brzezicka et al., 2011; Fangmeier et al., 2006; Hobeika et al., 2016; L. Wang et al., 2020; Wertheim & Ragni, 2018; Woods et al., 2014) **5** (Blos et al., 2012; Fangmeier et al., 2006; Luo et al., 2003) **6** (Blos et al., 2012; Brzezicka et al., 2011; Fangmeier et al., 2006; Fugelsang & Dunbar, 2005; Hobeika et al., 2016; L. Wang et al., 2020; Wertheim & Ragni, 2018) **7** (Brzezicka et al., 2011; Fangmeier et al., 2006; Fugelsang & Dunbar, 2005; Hobeika et al., 2016; Wertheim & Ragni, 2018; Woods et al., 2014)

6. Conclusion and future directions

This state-of-the-art overview has shown that a diverse and wide range of evidence supports interpretative models of introspection, while few speak for direct access models. However, this by no way means that our understanding of introspection is full and complete. Many open and exciting questions remain to be investigated. Here, I take the opportunity to summarize what I think are the most important unexplored tracks for future research.

6.1. The need for more specific computational models of introspection

A first need for introspection research is increasing the precision of introspective models. As we saw, models and data of interpretative introspection are abundant and varied. However, they mostly remain at a quite general, verbal, level of description. That is, they don't propose a precise mathematical model. As philosophers of science have suggested (Hempel, 1966), quantity and variety of evidence is only one part of what's needed to support a theory. The precision of the predictions - hence of the underlying model - matters a lot and remain, in my opinion, the main aspect in need for improvement in introspection research. A first step in the direction of building computational models of introspection can be found in the Bayesian cue integration model of the sense of agency (Moore & Fletcher, 2012). However, as we mentioned in Paper 1, 2, 3, this model has not been properly tested empirically yet. A natural approach would be to follow the same steps that were followed in the judgment and decision-making literature, by comparing this "optimal inference" model to "satisficing" heuristic models (Brandstätter et al., 2006; Gigerenzer & Goldstein, 1996). As I see it, the fields germane to introspection (sense of agency, metacognition, theory of mind) would really benefit from being considered not as specialized cognitive functions, but as special cases of reasoning and judgment processes. To support this view, I highlighted the extensive overlap between the neural correlates of these cognitive activities and reasoning (see section 4, table 3). From this perspective, it is a fair assumption that these fields would eventually merge.

A consequence of increasing modelling precision is that we may need specialized models for specific target "mental states" or processes to be introspected. Only one general model may not be suited to account for introspection of all mental states (e.g. attitudes, emotions, preferences, beliefs, etc..). For example, as I

mentioned above in the discussion of the dual method model (section 4), attitudes may not really be mental states, or some representation at all. They would rather be generalizations of past behaviour and evaluations. So, their "introspection" would most likely need to involve a memory sampling/retrieval component. Alternatively, if one wants to study introspection of current decision-making process, long term memory does not have to be involved. Similarly, reports about longstanding preferences versus ongoing evaluations during a decision process (i.e. evidence accumulation in a diffusion drift model) most likely rely on different processes, despite the fact that both these activities could be loosely labelled "introspection of preferences". Indeed, asking questions about longstanding preferences involves generalization from past behaviour. By opposition, introspecting ongoing evaluations does not have, in principle, to rely on generalization from long term memory, although, it may rely on inferences based on, for example, current bodily sensations, content available in consciousness, causal theories, etc...

6.2. Modelling choice blindness: detection, confabulation, and preference change

With respect to modelling introspection, this brings us back to Paper 1 and my preliminary attempt at modelling choice blindness. As mentioned above, choice blindness has three components: 1) Detection, 2) Confabulation, 3) Preference change (Strandberg, 2020). In Paper 1, I attempted to account for the first component, i.e. what drives false feedback detection. As this framework is the first of its kind, it remained at a general level, supposing that failures to detect false feedback stem from the interaction of monitoring and inferential mechanisms. An important point here is that a model of detection in choice blindness would not only involve introspective mechanisms, but also outcome monitoring and memory monitoring mechanisms. The mere *detection* of false feedback could in principle occur without any introspective access to one's preferences and attitudes. A system would only have to be able to detect unexpected change in the environment to detect manipulations (e.g. face A should have been there, but now it's face B).

Currently, an open question is "which specific neurocognitive mechanisms are involved in the routine outcome monitoring that leads to the spontaneous detection of false feedback in choice blindness?". In Paper 1, I suggested a general

principle of cue integration and reasoning. I also suggested various possible monitoring mechanisms: 1) prediction of reward value by the basal ganglia (Alexander & Brown, 2011), 2) checking if some abstract environmental conditions satisfy the system's goal with so-called TOTE units (test-operate-test-exit) (Botvinick, 2008) 3) non motor based sensory predictions (Dogge et al., 2019), 4) incidental detection of incongruences with memory (long term or working memory). The brain bases of detection that I reported in paper 2 are consistent with all these hypotheses. We saw the activation of the basal ganglia (reward prediction), insula, superior temporal sulcus and lateral parietal cortex (sensory prediction, sense of agency), rostral prefrontal cortex (memory monitoring), and lateral prefrontal and lateral parietal cortices (working memory, executive functions, reasoning). Hence, it is possible that false feedback detection may rely on a whole set of monitoring and reasoning mechanisms. More precise research aimed at teasing their contribution apart would be required.

In addition, using choice blindness to model confabulation and preference change are very promising and relatively unexplored avenues. One may conceive of confabulation as a search for causal factors that can explain one's behaviour. In this vein, it could be interesting to try to model how people rely on their long-term memory, current perception, or self-models/narratives to find causal candidates for their choices. To test how causal factors are searched in confabulation, one approach would be to create a more controlled situation, in which people's possible reasons are more limited. A multi attribute decision making task would be one way to control the factors that people may mention during their confabulation and explore what influences their choices of factors while confabulating. One could also suggest or prime specific causal theories to participants or make a possible reason more salient and observe if it influences people's confabulations. The question of what impacts (and in which way) people's confabulatory reports in choice blindness has been virtually unexplored.

Confabulation is the component of choice blindness that is the more clearly related to introspection. However, as discussed in Paper 4, choice blindness induced confabulations not only involves introspection, but also misinformation. That is, in the choice blindness paradigm, participants don't produce confabulation spontaneously; they confabulate because they are given false information about their true choice. I think that this is one plausible explanation of the results observed in the fMRI study of confabulation in Paper 4. Contrary

to what a radical self-perception theorist would expect, several differences emerged at the neural level between confabulation and non-confabulation; i.e. if confabulation and introspection relied on the same mechanism, they should also involve the same brain regions. However, one possible explanation for this result is that the misinformation used in choice blindness creates a conflict between true memories and false beliefs about one's own choice, resulting in a more effortful rationalization. Longer reaction times as well as the involvement of areas related to memory monitoring (rPFC) and executive function (dlPFC) in confabulation are consistent with this interpretation (see Paper 4).

I also want to stress that the results in Paper 4 are more consistent with self-perception than direct-access theories. The fact that no brain area was more activated in the non-confabulated condition suggests that there is no specific introspective mechanism at play when people don't confabulate. Indeed, direct access theories would predict specific activations related to direct-access introspection in the non-confabulation condition, which would contrast with the interpretation-related activations in the confabulation condition. To me, the most parsimonious interpretation is that the activations in the mentalizing network (TPJ, mPFC, precuneus) during confabulation reflects an increased activation of interpretative mechanisms which are already at play in the non-confabulation condition, and not the recruitment of a different set of mechanisms. This would be consistent with the numerous studies showing that self-related processing also relies on the mentalizing network (Denny et al., 2012; Kestemont et al., 2015; Legrand & Ruby, 2009). In addition, when I tried to contrast both confabulation and non-confabulation to various baselines which either did not involve mentalizing or involved thinking of other's mental states, overlapping activations of the mentalizing network appeared in confabulation and non-confabulation. Hence, it would seem that both confabulation and non-confabulation recruit the mentalizing network, but it would be more activated when people confabulate. However, I admit that this interpretation of my results is not clear-cut and still tentative. Developing a formal model breaking down the different processing steps at play in direct access or interpretative introspection may help formulating more precise predictions of brain activity and drawing more definitive conclusions.

In future studies, one possible way to investigate more precisely the question of whether confabulation and non-confabulation rely on the same exact neural correlates would be to remove the misinformation factor which was confounded

with confabulation in my study. To do so, one would need to study spontaneous confabulation, without misinformation. This would require developing a new methodology to assess the validity of introspective reports. Indeed, what is the strength of the choice blindness paradigm may also be its weakness in this context. The way we make sure that participants reports are wrong is by presenting them with an option they did not choose as their true choice. That is, in choice blindness, the assessment of verbal reports always relies on misinformation.

However, if we could study confabulation without the misinformation component that exists in choice blindness, it could still be possible that the radical predictions of self-perception theory hold. That is, no brain-based differences would be observable between confabulated and non-confabulated reports. Below, I mention some possible tracks to spontaneous confabulation.

6.3. Beyond choice blindness: spontaneous confabulation and strategic self-deception

In my opinion, the strategic self-deception theory provides a promising framework to investigate spontaneous confabulation by telling us in which contexts confabulation is the most prone to occur spontaneously. The strategic self-deception theory gives an interesting new perspective: confabulation and lack of direct introspective access are not a flaw of our cognitive machinery, but a beneficial adaptation. This evolutionary hypothesis has been spelled out only recently (Hippel & Trivers, 2011; Kurzban, 2012; Trivers, 2011), however its root can be traced back to work on mixed-motive games in economics and game theoretics by the Nobel prize winner Schelling (Schelling, 1960; Simler & Hanson, 2017). Mixed games are special situations where agents' interests partly overlap and partly diverge. These games, typical of social situations, are ripe for behaviours close to self-deception. To take Simler & Hanson (2017)'s examples, a general may have incentives to purposefully adopt the false belief that his army can win in order to intimidate his enemy. The point is that holding false belief can prove highly beneficial in social games, which can provide an incentive, if not an evolutionary pressure, for self-deception.

What is the state of research on the strategic deception hypothesis? As I mentioned, the hypothesis is fairly new. In my opinion, an increasing amount of data is compatible with this hypothesis. However, there is a need for more clear-

cut and controlled experiments. Simler and Hanson (2017) reviewed an impressive array of phenomena that may be accounted for by self-deception. For example, excessive health-related expenses may be seen as a signalling goal of conspicuous caring. Doctors would tend to prescribe more medicine than strictly required to signal that they care for their patient. This would stem from a need to show that we care for our ally, even if our actual contribution or their needs may be negligible. Simler and Hanson reviewed evidence for such signalling and self-deception related behaviour in political, religious, environmental, educational, charity donation domains. To me, they may not provide conclusive evidence, but they make a strong case for strategic self-deception as very viable and promising hypothesis.

On the experimental side, a few attempts have been made, although operationalizations of self-deception may still be tentative. In the first empirical paper on self-deception, "self-deception: a concept in search of a phenomenon", Gur and Sackeim (1979) tried to show that, on simple tasks such as recognizing one's own voice, people acted in a self-deceptive way. They wanted to show that when people failed to recognize their own voices, they were respecting four criteria of self-deception. That is, they i) were holding two contradictory beliefs (this is my voice and not my voice), ii) simultaneously, iii) were unaware of holding one of them, and iv) this unawareness was motivated. They argued that the information that the voice they heard was their own was sometimes represented even when they failed to detect it using galvanic skin responses, that tend to be higher when one perceives her own voice.

Attempts at studying self-deception in social context sometimes used questionnaires such as the lie acceptability scale or the self-deception scale of the Balanced inventory of desirable responding (Lynch & Trivers, 2012; Wright et al., 2015). As an interesting side note, investigations of social desirability in choice blindness used other questionnaires (the Marlowe–Crowne Social Desirability Scale [Crowne & Marlowe, 1960], or the SDS-17 [Stöber, 2001]), but not the BIDS, which include the self-deception scale (Paulhus & Reid, 1991). Hence, whether some components of choice blindness (e.g., detection, confabulation) correlate with self-deception tendencies remains an interesting and open question.

Other operationalizations of self-deception have equated it with overconfidence and biased information search, in the context of persuading others (Anderson et al., 2012; Kennedy et al., 2013; Schwardmann & van der Weele, 2019; Smith et

al., 2017). However, these phenomena don't exactly capture the core of the idea of self-deception, that is: two contradictory beliefs need to be held, and only one reaches awareness for strategic purposes. Similarly, Pinker (2011) argued that it is important to distinguish error and biases from self-deception, which is not always clearly done in this emerging literature.

To fill this gap, a new experimental approach based on the self-insight methodology described in section 3.2. would be valuable. The basic idea of this methodology is to let people evaluate the weight they gave to various attributes in a decision-making context and compare these with the objective weights derived from a regression model ran on their actual choices. We could for example ask people to choose between different charities based on multiple attributes. After each choice, people would rate each attributes importance in their decision. One or few attributes would be socially undesirable (e.g. signalling, status, attractiveness enhancement). To show that people's introspection is distorted in a motivated way, we could show that people tend to underweight socially undesirable attribute more than neutral ones. We could compare a condition in which people's self-evaluation are private VS public, even possibly discussed with a confederate. We would expect that socially undesirable attributes would be even more underweight in public situations. Further studies with neuroimaging such as fMRI would be valuable to see if two contradictory beliefs can be decoded from brain activity.

6.4. Choice blindness induced preference change

In several parts of the thesis, I have mentioned the third component of the choice blindness methodology: how it allows for the study of preference and attitude change. Many studies have shown that choice blindness manipulations can change people's facial attractiveness preferences (Johansson et al., 2014; Mouratidou et al., 2022; Pärnamets et al., 2020; Remington et al., 2024; Taya et al., 2014; Vogel et al., 2023 [paper 1]), aesthetic preferences (Mouratidou et al., 2022), risk preferences (Kusev et al., 2022; Muda et al., 2020), financial preference (McLaughlin & Somerville, 2013), and political attitudes (Hall et al., 2013; Strandberg, 2020; Strandberg et al., 2018). It has also been used to show that it is possible to change people's symptom report or their responses about experiential avoidance in the health domain (Artenie et al., 2023; Merckelbach et al., 2011b).

These changes are typically assessed by analysing how people change their later choices (i.e. choice consistency), their preference or attitude ratings, or their verbal reports after failing to notice false feedback on their choices.

This is certainly an important component of choice blindness. It shows that faulty introspection can have downstream effects and influences our later choices and behaviour. To some degree, who we *think* we are impacts who we actually are. This aspect has been stressed by agencialist accounts of self-knowledge in philosophy (Moran, 2001). Properties of the self are not set in stone, and one may change oneself through its actions. Hence, even if our judgments about ourselves are wrong in the moment when we formulate them, we have to some extent the ability to make them right by acting in a way that is consistent with them later on (Moran, 2001). Although this facet of choice blindness was not at the centre of my project, it surfaces as a natural side-effect. All my studies included an assessment of preference change through the measure of choice consistency, showing how failures to detect manipulation increases the likelihood to later make choices against our initial preferences.

For example, the fMRI study of choice blindness also included an assessment of preference change. Interestingly, the largest preference change effect was observed in this study: when comparing non-detected and non-manipulated trials, we saw a 25% change in preference consistency (see figure 4).

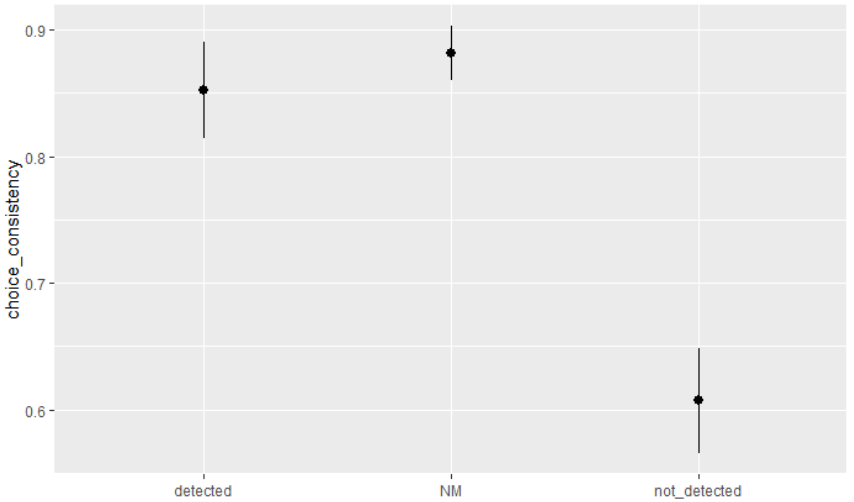


Figure 4. Choice consistency in the fMRI experiment. NM stands for non-manipulated trials.

In other projects that were not included in this thesis, I have looked further at preference change in relation to choice blindness. For example, I collaborated on a project which studied whether choice blindness could change *collective* preferences in dyads (Pärnamets et al., 2020). Initially, the participants teamed up in pairs. The dyads were then shown two pictures and were tasked to jointly choose one person as a roommate based on the picture. First of all, we found that a large proportion of the manipulations remained undetected, despite the dyad discussing and explaining the reversed outcome of the choice (what we call "collaborative confabulation"). But our main result was an interesting dissociation between group preferences and individual preferences. Manipulations that were collectively accepted led to a preference change when the dyad was making a collective choice at a later stage. However, when the members of the dyad were separated when making the second-round choice, their individual preferences remained unaltered, despite having changed at the collective level.

In an ongoing work, I am also studying more specific properties of the preference change effect. I wanted to investigate whether the choice blindness-induced preference change would respect the principle of transitivity of preference. The transitivity of preference means that if you prefer A to B and B to C, then you should also prefer A to C. The transitivity of preference is a fundamental axiom of economic theories of decision making (i.e. expected utility theory). Failure to abide by this principle could, it is argued, lead to irrational behaviour; for example, people could be exploited through money pumps (infinite loops of spending). To test if people would really respect the axiom of transitivity, we attempted to implant intransitive preferences in people's mind, using a choice blindness manipulation on specific trials where a change of preference would lead to an intransitive preference pattern. Although the project is not finished yet, preliminary results suggested, to my surprise, the existence of a resistance to intransitivity. When participants changed a preference relation that should have created an intransitive pattern, they tended to change their other preference relations in order to preserve the transitivity of their preference pattern. Oppositely, when a preference relation with no bearing on transitivity changed, the other preference relations did not change much. This might suggest that preferences and their change are constrained by certain structures and principles, among which may be transitivity (Zander, Vogel, & Johansson, 2023).

It is beyond the scope of this thesis to properly review the postulated mechanisms of preference change in the scientific literature, and to assess how our result might fit or not with the leading theories. But, in brief, I will mention a few theories of why preference change arises through choice. First of all, preference change might stem from cognitive dissonance (Festinger, 1957). According to this, people would attempt to resolve the conflict between their false beliefs about their choice and actual preferences by changing their preferences accordingly (Harmon-Jones & Mills, 2019). Self-perception theory on the other hand would suggest that we infer our own preferences by observing our own behaviour. Seeing us choosing an alternative (even though that what not the one we preferred) would lead us to infer a preference for this alternative an act consistently with this in the future (Brehm, 1972; Chammat et al., 2017). Memory could also play a role in preference change through choice blindness. For instance, the mere illusion of choice (when no choice was actually made) can provide a memory boost (Murty et al., 2015). Another interesting hypothesis that does not tend to be considered so much in the literature relies on reinforcement learning (Daw & Tobler, 2013). According to reinforcement learning models, preferences are updated after each choice, when its actual outcome is observed. The brain would compute the difference between the expected reward and the actual reward to update its preference. For example, if the reward was lower than expected, the agent would reduce her preference for the chosen alternative. It is an interesting possibility that preference change through choice blindness would stem from an incorrect updating of preference. Further modelling work would be required to spell out how this would specifically produce the choice blindness induced preference change. One may speculate, for example, that the prior value of the originally preferred alternative is misattributed to the unchosen one, which would increase the preference for it.

This concludes the theoretical introduction of my thesis. I hope that the reader will enjoy my papers included in the next sections.

References

Aardema, F., Johansson, P., Hall, L., Paradisis, S.-M., Zidani, M., & Roberts, S. (2014). Choice Blindness, Confabulatory Introspection, and Obsessive-Compulsive Symptoms: A New Area of Investigation. *International Journal of Cognitive Therapy*, *7*(1), Article 1. https://doi.org/10.1521/ijct.2014.7.1.83

Albarracín, D., & Wyer, R. S. (2000). The cognitive impact of past behavior: Influences on beliefs, attitudes, and future behavioral decisions. *Journal of Personality and Social Psychology*, *79*(1), 5–22.

Alexander, W. H., & Brown, J. W. (2011). Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience*, *14*(10), Article 10. https://doi.org/10.1038/nn.2921

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the Tribes of Fluency to Form a Metacognitive Nation. *Personality and Social Psychology Review*, *13*(3), 219–235. https://doi.org/10.1177/1088868309341564

Anderson, C., Brion, S., Moore, D. A., & Kennedy, J. A. (2012). A status-enhancement account of overconfidence. *Journal of Personality and Social Psychology*, *103*(4), 718–735. https://doi.org/10.1037/a0029395

Artenie, D. Z., Olson, J. A., Dupuis, G., Suisman, C. C., Casagrande, S. A. G., Akberdina, S., Roy, M., & Langer, E. J. (2023). Exploring the clinical utility of choice blindness: Generalization of effects and necessity of deception. *Psychology of Consciousness: Theory, Research, and Practice*, No Pagination Specified-No Pagination Specified. https://doi.org/10.1037/cns0000372

Baranski, J. V., & Petrusic, W. M. (1998). Probing the locus of confidence judgments: Experiments on the time to determine confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(3), 929–945. https://doi.org/10.1037/0096-1523.24.3.929

Bar-On, D. (2004). *Speaking my mind: Expression and self-knowledge*. Oxford University Press.

Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage*, *76*, 412–427. https://doi.org/10.1016/j.neuroimage.2013.02.063

Bem, D. J. (1972). Self-Perception Theory11Development of self-perception theory was supported primarily by a grant from the National Science Foundation (GS 1452) awarded to the author during his tenure at Carnegie-Mellon University. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 6, pp. 1–62). Academic Press. https://doi.org/10.1016/S0065-2601(08)60024-6

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General*, *127*(1), 55–68. https://doi.org/10.1037/0096-3445.127.1.55

Besken, M., & Mulligan, N. W. (2013). Easily perceived, easily remembered? Perceptual interference produces a double dissociation between metamemory and memory performance. *Memory & Cognition*, *41*(6), 897–903. https://doi.org/10.3758/s13421-013-0307-8

Blos, J., Chatterjee, A., Kircher, T., & Straube, B. (2012). Neural correlates of causality judgment in physical and social context—The reversed effects of space and time. *NeuroImage*, *63*(2), 882–893. https://doi.org/10.1016/j.neuroimage.2012.07.028

Botvinick, M. M. (2008). Hierarchical models of behavior and prefrontal function. *Trends in Cognitive Sciences*, *12*(5), 201–208. https://doi.org/10.1016/j.tics.2008.02.009

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The Priority Heuristic: Making Choices Without Trade-Offs. *Psychological Review*, *113*(2), 409–432. https://doi.org/10.1037/0033-295X.113.2.409

Briñol, P., & Petty, R. E. (2003). Overt head movements and persuasion: A self-validation analysis. *Journal of Personality and Social Psychology*, *84*(6), 1123–1139. https://doi.org/10.1037/0022-3514.84.6.1123

Briñol, P., & Petty, R. E. (2022). Self-validation theory: An integrative framework for understanding when thoughts become consequential. *Psychological Review*, *129*(2), 340–367. https://doi.org/10.1037/rev0000340

Brock, A. C. (2013). The history of introspection revisited. *Self-Observation in the Social Sciences*, 25–43.

Brzezicka, A., Sedek, G., Marchewka, A., Gola, M., Jednoróg, K., Krolicki, L., & Wrobel, A. (2011). A role for the right prefrontal and bilateral parietal cortex in four-term transitive reasoning: An fMRI study with abstract linear syllogism tasks. *Acta Neurobiologiae Experimentalis*, *71*, 479–495.

Carruthers, P. (2011). *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. Oxford University Press.

Carruthers, P. (2017). The Illusion of Conscious Thought. *Journal of Consciousness Studies*, *24*(9–10), 228–252.

Chaiken, S., & Baldwin, M. W. (1981). Affective-cognitive consistency and the effect of salient behavioral information on the self-perception of attitudes. *Journal of Personality and Social Psychology*, *41*(1), 1–12. https://doi.org/10.1037/0022-3514.41.1.1

Chammat, M., El Karoui, I., Allali, S., Hagège, J., Lehongre, K., Hasboun, D., Baulac, M., Epelbaum, S., Michon, A., Dubois, B., Navarro, V., Salti, M., & Naccache, L. (2017). Cognitive dissonance resolution depends on episodic memory. *Scientific Reports*, *7*, 41320. https://doi.org/10.1038/srep41320

Charalampaki, A., Ninija Karabanov, A., Ritterband-Rosenbaum, A., Bo Nielsen, J., Roman Siebner, H., & Schram Christensen, M. (2022). Sense of agency as synecdoche: Multiple neurobiological mechanisms may underlie the phenomenon summarized as sense of agency. *Consciousness and Cognition*, *101*, 103307. https://doi.org/10.1016/j.concog.2022.103307

Cheung, T. T. L., Junghans, A. F., Dijksterhuis, G. B., Kroese, F., Johansson, P., Hall, L., & De Ridder, D. T. D. (2016). Consumers' choice-blindness to ingredient information. *Appetite*, *106*, 2–12. https://doi.org/10.1016/j.appet.2015.09.022

Clithero, J. A., & Rangel, A. (2014). Informatic parcellation of the network involved in the computation of subjective value. *Social Cognitive and Affective Neuroscience*, *9*(9), 1289–1302. https://doi.org/10.1093/scan/nst106

Cochran, K. J., Greenspan, R. L., Bogart, D. F., & Loftus, E. F. (2016). Memory blindness: Altered memory reports lead to distortion in eyewitness memory. *Memory & Cognition*, *44*(5), 717–726. https://doi.org/10.3758/s13421-016-0594-y

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*(4), 349–354. https://doi.org/10.1037/h0047358

Daw, N. D., & Tobler, P. N. (2013). *Neuroeconomics: Chapter 15. Value Learning through Reinforcement: The Basics of Dopamine and Reinforcement Learning*. Elsevier Inc. Chapters.

Dehaene, S., Kerszberg, M., & Changeux, J.-P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences*, *95*(24), 14529–14534. https://doi.org/10.1073/pnas.95.24.14529

Denny, B. T., Kober, H., Wager, T. D., & Ochsner, K. N. (2012). A Meta-analysis of Functional Neuroimaging Studies of Self- and Other Judgments Reveals a Spatial Gradient for Mentalizing in Medial Prefrontal Cortex. *Journal of Cognitive Neuroscience*, *24*(8), 1742–1752. https://doi.org/10.1162/jocn_a_00233

Dogge, M., Custers, R., & Aarts, H. (2019). Moving Forward: On the Limits of Motor-Based Forward Models. *Trends in Cognitive Sciences*, *23*(9), 743–753. https://doi.org/10.1016/j.tics.2019.06.008

Douglass, M. D., Bain, S. A., Boland, J., Cooke, D. J., & McCarthy, P. (2023). *Investigating Individual Differences in Confession Decisions Using a Choice-Blindness Paradigm* (SSRN Scholarly Paper 4476912). https://doi.org/10.2139/ssrn.4476912

Dulany Jr., D. E., & O'Connell, D. C. (1963). Does partial reinforcement dissociate verbal rules and the behavior they might be presumed to control? *Journal of Verbal Learning & Verbal Behavior*, *2*(4), 361–372. https://doi.org/10.1016/S0022-5371(63)80105-X

Erdelyi, M. H. (1985). *Psychoanalysis: Freud's cognitive psychology* (pp. xv, 303). W H Freeman/Times Books/ Henry Holt & Co.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*(3), Article 3. https://doi.org/10.1037/0033-295X.87.3.215

Fangmeier, T., Knauff, M., Ruff, C. C., & Sloutsky, V. (2006). fMRI Evidence for a Three-Stage Model of Deductive Reasoning. *Journal of Cognitive Neuroscience*, *18*(3), 320–334. https://doi.org/10.1162/jocn.2006.18.3.320

Festinger, L. (1957). *A theory of cognitive dissonance* (pp. xi, 291). Stanford University Press.

Fleming, S. M. (2024). Metacognition and Confidence: A Review and Synthesis. *Annual Review of Psychology*, *75*(1), null. https://doi.org/10.1146/annurev-psych-022423-032425

Frith, U., & Happé, F. (1999). Theory of mind and self-consciousness: What is it like to be autistic? *Mind & Language*, *14*(1), 1–22. https://doi.org/10.1111/1468-0017.00100

Fugelsang, J. A., & Dunbar, K. N. (2005). Brain-based mechanisms underlying complex causal thinking. *Neuropsychologia*, *43*(8), 1204–1213. https://doi.org/10.1016/j.neuropsychologia.2004.10.012

Gazzaniga, M. S. (1985). *The social brain: Discovering the networks of the mind*. New York : Basic Books. http://archive.org/details/socialbraindisco0000gazz

Gazzaniga, M. S. (2014). The split-brain: Rooting consciousness in biology. *Proceedings of the National Academy of Sciences*, *111*(51), 18093–18094. https://doi.org/10.1073/pnas.1417892111

Gertler, B. (2021). Self-Knowledge. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2021/entries/self-knowledge/

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*(4), 650–669. https://doi.org/10.1037/0033-295x.103.4.650

Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences*, *16*(1), 1–14. https://doi.org/10.1017/S0140525X00028636

Goupil, L., Johansson, P., Hall, L., & Aucouturier, J.-J. (2021). Vocal signals only impact speakers' own emotions when they are self-attributed. *Consciousness and Cognition*, *88*, 103072. https://doi.org/10.1016/j.concog.2020.103072

Gur, R. C., & Sackeim, H. A. (1979). Self-deception: A concept in search of a phenomenon. *Journal of Personality and Social Psychology*, *37*(2), 147.

Haggard, P. (2017). Sense of agency in the human brain. *Nature Reviews Neuroscience*, *18*(4), Article 4. https://doi.org/10.1038/nrn.2017.14

Haggard, P., & Chambon, V. (2012). Sense of agency. *Current Biology: CB*, *22*(10), R390-392. https://doi.org/10.1016/j.cub.2012.02.040

Hall, L., Johansson, P., & Strandberg, T. (2012). Lifting the Veil of Morality: Choice Blindness and Attitude Reversals on a Self-Transforming Survey. *PloS One*, *7*, e45457. https://doi.org/10.1371/journal.pone.0045457

Hall, L., Johansson, P., Tärning, B., Sikström, S., & Deutgen, T. (2010). Magic at the marketplace: Choice blindness for the taste of jam and the smell of tea. *Cognition*, *117*(1), Article 1. https://doi.org/10.1016/j.cognition.2010.06.010

Hall, L., Strandberg, T., Pärnamets, P., Lind, A., Tärning, B., & Johansson, P. (2013). How the Polls Can Be Both Spot On and Dead Wrong: Using Choice Blindness to Shift Political Attitudes and Voter Intentions. *PloS One*, *8*, e60554. https://doi.org/10.1371/journal.pone.0060554

Happé, F. (2003). Theory of Mind and the Self. *Annals of the New York Academy of Sciences*, *1001*(1), 134–144. https://doi.org/10.1196/annals.1279.008

Harmon-Jones, E., & Mills, J. (2019). An introduction to cognitive dissonance theory and an overview of current perspectives on the theory. In *Cognitive dissonance: Reexamining a pivotal theory in psychology, 2nd ed* (pp. 3–24). American Psychological Association. https://doi.org/10.1037/0000135-001

Harris, C. L. (2006). Language and cognition. *Encyclopedia of Cognitive Science*, 1–6.

Hippel, W. von, & Trivers, R. (2011). The evolution and psychology of self-deception. *Behavioral and Brain Sciences*, *34*(1), 1–16. https://doi.org/10.1017/S0140525X10001354

Hobeika, L., Diard-Detoeuf, C., Garcin, B., Levy, R., & Volle, E. (2016). General and specialized brain correlates for analogical reasoning: A meta-analysis of functional imaging studies. *Human Brain Mapping*, *37*(5), 1953–1969. https://doi.org/10.1002/hbm.23149

Huettel, S. A., Song, A. W., McCarthy, G., Huettel, S. A., Song, A. W., & McCarthy, G. (2014). *Functional Magnetic Resonance Imaging* (Third Edition, Third Edition). Oxford University Press.

Jauhar, S., Fortea, L., Solanes, A., Albajes-Eizagirre, A., McKenna, P. J., & Radua, J. (2021). Brain activations associated with anticipation and delivery of monetary reward: A systematic review and meta-analysis of fMRI studies. *PloS One*, *16*(8), e0255292. https://doi.org/10.1371/journal.pone.0255292

Johansson, P., & Hall, L. (2008). From change blindness to choice blindness. *PSYCHOLOGIA*, *51*, 142–155. https://doi.org/10.2117/psysoc.2008.142

Johansson, P., Hall, L., Gulz, A., Haake, M., & Watanabe, K. (2007). Choice blindness and trust in the virtual world. *Technical Report of IEICE: HIP*, *107*(60), 83–86.

Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science (New York, N.Y.)*, *310*(5745), Article 5745. https://doi.org/10.1126/science.1111709

Johansson, P., Hall, L., Sikström, S., Tärning, B., & Lind, A. (2006). How something can be said about telling more than we can know: On choice blindness and introspection. *Consciousness and Cognition*, *15*(4), Article 4. https://doi.org/10.1016/j.concog.2006.09.004

Johansson, P., Hall, L., Tärning, B., Sikström, S., & Chater, N. (2013). Choice Blindness and Preference Change: You Will Like This Paper Better If You (Believe You) Chose to Read It! *Journal of Behavioral Decision Making*, *27*. https://doi.org/10.1002/bdm.1807

Johansson, P., Hall, L., Tärning, B., Sikström, S., & Chater, N. (2014). Choice Blindness and Preference Change: You Will Like This Paper Better If You (Believe You) Chose to Read It! *Journal of Behavioral Decision Making*, *27*(3), Article 3. https://doi.org/10.1002/bdm.1807

Kennedy, J. A., Anderson, C., & Moore, D. A. (2013). When overconfidence is revealed to others: Testing the status-enhancement theory of overconfidence. *Organizational Behavior and Human Decision Processes*, *122*(2), 266–279. https://doi.org/10.1016/j.obhdp.2013.08.005

Kestemont, J., Ma, N., Baetens, K., Clément, N., Van Overwalle, F., & Vandekerckhove, M. (2015). Neural correlates of attributing causes to the self, another person and the situation. *Social Cognitive and Affective Neuroscience*, *10*(1), 114–121. https://doi.org/10.1093/scan/nsu030

Kornell, N. (2014a). Where is the "meta" in animal metacognition? *Journal of Comparative Psychology*, *128*(2), 143–149. https://doi.org/10.1037/a0033444

Kornell, N. (2014b). Where to Draw the Line on Metacognition: A Taxonomy of Metacognitive Cues. *Journal of Comparative Psychology (Washington, D.C. : 1983)*, *128*, 160–162. https://doi.org/10.1037/a0036194

Kurzban, R. (2012). *Why Everyone (Else) Is a Hypocrite: Evolution and the Modular Mind*. Princeton University Press. https://doi.org/10.1515/9781400835997

Kusev, P., van Schaik, P., Teal, J., Martin, R., Hall, L., & Johansson, P. (n.d.). How false feedback influences decision-makers' risk preferences. *Journal of Behavioral Decision Making*, *n/a*(n/a). https://doi.org/10.1002/bdm.2278

Kusev, P., van Schaik, P., Teal, J., Martin, R., Hall, L., & Johansson, P. (2022). How false feedback influences decision-makers' risk preferences. *Journal of Behavioral Decision Making*, *35*(5), e2278. https://doi.org/10.1002/bdm.2278

Lachaud, L., Jacquet, B., & Baratgin, J. (2022). Reducing Choice-Blindness? An Experimental Study Comparing Experienced Meditators to Non-Meditators. *European Journal of Investigation in Health, Psychology and Education*, *12*(11), 1607–1620. https://doi.org/10.3390/ejihpe12110113

Lagnado, D. A., Newell, B. R., Kahan, S., & Shanks, D. R. (2006). Insight and strategy in multiple-cue learning. *Journal of Experimental Psychology: General*, *135*(2), 162–183. https://doi.org/10.1037/0096-3445.135.2.162

Law, E. H., Pickard, A. L., Kaczynski, A., & Pickard, A. S. (2017). Choice Blindness and Health-State Choices among Adolescents and Adults. *Medical Decision Making*, *37*(6), Article 6. https://doi.org/10.1177/0272989X17700847

Legrand, D., & Ruby, P. (2009). What is self-specific? Theoretical investigation and critical review of neuroimaging results. *Psychological Review*, *116*(1), 252–282. https://doi.org/10.1037/a0014172

Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory*, *12*(4), 361–366. https://doi.org/10.1101/lm.94705

Luo, Q., Perry, C., Peng, D., Jin, Z., Xu, D., Ding, G., & Xu, S. (2003). The neural substrate of analogical reasoning: An fMRI study. *Cognitive Brain Research*, *17*(3), 527–534. https://doi.org/10.1016/S0926-6410(03)00167-8

Lynch, R. F., & Trivers, R. L. (2012). Self-deception inhibits laughter. *Personality and Individual Differences*, *53*(4), 491–495. https://doi.org/10.1016/j.paid.2012.02.017

Mashour, G. A., Roelfsema, P., Changeux, J.-P., & Dehaene, S. (2020). Conscious Processing and the Global Neuronal Workspace Hypothesis. *Neuron*, *105*(5), 776–798. https://doi.org/10.1016/j.neuron.2020.01.026

McLaughlin, O., & Somerville, J. (2013). Choice blindness in financial decision making. *Judgment and Decision Making*, *8*(5), Article 5.

Merckelbach, H., Jelicic, M., & Pieters, M. (2011a). The residual effect of feigning: How intentional faking may evolve into a less conscious form of symptom reporting. *Journal of Clinical and Experimental Neuropsychology*, *33*(1), 131–139. https://doi.org/10.1080/13803395.2010.495055

Merckelbach, H., Jelicic, M., & Pieters, M. (2011b). The residual effect of feigning: How intentional faking may evolve into a less conscious form of symptom reporting. *Journal of Clinical and Experimental Neuropsychology*, *33*(1), Article 1. https://doi.org/10.1080/13803395.2010.495055

Moore, J. W. (2016). What Is the Sense of Agency and Why Does it Matter? *Frontiers in Psychology*, *7*, 1272. https://doi.org/10.3389/fpsyg.2016.01272

Moore, J. W., & Fletcher, P. C. (2012). Sense of agency in health and disease: A review of cue integration approaches. *Consciousness and Cognition*, *21*(1), 59–68. https://doi.org/10.1016/j.concog.2011.08.010

Moran, R. (2001). *Authority and Estrangement: An Essay on Self-Knowledge*. Princeton University Press. https://www.jstor.org/stable/j.ctt7sgs5

Mouratidou, A., Zlatev, J., & van de Weijer, J. (2022). How Much Do We Really Care What We Pick? Pre-verbal and Verbal Investment in Choices Concerning Faces and Figures. *Topoi*, *41*(4), 695–713. https://doi.org/10.1007/s11245-022-09807-z

Muda, R., Niszczota, P., & Augustynowicz, P. (2020). The effect of imperfect memory recall on risk preferences. *Journal of Behavioral Decision Making*, *33*(5), 683–690. https://doi.org/10.1002/bdm.2185

Murty, V. P., DuBrow, S., & Davachi, L. (2015). The simple act of choosing influences declarative memory. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *35*(16), 6255–6264. https://doi.org/10.1523/JNEUROSCI.4181-14.2015

Newell, B. R., & Shanks, D. R. (2014). Unconscious influences on decision making: A critical review. *The Behavioral and Brain Sciences*, *37*(1), 1–19. https://doi.org/10.1017/S0140525X12003214

Nichols, S., & Stich, S. P. (2003). *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*. Oxford University Press. http://www.oxfordscholarship.com/view/10.1093/0198236107.001.0001/acprof-9780198236108

Nisbett, R. E., & Wilson, T. D. (1977). Telling More Than We Can Know: Verbal Reports on Mental Processes. *Psychological Review*, *84*(3), 231–259.

Norman, E., Pfuhl, G., Sæle, R. G., Svartdal, F., Låg, T., & Dahl, T. I. (2019). Metacognition in Psychology. *Review of General Psychology*, *23*(4), 403–424. https://doi.org/10.1177/1089268019883821

Pärnamets, P., Hall, L., & Johansson, P. (2015). Memory distortions resulting from a choice blindness task. *CogSci*.

Pärnamets, P., Johansson, P., Strandberg, T., Balkenius, C., & Hall, L. (2023). *Looking at choice blindness: Evidence from gaze patterns and pupil dilation*. PsyArXiv. https://doi.org/10.31234/osf.io/v85sz

Pärnamets, P., Zimmermann, J. von, Raafat, R., Vogel, G., Hall, L., Chater, N., & Johansson, P. (2020). *Choice blindness and choice-induced preference change in groups*. PsyArXiv. https://doi.org/10.31234/osf.io/zut93

Paulhus, D. L., & Reid, D. B. (1991). Enhancement and denial in socially desirable responding. *Journal of Personality and Social Psychology*, *60*(2), 307–317. https://doi.org/10.1037/0022-3514.60.2.307

Perlovsky, L., & Sakai, K. L. (2014). Language and Cognition. *Frontiers in Behavioral Neuroscience*, *8*, 436. https://doi.org/10.3389/fnbeh.2014.00436

Petitmengin, C., Remillieux, A., Cahour, B., & Carter-Thomas, S. (2013). A gap in Nisbett and Wilson's findings? A first-person access to our cognitive processes. *Consciousness and Cognition*, *22*(2), 654–669. https://doi.org/10.1016/j.concog.2013.02.004

Petitmengin, C., Remillieux, A., & Valenzuela-Moguillansky, C. (2019). Discovering the structures of lived experience. *Phenomenology and the Cognitive Sciences*, *18*(4), 691–730. https://doi.org/10.1007/s11097-018-9597-4

Pinker, S. (2011). Representations and decision rules in the theory of self-deception. *Behavioral and Brain Sciences*, *34*(1), 35–37. https://doi.org/10.1017/S0140525X1000261X

Poorun, T., Almeida-Lopez, P., Fernanda-Cadena, L., Jahn, N., Opoku, M., Johansson, P., & Hall, L. (2018). *Is Choice Blindness accounted for by Individual Differences in Personality, Working Memory and Visual Working Memory?* https://doi.org/10.13140/RG.2.2.32717.13282

Pronin, E., Wegner, D. M., McCarthy, K., & Rodriguez, S. (2006). Everyday magical powers: The role of apparent mental causation in the overestimation of personal influence. *Journal of Personality and Social Psychology*, *91*(2), 218–231. https://doi.org/10.1037/0022-3514.91.2.218

Rebouillat, B., Leonetti, J. M., & Kouider, S. (2021). People confabulate with high confidence when their decisions are supported by weak internal variables. *Neuroscience of Consciousness*, *2021*(1). https://doi.org/10.1093/nc/niab004

Remington, A., White, H., Fairnie, J., Hall, L., & Johansson, P. (2024). *Choice Blindness in Autistic and Non-autistic People*. https://doi.org/10.31219/osf.io/s4d6w

Ridderinkhof, K. R., Ullsperger, M., Crone, E. A., & Nieuwenhuis, S. (2004). The Role of the Medial Frontal Cortex in Cognitive Control. *Science*, *306*(5695), Article 5695. https://doi.org/10.1126/science.1100301

Rieznik, A., Moscovich, L., Frieiro, A., Figini, J., Catalano, R., Garrido, J. M., Alvarez Heduan, F., Sigman, M., & Gonzalez, P. A. (2017). A massive experiment on choice blindness in political decisions: Confidence, confabulation, and unconscious detection of self-deception. *PloS One*, *12*(2), e0171108. https://doi.org/10.1371/journal.pone.0171108

Sagana, A. (2015). *A blind man's bluff: Choice blindness in eyewitness testimony* [Doctoral Thesis, Maastricht University]. https://doi.org/10.26481/dis.20150917as

Sagana, A., Sauerland, M., & Merckelbach, H. (2013). Witnesses' blindness for their own facial recognition decisions: A field study. *Behavioral Sciences & the Law*, *31*(5), 624–636. https://doi.org/10.1002/bsl.2082

Sagana, A., Sauerland, M., & Merckelbach, H. (2014a). Memory impairment is not sufficient for choice blindness to occur. *Frontiers in Psychology*, *5*, 449. https://doi.org/10.3389/fpsyg.2014.00449

Sagana, A., Sauerland, M., & Merckelbach, H. (2014b). Memory impairment is not sufficient for choice blindness to occur. *Frontiers in Psychology*, *5*, 449. https://doi.org/10.3389/fpsyg.2014.00449

Sagana, A., Sauerland, M., & Merckelbach, H. (2014c). 'This Is the Person You Selected': Eyewitnesses' Blindness for Their Own Facial Recognition Decisions: Witnesses' blind facial recognition decisions. *Applied Cognitive Psychology*, *28*(5), 753–764. https://doi.org/10.1002/acp.3062

Sagana, A., Sauerland, M., & Merckelbach, H. (2016). The effect of choice reversals on blindness for identification decisions. *Psychology, Crime & Law*, *22*(4), 303–314. https://doi.org/10.1080/1068316X.2015.1085984

Sagana, A., Sauerland, M., & Merckelbach, H. (2018). Warnings to Counter Choice Blindness for Identification Decisions: Warnings Offer an Advantage in Time but Not in Rate of Detection. *Frontiers in Psychology*, *9*, 981. https://doi.org/10.3389/fpsyg.2018.00981

Sauerland, M., Sagana, A., & Otgaar, H. (2013). Theoretical and legal issues related to choice blindness for voices. *Legal and Criminological Psychology*, *18*(2), Article 2. https://doi.org/10.1111/j.2044-8333.2012.02049.x

Sauerland, M., Sagana, A., Otgaar, H., & Broers, N. J. (2014). Self-Relevance Does Not Moderate Choice Blindness in Adolescents and Children. *PLOS ONE*, *9*(6), Article 6. https://doi.org/10.1371/journal.pone.0098563

Sauerland, M., Sagana, A., Siegmann, K., Heiligers, D., Merckelbach, H., & Jenkins, R. (2016). These two are different. Yes, they're the same: Choice blindness for facial identity. *Consciousness and Cognition*, *40*, 93–104. https://doi.org/10.1016/j.concog.2016.01.003

Sauerland, M., Schell, J. M., Collaris, J., Reimer, N. K., Schneider, M., & Merckelbach, H. (2013). 'Yes, I have sometimes stolen bikes': Blindness for norm-violating behaviors and implications for suspect interrogations. *Behavioral Sciences & the Law*, *31*(2), Article 2. https://doi.org/10.1002/bsl.2063

Scaife, R. (2014). A Problem for Self-Knowledge: The Implications of Taking Confabulation Seriously. *Acta Analytica*, *29*(4), 469–485.

Schelling, T. C. (1960). *The strategy of conflict* (pp. vii, 303). Harvard Univer. Press.

Schultheiss, O. C., & Strasser, A. (2012). Referential processing and competence as determinants of congruence between implicit and explicit motives. In *Handbook of self-knowledge* (pp. 39–62). The Guilford Press.

Schwardmann, P., & van der Weele, J. (2019). Deception and self-deception. *Nature Human Behaviour*, *3*(10), Article 10. https://doi.org/10.1038/s41562-019-0666-7

Schwartz, B. L., Benjamin, A. S., & Bjork, R. A. (1997). The Inferential and Experiential Bases of Metamemory. *Current Directions in Psychological Science*, *6*(5), 132–137. https://doi.org/10.1111/1467-8721.ep10772899

Schwartz, S. H. (1966). Trial-by-trial analysis of processes in simple and disjunctive concept-attainment tasks. *Journal of Experimental Psychology*, *72*(3), 456–465. https://doi.org/10.1037/h0023652

Seghezzi, S., Zirone, E., Paulesu, E., & Zapparoli, L. (2019). The Brain in (Willed) Action: A Meta-Analytical Comparison of Imaging Studies on Motor Intentionality and Sense of Agency. *Frontiers in Psychology*, *10*, 804. https://doi.org/10.3389/fpsyg.2019.00804

Simler, K., & Hanson, R. (2017). *The elephant in the brain: Hidden motives in everyday life*. Oxford University Press.

Slovic, P., & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance*, *6*(6), 649–744. https://doi.org/10.1016/0030-5073(71)90033-X

Smith, M. K., Trivers, R., & von Hippel, W. (2017). Self-deception facilitates interpersonal persuasion. *Journal of Economic Psychology*, *63*, 93–101. https://doi.org/10.1016/j.joep.2017.02.012

Somerville, J., & McGowan, F. (2016). Can chocolate cure blindness? Investigating the effect of preference strength and incentives on the incidence of Choice Blindness. *Journal of Behavioral and Experimental Economics*, *61*, 1–11. https://doi.org/10.1016/j.socec.2016.01.001

Sperduti, M., Delaveau, P., Fossati, P., & Nadel, J. (2011). Different brain structures related to self- and external-agency attribution: A brief review and meta-analysis. *Brain Structure and Function*, *216*(2), 151–157. https://doi.org/10.1007/s00429-010-0298-1

Steenfeldt-Kristensen, C., & Thornton, I. M. (2013). Haptic Choice Blindness. *I-Perception*, *4*(3), Article 3. https://doi.org/10.1068/i0581sas

Stille, L., Norin, E., & Sikstrom, S. (2017). Self-delivered misinformation—Merging the choice blindness and misinformation effect paradigms. *PloS One*, *12*(3), e0173606. https://doi.org/10.1371/journal.pone.0173606

Stöber, J. (2001). The Social Desirability Scale-17 (SDS-17): Convergent validity, discriminant validity, and relationship with age. *European Journal of Psychological Assessment*, *17*(3), 222–232. https://doi.org/10.1027/1015-5759.17.3.222

Strandberg, T. (2020). *The malleability of political attitudes: Choice blindness, confabulation and attitude change*. https://portal.research.lu.se/en/publications/the-malleability-of-political-attitudes-choice-blindness-confabul

Strandberg, T., Hall, L., Johansson, P., Björklund, F., & Pärnamets, P. (2019, July 8). Correction of manipulated responses in the choice blindness paradigm: What are the predictors? *Proceedings of the 41st Annual Conference of the Cognitive Science Society*. CogSci 2019: The 41st Annual Meeting of the Cognitive Science Society. https://portal.research.lu.se/portal/en/publications/correction-of-manipulated-responses-in-the-choice-blindness-paradigm(c5497a38-1d15-44bc-8b23-8a501e5dc27f).html

Strandberg, T., Olson, J. A., Hall, L., Woods, A., & Johansson, P. (2020). Depolarizing American voters: Democrats and Republicans are equally susceptible to false attitude feedback. *PLOS ONE*, *15*(2), e0226799. https://doi.org/10.1371/journal.pone.0226799

Strandberg, T., Sivén, D., Hall, L., Johansson, P., & Pärnamets, P. (2018). False beliefs and confabulation can lead to lasting changes in political attitudes. *Journal of Experimental Psychology. General*, *147*, 1382–1399. https://doi.org/10.1037/xge0000489

Synofzik, M., Vosgerau, G., & Newen, A. (2008). Beyond the comparator model: A multifactorial two-step account of agency. *Consciousness and Cognition*, *17*(1), Article 1. https://doi.org/10.1016/j.concog.2007.03.010

Taya, F., Gupta, S., Farber, I., & Mullette-Gillman, O. A. (2014). Manipulation Detection and Preference Alterations in a Choice Blindness Paradigm. *PLOS ONE*, *9*(9), Article 9. https://doi.org/10.1371/journal.pone.0108515

Tenney, E. R., Vazire, S., & Mehl, M. R. (2013). This examined life: The upside of self-knowledge for interpersonal relationships. *PloS One*, *8*(7), e69605. https://doi.org/10.1371/journal.pone.0069605

Trivers, R. (2011). *The folly of fools: The logic of deceit and self-deception in human life*. Basic Books (AZ).

Trouche, E., Johansson, P., Hall, L., & Mercier, H. (2015). The Selective Laziness of Reasoning. *Cognitive Science*, *40*. https://doi.org/10.1111/cogs.12303

Trouche, E., Johansson, P., Hall, L., & Mercier, H. (2018). Vigilant conservatism in evaluating communicated information. *PLOS ONE*, *13*, e0188825. https://doi.org/10.1371/journal.pone.0188825

Undorf, M., & Erdfelder, E. (2011). Judgments of learning reflect encoding fluency: Conclusive evidence for the ease-of-processing hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1264–1269. https://doi.org/10.1037/a0023719

Vaccaro, A. G., & Fleming, S. M. (2018). Thinking about thinking: A coordinate-based meta-analysis of neuroimaging studies of metacognitive judgements. *Brain and Neuroscience Advances*, *2*, 2398212818810591. https://doi.org/10.1177/2398212818810591

Valk, S. L., Bernhardt, B. C., Böckler, A., Kanske, P., & Singer, T. (2016). Substrates of metacognition on perception and metacognition on higher-order cognition relate to different subsystems of the mentalizing network. *Human Brain Mapping*, *37*(10), 3388–3399. https://doi.org/10.1002/hbm.23247

Van Overwalle, F. (2011). A dissociation between social mentalizing and general reasoning. *NeuroImage*, *54*(2), 1589–1599. https://doi.org/10.1016/j.neuroimage.2010.09.043

VandenBos, G. R. (2007). *APA dictionary of psychology*. American Psychological Association. https://psycnet.apa.org/record/2006-11044-000

Vogel, G., Hall, L., Moore, J., & Johansson, P. (2024). The right face at the wrong place: How motor intentions can override outcome monitoring. *iScience*, *27*(1), 108649. https://doi.org/10.1016/j.isci.2023.108649

Vogel, G., Mårtensson, J., Mannfolk, P., Hall, L., Westen, D. van, & Johansson, P. (2024a). *Catching the brain in the act of confabulation: A fMRI study*. PsyArXiv. https://doi.org/10.31234/osf.io/e9vg7

Vogel, G., Mårtensson, J., Mannfolk, P., Hall, L., Westen, D. van, & Johansson, P. (2024b). *The neural correlates of outcome monitoring and false feedback detection in choice blindness: A fMRI study*. PsyArXiv. https://doi.org/10.31234/osf.io/smecn

Vogel, G., Pärnamets, P., Hall, L., & Johansson, P. (2023). *Choice blindness without deception: Failures to notice false-feedbacks persist in explicit detection tasks*.

Vranka, M. A., & Bahnik, S. (2016). Is the Emotional Dog Blind to Its Choices? *Experimental Psychology*, *63*(3), 180–188. https://doi.org/10.1027/1618-3169/a000325

Wang, L., Zhang, M., Zou, F., Wu, X., & Wang, Y. (2020). Deductive-reasoning brain networks: A coordinate-based meta-analysis of the neural signatures in deductive reasoning. *Brain and Behavior*, *10*(12), e01853. https://doi.org/10.1002/brb3.1853

Wang, Y., Zhao, S., Zhang, Z., & Feng, W. (2018). Sad Facial Expressions Increase Choice Blindness. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.02300

Wegner, D. M. (2002). *The illusion of conscious will* (pp. xi, 405). MIT Press.

Wegner, D. M., & Wheatley, T. (1999). Apparent mental causation. Sources of the experience of will. *The American Psychologist*, *54*(7), 480–492.

Wells, G. L., & Petty, R. E. (1980). The effects of overt head movements on persuasion: Compatibility and incompatibility of responses. *Basic and Applied Social Psychology*, *1*(3), 219–230. https://doi.org/10.1207/s15324834basp0103_2

Wertheim, J., & Ragni, M. (2018). The Neural Correlates of Relational Reasoning: A Meta-analysis of 47 Functional Magnetic Resonance Studies. *Journal of Cognitive Neuroscience*, *30*(11), 1734–1748. https://doi.org/10.1162/jocn_a_01311

Wilson, T. D. (2009). Know Thyself. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *4*(4), 384–389. https://doi.org/10.1111/j.1745-6924.2009.01143.x

Wilson, T. D., & Dunn, E. W. (2004). Self-Knowledge: Its Limits, Value, and Potential for Improvement. *Annual Review of Psychology*, *55*(1), 493–518. https://doi.org/10.1146/annurev.psych.55.090902.141954

Wirth, M., de Paula Couto, M. C., Pavlova, M. K., & Rothermund, K. (2023). Manipulating prescriptive views of active aging and altruistic disengagement. *Psychology and Aging*, *38*(8), 854–881. https://doi.org/10.1037/pag0000763

Wong, S. F., Aardema, F., Giraldo-O'Meara, M., Hall, L., & Johansson, P. (2020). Choice Blindness, Confabulatory Introspection, and Obsessive–Compulsive Symptoms: Investigation in a Clinical Sample. *Cognitive Therapy and Research*, *44*(2), 376–385. https://doi.org/10.1007/s10608-019-10066-3

Woods, A. J., Hamilton, R. H., Kranjec, A., Minhaus, P., Bikson, M., Yu, J., & Chatterjee, A. (2014). Space, time, and causality in the human brain. *NeuroImage*, *92*, 285–297. https://doi.org/10.1016/j.neuroimage.2014.02.015

Wright, G. R. T., Berry, C. J., Catmur, C., & Bird, G. (2015). Good Liars Are Neither 'Dark' Nor Self-Deceptive. *PLoS ONE*, *10*(6), e0127315. https://doi.org/10.1371/journal.pone.0127315

Yeung, N., & Summerfield, C. (2014). Shared Mechanisms for Confidence Judgements and Error Detection in Human Decision Making. In S. M. Fleming & C. D. Frith (Eds.), *The Cognitive Neuroscience of Metacognition* (pp. 147–167). Springer. https://doi.org/10.1007/978-3-642-45190-4_7

Zander, D., Vogel, G., & Johansson, P. (2023). *Mechanisms for preserving transitive preferences: Transitivity in the Context of Choice Blindness*. https://doi.org/10.13140/RG.2.2.22427.85286

Zhang, Q., Lu, Y., Huangfu, H., & Fu, S. (2020). Impacts of the Time Interval on the Choice Blindness Persistence: A Visual Cognition Test-Based Study. In H. Ayaz (Ed.), *Advances in Neuroergonomics and Cognitive Engineering* (pp. 233–242). Springer International Publishing.

Zylberberg, A., Barttfeld, P., & Sigman, M. (2012). The construction of confidence in a perceptual decision. *Frontiers in Integrative Neuroscience*, *6*. https://www.frontiersin.org/articles/10.3389/fnint.2012.00079

Annex:
List of choice blindness studies

Table annex 1. List of choice blindness studies till 2024.

Study	Domain	Stimuli	Attitude change	Detail attitude change
(Johansson et al., 2005)	Facial attractiveness	Pictures of faces	NO	
(Johansson et al., 2006)	Facial attractiveness	Pictures of faces	NO	
(Johansson et al., 2007)	Facial attractiveness	Pictures of faces	NO	
(Johansson & Hall, 2008)	Aesthetic judgment	Abstract Patterns	NO	
(Johansson & Hall, 2008)	Facial attractiveness	Pictures of faces	NO	
(Hall et al., 2010)	Consumer decision	Jam and tea	NO	
(Hall et al., 2010)	Consumer decision	Tea	NO	
(Merckelbach et al., 2011)	Health	Psychiatric symptoms	YES	Symptom report
(Hall et al., 2012)	Moral judgment	Moral principles and issues	NO	
(Sauerland, Sagana, et al., 2013)	Sympathy judgment	Voice recordings	NO	
(Sauerland, Schell, et al., 2013)	Norm violation	Norm violating behaviour questionnaire	YES	Change of answer about norm violating behaviour
(Hall et al., 2013)	Political judgment	Political survey	YES	Voting intentions
(McLaughlin & Somerville, 2013)	Financial decision	Investment portfolios	YES	From verbal reports
(Sagana et al., 2013)	Eyewitness/forensic psychology	Pictures of faces	NO	
(Steenfeldt-Kristensen & Thornton, 2013)	Tactile preference	Objects	NO	
(Petitmengin et al., 2013)	Facial attractiveness	Pictures of faces	NO	
(Johansson et al., 2014)	Facial attractiveness	Pictures of faces	YES	Ratings and Choice Consistency
(Sagana et al., 2014b)	Sympathy judgment	Pictures of faces	NO	
(Sagana et al., 2014c)	Eyewitness/forensic psychology	Pictures of faces	NO	
(Sauerland et al., 2014)	Preference for objects	Pictures of objects	NO	
(Aardema et al., 2014)	Clinical	Scenario of accident	NO	
(Taya et al., 2014)	Facial attractiveness	Pictures of faces	YES	Ratings
(Sagana, 2015)	Eyewitness/forensic psychology	Pictures of faces	NO	
(Pärnamets et al., 2015)	Memory	Pictures of faces	NO	

(Trouche et al., 2015)	Reasoning	Reasoning problems, Arguments	NO	
(Somerville & McGowan, 2016)	Food preference	Pictures of chocolates	NO	
(Somerville & McGowan, 2016)	Facial attractiveness	Pictures of faces	NO	
(Cheung et al., 2016)	Consumer decision	Can of soup	NO	
(Vranka & Bahnik, 2016)	Moral judgment	Descriptions of morally ambiguous behaviours	NO	
(Sauerland et al., 2016)	Eyewitness/forensic psychology	Pictures of faces	NO	
(Sagana et al., 2016)	Eyewitness/forensic psychology	Lineups	NO	
(Cochran et al., 2016)	Eyewitness/forensic psychology	Lineups	NO	
(Rieznik et al., 2017)	Political judgment	Political survey	YES	Voting intentions
(Law et al., 2017)	Health	Health state scenarios	NO	
(Stille et al., 2017)	Memory	Videos of events	NO	
(Strandberg et al., 2018)	Political judgment	Political issues	YES	Ratings
(Trouche et al., 2018)	Reasoning	General knowledge	NO	
(Y. Wang et al., 2018)	Facial attractiveness	Pictures of faces	NO	
(Sagana et al., 2018)	Eyewitness/forensic psychology	Pictures of faces	NO	
(Poorun et al., 2018)	Facial attractiveness	Pictures of faces	NO	
(Strandberg et al., 2019)	Political judgment	Political survey	NO	
(Strandberg et al., 2020)	Political judgment	Political survey	YES	Verbal reports
(Zhang et al., 2020)	Aesthetic judgment	Pictures of natural sceneries	NO	
(Wong et al., 2020)	Clinical	Scenario of accident	NO	
(Muda et al., 2020)	Risk preference	Monetary gambles	YES	Choice consistency
(Rebouillat et al., 2021)	Selective attention	Pictures of faces	NO	
(Mouratidou et al., 2022)	Facial attractiveness	Pictures of faces	YES	Verbal reports
(Mouratidou et al., 2022)	Aesthetic judgment	Abstract Patterns	YES	Verbal reports
(Kusev et al., 2022)	Risk preference	Monetary gambles	YES	Choice consistency
(Lachaud et al., 2022)	Facial attractiveness	Pictures of faces	NO	
(Wirth et al., 2023)	Behavioural norms	Prescriptive view of aging questionnaire	NO	
(Artenie et al., 2023)	Experiential avoidance	Experiential avoidance questionnaire	YES	Choice consistency
(Vogel, Hall, et al., 2024)	Facial attractiveness	Pictures of faces	YES	Choice consistency
McKay, R., Strandberg, T., Hall, L., & Johansson, P. (in prep)	Religious attitude	Religious statements	NO	
Ambrus, E., Hartig, B., Johansson, P. & McKay, R. (Submitted)	Own Personality	Personality traits	NO	
(Remington et al., 2024)	Clinical	Pictures of faces	YES	Choice consistency
(Pärnamets et al., 2023)	Facial attractiveness	Pictures of faces	NO	
(Vogel et al., 2023) [paper 1]	Facial attractiveness	Pictures of faces	YES	Choice consistency
(Vogel, Mårtensson, et al., 2024b) [paper 2]	Neuroimaging	Pictures of faces	YES	Choice consistency

(Vogel, Mårtensson, et al., 2024a) [paper 4]	Neuroimaging	Pictures of faces	YES	Choice consistency
(Pärnamets et al., 2020)	Flatmate choice	Pictures of faces	YES	Choice consistency
Pärnamets, Hall, L., & Johansson, P. (in prep).	Reasoning	Reasoning problems	NO	
(Douglass et al., 2023)	Norm violation	Norm violating behaviour questionnaire	NO	